Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Análisis de sentimientos
Análisis de sentimiento en el dominio salud: analizando comentarios sobre fármacos
Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, Paloma Martínez 15
On the effect of word order on cross-lingual sentiment analysis
Àlex R. Atrio, Toni Badia, Jeremy Barnes 23

Traducción automática
Adapting NMT to caption translation in Wikimedia Commons for low-resource languages
Alberto Poncelas, Kepa Sarasola, Meghan Dowling, Andy Way, Gorka Labaka, Iñaki Alegria 33
Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon
Ximena Gutierrez-Vasques, Alfonso Medina-Urrea, Gerardo Sierra 41

Análisis textual
Evaluation of distributional semantic models for the extraction of semantic relations for named rivers from a small specialized corpus
Juan Rojas-García, Pamela Faber 51
Spelling normalisation of Basque historical texts
Ainara Estarrona, Izaskun Etxeberria, Ander Soraluze, Manuel Padilla 59
Automatic proficiency classification in L2 Portuguese
Iria del Río 67

Procesamiento multilingüe
Cross-lingual diachronic distance: application to Portuguese and Spanish
José Ramom Pichel, Pablo Gamallo, Iñaki Alegría 77
Action type induction from multilingual lexical features
Lorenzo Gregori, Rossella Varvara, Andrea Amelio Ravelli 85

Técnicas de generación de resúmenes y simplificación textual
Lexical simplification approach using easy-to-read resources
Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, Paloma Martínez 95
Generación de resúmenes extractivos de múltiples documentos usando grafos semánticos
Oleyda del Camino Valle, Alfredo Simón-Cuevas, Eduardo Valladares-Valdés, José A. Olivas, Francisco P. Romero 103
Applying siamese hierarchical attention neural networks for multi-document summarization
José Angel González, Julien Delonca, Emilio Sanchis, Fernando García-Granada, Encarna Segarra . 111

Procesamiento del lenguaje y técnicas basadas en grafos
Benchmarking entity linking for question answering over knowledge graphs
Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas 121
Landscaping language technologies using topic modeling and graph analysis: overview of the Spanish contribution
Doaa Samy, David Pérez-Fernández, Jerónimo Arenas-García 129

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

L. Alfonso Ureña López	Universidad de Jaén	laurena@ujaen.es	(Director)
Patricio Martínez Barco	Universidad de Alicante	patricio@dlsi.ua.es	(Secretario)
Manuel Palomar Sanz	Universidad de Alicante	mpalomar@dlsi.ua.es	
Felisa Verdejo Maíllo	UNED	felisa@lsi.uned.es	

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad del País Vasco
Año de edición: 2019
Editores: Inmaculada Hernáez Rioja, Escuela de Ingeniería de Bilbao UPV/EHU, inma.hernaez@ehu.eus
Eva Navas Cordón, Escuela de Ingeniería de Bilbao UPV/EHU, eva.navas@ehu.eus
Jon Sánchez, Escuela de Ingeniería de Bilbao UPV/EHU, jon.sanchez@ehu.eus

Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Consejo asesor

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón Masalles	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández Rodríguez	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola Galletebeitia	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
Ramón López-Cozar Delgado	Universidad de Granada (España)
José Miguel Goñi Menoyo	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Antònia Martí Antonín	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Patricio Martínez-Barco	Universidad de Alicante (España)
Eugenio Martínez Cámara	Universidad de Granada (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lluís Padró Cirera	Universidad Politécnica de Cataluña (España)
Manuel Palomar Sanz	Universidad de Alicante (España)
Ferrán Pla Santamaría	Universidad Politécnica de Valencia (España)
German Rigau Claramunt	Universidad del País Vasco (España)
Horacio Rodríguez Hontoria	Universidad Politécnica de Cataluña (España)
Paolo Rosso	Universidad Politécnica de Valencia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Emilio Sanchís Arnal	Universidad Politécnica de Valencia (España)
Kepa Sarasola Gabiola	Universidad del País Vasco (España)
Encarna Segarra Soriano	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé Delor	Universidad de Barcelona (España)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares Ferro	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Rodrigo Agerri	Universidad del País Vasco (España)
Francisco J. Ribadas-Pena	Universidad de Vigo (España)
Manuel Maña	Universidad de Huelva (España)
Pilar López	Universidad de Jaén (España)
Lea Canales	Savana (España)
Gorka Labaka	Universidad del País Vasco (España)
Itziar González-Dios	Universidad del País Vasco (España)
Aitziber Atutxa	Universidad del País Vasco (España)
Soto Montalvo	Universidad Nacional de Educación a Distancia (UNED)
Salud María Jiménez-Zafra	Universidad de Jaén (España)
Noa Cruz	Centro de Excelencia de Inteligencia Artificial. Bankia (España)
Víctor Manuel Darriba Bilbao	Universidad de Vigo (España)
Fernando Martínez-Santiago	Universidad de Jaén (España)

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preámbulo

La revista *Procesamiento del Lenguaje Natural* pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. Además, se quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Anualmente la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) publica dos números de la revista, que incluyen artículos originales, presentaciones de proyectos en marcha, reseñas bibliográficas y resúmenes de tesis doctorales. Esta revista se distribuye gratuitamente a todos los socios, y con el fin de conseguir una mayor expansión y facilitar el acceso a la publicación, su contenido es libremente accesible por Internet.

Las áreas temáticas tratadas son las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Lingüística de corpus
- Desarrollo de recursos y herramientas lingüísticas
- Gramáticas y formalismos para el análisis morfológico y sintáctico
- Semántica, pragmática y discurso
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica
- Aprendizaje automático en PLN
- Generación textual monolingüe y multilingüe
- Traducción automática
- Reconocimiento y síntesis del habla
- Extracción y recuperación de información monolingüe, multilingüe y multimodal
- Sistemas de búsqueda de respuestas
- Análisis automático del contenido textual
- Resumen automático
- PLN para la generación de recursos educativos
- PLN para lenguas con recursos limitados
- Aplicaciones industriales del PLN
- Sistemas de diálogo
- Análisis de sentimientos y opiniones
- Minería de texto
- Evaluación de sistemas de PLN
- Implicación textual y paráfrasis

El ejemplar número 63 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a tres apartados diferentes: comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas (demostraciones).

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista. Queremos agradecer a los miembros del Comité Asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 42 trabajos para este número, de los cuales 25 eran artículos científicos y 17 correspondían a resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas. De entre los 25 artículos recibidos, 14 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 56%.

El Comité Asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato: se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso, cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones, sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del comité editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Septiembre de 2019
Los editores.

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, ongoing projects, book reviews and summaries of doctoral theses. All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 63th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, research project summaries and description of Natural Language Processing software tools. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Fourty-two papers were submitted for this issue, from which twenty-five were scientific papers and seventeen were either projects or tool description summaries. From these twenty-five papers, we selected fourteen (56%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation of those papers with a difference of three or more points out of seven in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit were supervised by the editorial board. The cut-off criterion adopted was the mean of the three scores given.

September 2019
Editorial board.

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Análisis de sentimientos
Análisis de sentimiento en el dominio salud: analizando comentarios sobre fármacos
Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, Paloma Martínez 15
On the effect of word order on cross-lingual sentiment analysis
Àlex R. Atrio, Toni Badia, Jeremy Barnes 23

Traducción automática
Adapting NMT to caption translation in Wikimedia Commons for low-resource languages
Alberto Poncelas, Kepa Sarasola, Meghan Dowling, Andy Way, Gorka Labaka, Iñaki Alegria 33
Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon
Ximena Gutierrez-Vasques, Alfonso Medina-Urrea, Gerardo Sierra 41

Análisis textual
Evaluation of distributional semantic models for the extraction of semantic relations for named rivers from a small specialized corpus
Juan Rojas-García, Pamela Faber 51
Spelling normalisation of Basque historical texts
Ainara Estarrona, Izaskun Etxeberria, Ander Soraluze, Manuel Padilla 59
Automatic proficiency classification in L2 Portuguese
Iria del Río 67

Procesamiento multilingüe
Cross-lingual diachronic distance: application to Portuguese and Spanish
José Ramom Pichel, Pablo Gamallo, Iñaki Alegría 77
Action type induction from multilingual lexical features
Lorenzo Gregori, Rossella Varvara, Andrea Amelio Ravelli 85

Técnicas de generación de resúmenes y simplificación textual
Lexical simplification approach using easy-to-read resources
Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, Paloma Martínez 95
Generación de resúmenes extractivos de múltiples documentos usando grafos semánticos
Oleyda del Camino Valle, Alfredo Simón-Cuevas, Eduardo Valladares-Valdés, José A. Olivas, Francisco P. Romero 103
Applying siamese hierarchical attention neural networks for multi-document summarization
José Angel González, Julien Delonca, Emilio Sanchis, Fernando García-Granada, Encarna Segarra .. 111

Procesamiento del lenguaje y técnicas basadas en grafos
Benchmarking entity linking for question answering over knowledge graphs
Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas 121
Landscaping language technologies using topic modeling and graph analysis: overview of the Spanish contribution
Doaa Samy, David Pérez-Fernández, Jerónimo Arenas-García 129

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Proyectos

MENHIR: mental health monitoring through interactive conversations
Zoraida Callejas, Kawtar Benghazi, Manuel Noguera, María Inés Torres, Raquel Justo 139
Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad
Alejandro Molina-Villegas, Oscar S. Siordia, Edwing Aldana-Bobadilla, César Aguilar, Olga Acosta . 143
LIHLITH: Learning to Interact with Humans by Lifelong Interaction with Humans
Eneko Agirre, Arantxa Otegi, Camille Pradel, Sophie Rosset, Anselmo Peñas y Mark Cieliebak..... 147
Unsupervised neural machine translation, a new paradigm solely based on monolingual text
Mikel Artetxe, Gorka Labaka, Eneko Agirre 151
Project CAVIAR CApturing VIewers' Affective Response
Fernando Fernández-Martínez, Zoraida Callejas, Ricardo Kleinlein, Cristina Luna Jiménez, Juan Manuel Montero, José Manuel Pardo 155
LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos
Itziar Aldabe, Josu Aztiria, Francho Beltrán, Myriam Bras, Klara Ceberio, Itziar Cortes, Jean-Baptiste Coyos, Benaset Dazeas, Luise Esher, Gorka Labaka, Igor Leturia, Kepa Sarasola, Aure Séguier, Jean Sibille..... 159
Prevención de enfermedades infecciosas basado en el análisis inteligente en RRSS y participación ciuidadana
Óscar Apolinario, José Medina-Moreira, Harry Luna-Aveiga, José Antonio García-Díaz, Rafael Valencia-Garcia, José Ignacio Estrade-Cabrera 163
Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del plan de Impulso de las tecnologías del lenguaje
Antonio Moreno, Doroteo Torre, Ana Valverde, Leonardo Campillos 167
EXTRAE: EXTRacción de Asociaciones entre Enfermedades y otros conceptos médicos
Lourdes Araujo, Juan Martínez-Romo, Andrés Duque, Fernando López-Ostenero, Ricardo Sánchez de Madariaga, Adolfo Muñoz Carrero, Mario Pascual Carrasco 171
Enriched communication across the lifespan
Martin Cooke, Simon King, Valerie Hazan, Yannis Stylianou, Esther Janse, Deniz Baskent, Volker Hohmann, Axel Winneke, Inma Hernaez..... 175
Social-Univ 2.0: tecnologías del lenguaje humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante
Isabel Moreno, Javi Fernández, Yoán Gutiérrez..... 179

Demos

Asistente conversacional móvil para la rehabilitación de la tartamudez
David Griol, Javier Font, Zoraida Callejas 185
Plataforma web para la evaluación de la comprensión lectora aplicada a tareas de búsqueda
Javier Perea de la Casa, Fernando Javier Martínez Santiago, Miguel Ángel García Cumbreras, Manuel Carlos Díaz Galiano 189
Corpus viewer: NLP and ML-based platform for public policy making and implementation
David Pérez-Fernández, Jerónimo Arenas-García, Doaa Samy, Antonio Padilla-Soler, Vanesa Gómez-Verdejo 193
Atención automatizada a estudiantes en el proceso de matriculación en la Universidad de Alicante
Isabel Moreno, Yoan Gutierrez, Andrés Montoyo 197
Social-Univ 2.0: monitorización del entorno digital social universitario
Isabel Moreno, Javi Fernández, Yoán Gutiérrez..... 201
Terminological and language resources for developing a virtual patient dialogue system in Spanish
Leonardo Campillos-Llanos, Catherine Thomasa, Éric Bilinski, Piere Zwigenbaum, Sophie Rosset 205

Información General

Información para los autores 211
Información adicional..... 213

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Análisis de Sentimientos

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 15-22 recibido 28-03-2019 revisado 07-05-2019 aceptado 17-05-2019

Análisis de Sentimiento en el dominio salud: analizando comentarios sobre fármacos

Sentiment analysis on health domain: analyzing patient comments on drugs

Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, Paloma Martínez
Computer Science Department, University Carlos III of Madrid
{ccolon, isegura, pmf}@inf.uc3m.es

Resumen: En el análisis de sentimiento, la mayoría de las investigaciones han sido llevadas a cabo en dominios generales tales como en el análisis de opiniones de películas, restaurantes y otros productos o servicios, con escasa representación en el ámbito médico. Cada vez más, los pacientes buscan información en internet sobre los posibles beneficios, efectos adversos y opiniones que otros pacientes tiene sobre diferentes fármacos. El objetivo de este trabajo es predecir el grado de satisfacción de los pacientes respecto a un determinado fármaco en base a sus comentarios. Para llevar a cabo la tarea, hemos utilizado una colección de comentarios sobre diferentes tipos de fármacos y aplicado una red convolucional para la clasificación de los mismos. Los resultados muestran que este tipo de redes proporciona mejores resultados en términos de precisión, recall y f1-score que empleando algoritmos clásicos de clasificación como las maquinas de soporte vectorial.
Palabras clave: Análisis de Sentimiento, Clasificación de textos multi-clase, Deep Learning, Redes convolucionales

Abstract: Most of the research in sentiment analysis has been conducted in general domains such as the analysis of film reviews, restaurants and other products or services, but without much representation in the medical field. Increasingly, patients are searching the internet for information about the potential benefits, adverse effects and opinions that other patients have about different drugs. The aim of this work is to predict the degree of patient satisfaction with a given drug based on their reviews. To carry out the task, we have used a collection of reviews on different types of drugs and applied a convolutional network for their classification. The results show that this type of network provides better results in terms of precision, recall and f1-score than using classical classification algorithms such as vector support machines.
Keywords: Sentiment Analysis, Multi-Class Text Classifiation, Deep Learning, Convolutional Neural Network

1 *Introducción*

Actualmente, debido al aumento de redes sociales, blogs, foros de discusión y webs especializadas, existe un creciente volumen de datos de opinión registrados digitalmente y disponibles para su estudio (Liu, 2012). El análisis de sentimientos conforma el campo dedicado al análisis de las emociones, y opiniones de personas a través del Procesamiento del Lenguaje Natural (PLN).

Los sistemas dedicados a la minería de opiniones y al análisis de sentimientos, son ampliamente utilizados en ámbitos sociales y empresariales debido al fuerte papel que representa la opinión de las personas sobre un determinado producto y servicio. En el ámbito de la medicina, se puede extraer información en lo que refiere al estado de salud y condición de los pacientes, así como información sobre los tratamientos que reciben (Denecke y Deng, 2015).

Dentro de este dominio, el análisis y vigilancia de los diferentes fármacos, una vez aprobados en el mercado, es clave en lo que respecta a la mejora de la seguridad. Durante los ensayos clínicos, la probabilidad de detectar todos los efectos adversos a fármacos es muy baja, ya que los ensayos clínicos tie-

ISSN 1135-5948. DOI 10.26342/2019-63-1 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

nen una duración limitada y se realizan sobre pequeños grupos de la población. El efecto de un fármaco puede variar en función de un elevado número de factores, tales como la edad, condición previa del paciente, enfermedad, tratamiento con otros fármacos, etc. El análisis automático de la opinión que los pacientes tienen sobre determinados fármacos no sólo puede proporcionar una visión sobre el grado de satisfacción que los pacientes tienen sobre un determinado fármaco, sino que también puede dar la pista sobre fármacos con un elevado número de efectos adversos.

El objetivo de este trabajo es predecir el grado de satisfacción que los pacientes tienen sobre determinados fármacos, en base a sus comentarios. Concretamente, realizamos una serie de experimentos sobre una colección de comentarios extraídos de la web Drugs.com (Gräßer et al., 2018). El corpus contiene una serie de comentarios y opiniones escritas en inglés, por usuarios no expertos, sobre diferentes fármacos. Además de los comentarios, los usuarios dan una valoración relativa a la efectividad y los posibles efectos adversos del fármaco. La valoración puede tomar valores entre 1 (valoración más negativa) y 10 (valoración más positiva). La representación de las diferentes clases suponen un problema de multiclase desbalanceado donde aquellas más representadas son las más polarizadas, siendo las clases 10, 9 y 1 las que presentan un mayor número de instancias.

Para esta tarea, proponemos un enfoque de aprendizaje profundo supervisado, en concreto, una red neuronal convencional (CNN) (Collobert y Weston, 2008), cuya entrada es un modelo pre-entrenado de word embeddings (Pyysalo et al., 2013), que nos permite representar los comentarios. Además, realizamos una comparativa entre dicho enfoque con un modelo Máquinas de Soporte Vectorial (SVM) (Joachims, 1998). En este caso, los textos son representados utilizando un modelo de bolsa de palabras con la frecuencia inversa de las palabras.

El documento está estructurado de la siguiente manera: la sección 2 presenta brevemente una serie de estudios relacionados con en análisis sentimental en comentarios sobre fármacos y medicamentos. En la sección 3, se describen la colección de textos utilizada para llevar a cabo el estudio y la metodología empleada. En las secciones 4 y 5, se muestran los resultado obtenidos y una discusión sobre los mismos. Finalmente, en la sección 6 se presentan las conclusiones y trabajos futuros.

2 Estado de la cuestión

Los principales trabajos en análisis de sentimiento aplicado al dominio de salud en fármaco-vigilancia resultan útiles a la hora de seleccionar que fármacos deben ser vigilados para identificar posibles efectos adversos. Dichos trabajos se han centrado en dos grandes tareas: (I) clasificación de comentarios y (II) extracción de aspectos de opiniones. En la clasificación de los comentarios se extrae el sentimiento u opinión global sobre el texto, en este caso, sobre los comentarios de fármacos (ejemplo: positivo, negativo, neutro). En la extracción de aspectos, se identifican las opiniones o sentimientos referidas a determinadas características o aspectos de la entidad (fármaco) sobre el que se realiza el comentario.

Los diferentes enfoques empleados para la clasificación de los comentarios pueden agruparse en enfoques de aprendizaje automático (supervisado o no supervisado), enfoques basados en el léxico y reglas o en enfoques híbridos (Pang y Lee, 2008).

En (Na et al., 2012), realizan una clasificación binaria de sentimientos (positivos y negativos) en comentarios de medicamentos extraídos de la pagina web DrugLib.com. En este trabajo, se aplicó un lexicón construido a partir del Subjetivity Lexicon (SL) (Wilson, Wiebe, y Hoffmann, 2005) y de SentiWordNet (SWN) (Baccianella, Esuli, y Sebastiani, 2010). Además, los autores también exploraron un enfoque basado en SVM, donde los textos habían sido representados utilizando el modelo de bolsa de palabras. Los experimentos muestran que el enfoque lingüístico obtiene mejores resultados (accuracy de 78 %), frente a un 73 % de accuracy obtenido por el modelo SVM.

En (Na y Kyaing, 2015), se emplea otro enfoque puramente lingüístico en clasificación multiclase de sentimientos (positivo, negativo y neutral) en comentarios extraídos de webmd.com. Los autores también utilizaron un lexicón extraído de SL y SWN, además de un árbol de dependencias, obtenido con la herramienta Stanford NLP library (De Marneffe, MacCartney, y Manning, 2006). En este artículo también se realiza una comparativa con modelos SVM y bolsas de palabras, obte-

niendo unos resultados de 69 % de accuracy en el enfoque lingüístico y 66 % de accuracy usando SVM.

En (Gopalakrishnan y Ramaswamy, 2017), se realiza una clasificación polarizada (positivo, negativo y neutral) en comentarios extraídos de la web askapatient.com. Los comentarios son preprocesados y representados con vectores de frecuencia inversa (Tf-idf), donde las diferentes características son unigramas, bigramas y trigramas. Los autores exploraron diferentes algoritmos de aprendizaje supervisado tales como SVM, Redes Neuronales Probabilísticas (PNN) y Redes Neuronales de Base Radial (RBFN), obteniendo los mejores resultados con los modelos RBFN, con una F1 en torno al 84-94 %, para cada una de las clases.

En (Yadav et al., 2018), abordan el problema de multiclasificación de comentarios sobre medicamentos. Las posibles clases son: medicamentos efectivos, ineficaces o con efectos adversos graves. El corpus es extraído de la web patient.info y los textos son preprocesados obteniendo unigramas, bigramas y trigramas. Además, cada comentario es asignado con una posible puntuación (positiva, negativa y neutral), obtenida empleando el lexicón SWN. En este artículo, se evalúan diferentes modelos tales como SVM, Random Forest, Perceptrón Multicapa (MLP) y CNN. El mejor resultado es obtenido por el modelo de CNN obteniendo un 82 % de macro-F1. En este modelo, los textos fueron representados utilizando el modelo pre-entrenado de word embeddings de Google news.

En el trabajo (Gräßer et al., 2018), realizan una predicción de la satisfacción de los pacientes (positivo, negativo y neutral) sobre una colección de comentarios extraídos de Drugs.com y Druglib.com. Los textos son representados utilizando un modelo de bolsa de bigramas y trigramas. Los autores aplican regresión logística y obtienen resultados del 92.24 % de accuracy sobre los comentarios de Drugs.com y del 69.88 % sobre Druglib.com.

3 Metodología

3.1 Dataset

El corpus consiste en una colección de comentarios escritos en inglés, extraídos de forma automática por (Gräßer et al., 2018) de la web Drugs.com, que proporciona información sobre medicamentos tanto a pacientes como a profesionales de la salud.

El contenido del corpus consiste en comentarios de pacientes sobre un fármaco específico, así como una puntación del 1 al 10, que refleja el grado de satisfacción del paciente con el fármaco. A continuación, se describe con más detalle la información que proporciona el corpus para cada uno de los comentarios:

- Nombre del fármaco.
- Condición o estado en el que se encuentra el paciente; por ejemplo, *insomnio, depresión, trastorno bipolar, etc.*
- Texto del comentario del paciente sobre el fármaco.
- Puntuación del 1 al 10 del grado de satisfacción del paciente sobre el fármaco.
- Fecha de en la que se registró el comentario.
- Número de usuarios, consumidores y profesionales, que marcaron el comentario como útil.

Por ejemplo, un paciente con fibrilación auricular publicó el siguiente comentario: "*Only on it for 8 days. After 5 days started having shortness of breath, muscle spasms in upper back, pounding heart rate, fatigue, stiffness in neck and face*". Dicho comentario hace referencia al fármaco Flecainide y describe una serie de efectos adversos que éste le ha producido a los pocos días. La puntuación del fármaco proporcionada por el paciente es negativa con un valor de 1.

El corpus empleado está formado por una total de 215.063 comentarios. El corpus ha sido dividido en datasets de entrenamiento y test, manteniendo en ambos la proporción de las clases mediante un muestreo aleatorio estratificado. Los datasets de entrenamiento y test se encuentran en una proporción de 75 % y 25 %, respectivamente. Adicionalmente, un 15 % del total del dataset de entrenamiento ha sido reservado como dataset de validación, que nos permite aprender los mejores hiperparámetros de los distintos modelos.

La distribución de los diferentes comentarios en función de su clase (grado de satisfacción) puede ser observada en la figura 1. Exista una fuerte desproporción del número de instancias entre las distintas clases, predominando aquellas con valoraciones más polarizadas.

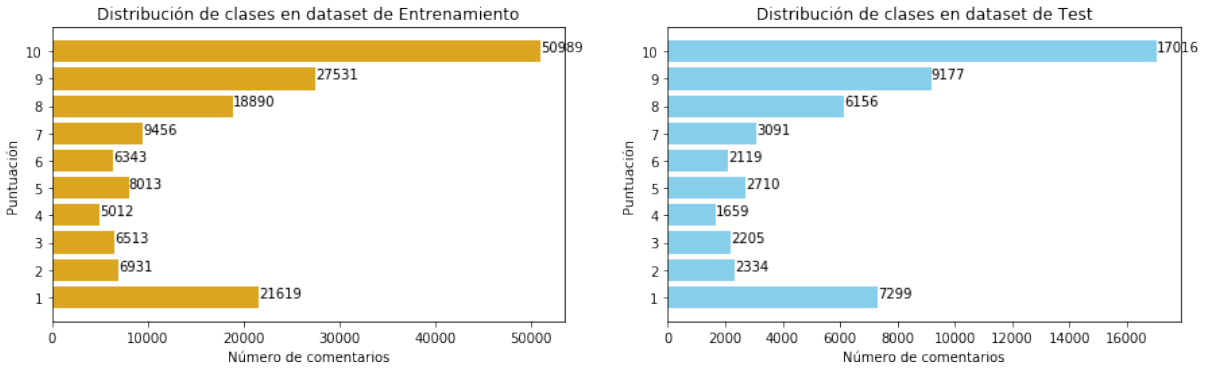


Figura 1: Distribución de las clases en los datasets de Entrenamiento y Test

3.2 Sistema

La finalidad de esta tarea consiste en resolver un problema de multiclasificación de textos. En particular, el objetivo es inferir un modelo capaz de predecir la puntuación (clase) o grado de satisfacción para un comentario concreto.

Los enfoques de aprendizaje profundo han demostrado su capacidad de aprender las características más relevantes para la tarea de PLN a resolver. En concreto, las redes convolucionales (CNN) han demostrado ser muy efectivas a la hora de realizar tareas de clasificación de textos. Dicha eficacia reside en su capacidad para extraer patrones representativos que describen el texto en forma de n-gramas (Kim, 2014), (Wang et al., 2017). Por este motivo, en este trabajo proponemos una arquitectura basada en CNN.

Para el preprocesado de los comentarios, el texto es tokenizado y las expresiones numéricas son cegadas, obteniendo un vocabulario de 49.339 tokens. Las CNN requieren que su entrada siempre tenga la misma longitud. Esto último es debido a la necesidad de definir unos filtros convolucionales donde los pesos se modificarán teniendo en cuenta un tamaño de ventana concreto para todos y cada uno de los textos (comentarios). Sin embargo, debido a la longitud variable de los comentarios, es necesario realizar un padding o rellenado de tokens, que no aporten información (zero-padding) al final de cada uno de los textos para igualar su longitud.

Basándonos en la función de distribución acumulada sobre la longitud de los comentarios, el 90 % de los textos preprocesados tienen una longitud menor o igual a 250 tokens. Sin embargo, como el tamaño máximo es relativamente pequeño (2.006 tokens), hemos decidido definir el tamaño de la entrada como este tamaño máximo, y evitando así realizar truncado.

La arquitectura propuesta consta de una capa de entrada donde cada uno de los textos, ya procesados, son representados como una matriz de embeddings gracias a un modelo pre-entrenado de word embeddings (Pyysalo et al., 2013). Este modelo de word embeddings fue entrenado sobre una colección de textos de PubMed, PMC y de la Wikipedia en inglés (versión de 2013), utilizando word2vec (Mikolov et al., 2013). El modelo contiene un vocabulario de 5.443.656 de palabras. La dimensión de los word embeddings es 200.

A continuación, la arquitectura cuenta con una capa de convolución, donde diferentes filtros o sub-matrices operan deslizándose a lo largo de la matriz de entrada, produciendo como salida un mapeo de características de los textos. Para obtener los mejores hiperparámetros de la red, aplicamos grid-search, obteniendo 64 filtros con un tamaño de ventana de 2, 3 y 5 vectores de palabras. Como función de activación empleamos una unidad de rectificación linear (ReLU) (Nair y Hinton, 2010) para evitar problemas de desvanecimiento de gradiente. (Ide y Kurita, 2017).

Una vez obtenido el mapeo de características, se emplea una capa de agrupación (pooling) para obtener aquellas características más relevantes a lo largo de las salidas de los diferentes filtros. Existen diferentes métodos de agrupamiento, tales como la media, el máximo de las entradas de la capa o modelos de atención (Boureau et al., 2010), (Er et al., 2016). Para nuestra experimentación, empleamos un agrupamiento basado en el máximo de los valores de cada convolución. Dicho método es empleado con el fin evitar que

la operación de padding pueda afectar negativamente a la representación de los textos (Suárez-Paniagua y Segura-Bedmar, 2018).

Finalmente, las salidas de las capas de agrupamiento son concatenadas y enviadas a una capa de perceptrones totalmente conectados. En esta capa, se utiliza la función de activación ReLU con el objetivo de mejorar la predicción de las clases. En este punto, los diferentes vectores que representan cada uno de los textos, son conectados con una capa Softmax, que finalmente predice la puntuación del 1 al 10 para cada comentario. En la figura 2 podemos observar la arquitectura descrita junto con las diferentes capas de regulación (dropout) empleadas para evitar sobreajuste durante el entrenamiento.

4 Resultados

En la evaluación de nuestros sistemas, hemos utilizado las métricas estándar para las tareas de clasificación de textos: precisión, recall y F1 (Frakes y Baeza-Yates, 1992). Dado que nos encontramos ante un problema de multi-clasificación, donde las diferentes clases están desbalanceadas, empleamos las métricas de micro y macro-averages. Con macro-average, mostramos la media de los valores obtenidos para cada clase de forma independiente, mientras que en micro-average se agrega la contribución de cada una de las clases.

Como sistema baseline, consideramos un modelo de Support Vector Machines (SVM) con kernel lineal. En este sistema, los comentarios son representado el modelo extendido de bolsa de palabras con tf-idf. Los clasificadores SVM han demostrado buenos resultados en problemas de clasificación de textos, ya que suelen ser linealmente separables (Joachims, 1998).

Los experimentos con arquitecturas de CNN han sido realizados empleando la librería Keras con tensorflow. Los experimentos realizados SVM han sido realizados empleando la librería scikit-learn de Python.

En el caso de la CNN, además se utilizó un optimizador ADAM (Kingma y Ba, 2014), con un learning rate de 0.001, un batch size de 500 comentarios durante 200 epoch y categorical-crossentropy como función de pérdida.

En el caso del clasificador SVM se empleó un parametro de penalización C=10 con squared hinge como función de pérdida.

Para la configuración de los hiper-parámetros de los diferentes clasificadores empleamos grid search. Los diferentes modelos se han evaluado empleando el dataset de validación.

Label	**Precision**	**Recall**	**F1-score**
1	0.5987	0.7282	0.6571
2	0.4562	0.3303	0.3832
3	0.4211	0.3156	0.3608
4	0.4135	0.3140	0.3570
5	0.3964	0.3030	0.3434
6	0.4017	0.2728	0.3249
7	0.3847	0.2705	0.3176
8	0.3915	0.3236	0.3543
9	0.4302	0.3655	0.3952
10	0.6014	0.7675	0.6744
Micro	0.5197	0.5197	0.5197
Macro	0.4495	0.3991	0.4168

Tabla 1: Baseline. Resultados empleando tf-idf y LinearSVM

La tabla 1 muestra los resultados del sistema baseline. los mejores resultados se obtienen en aquellas clases con mayor representación y más polarizadas, obteniendo un 67.44 % y 65.71 % de F1 para las clases 10 y 1 respectivamente. De igual forma, los peores resultados son aquellos obtenidos para las clases menos representadas, llegando a obtener un 32.49 % de F1 para la clase 6 y un 31.76 % para la clase número 7.

Las tablas 2 y 3 muestran los resultados obtenidos por el modelo CNN. En la tabla 2 se muestran los resultados obtenidos al mantener los word embedding estáticos en el entrenamiento del modelo CNN. Mientras, en la tabla 3, podemos observar los resultados obtenidos al permitir que los valores de los word embeddings sean ajustados durante el entrenamiento de la red.

Como podemos comprobar en la tabla 2, los resultados siguen el mismo patrón que los obtenidos con el modelo SVM. Los mejores resultados son aquellos obtenidos para las clases mayoritarias (68.07 % y 64.74 %). No obstante, los resultados de las clases minoritarias son incluso más bajos que los proporcionados por el modelo SVM.

Si comparamos el modelo SVM y el modelo CNN con word embeddings estáticos descritos hasta el momento, el modelo SVM obtiene un 51.97 % de micro-F1, mientras que el modelo CNN obtiene un 47.50 % de micro-

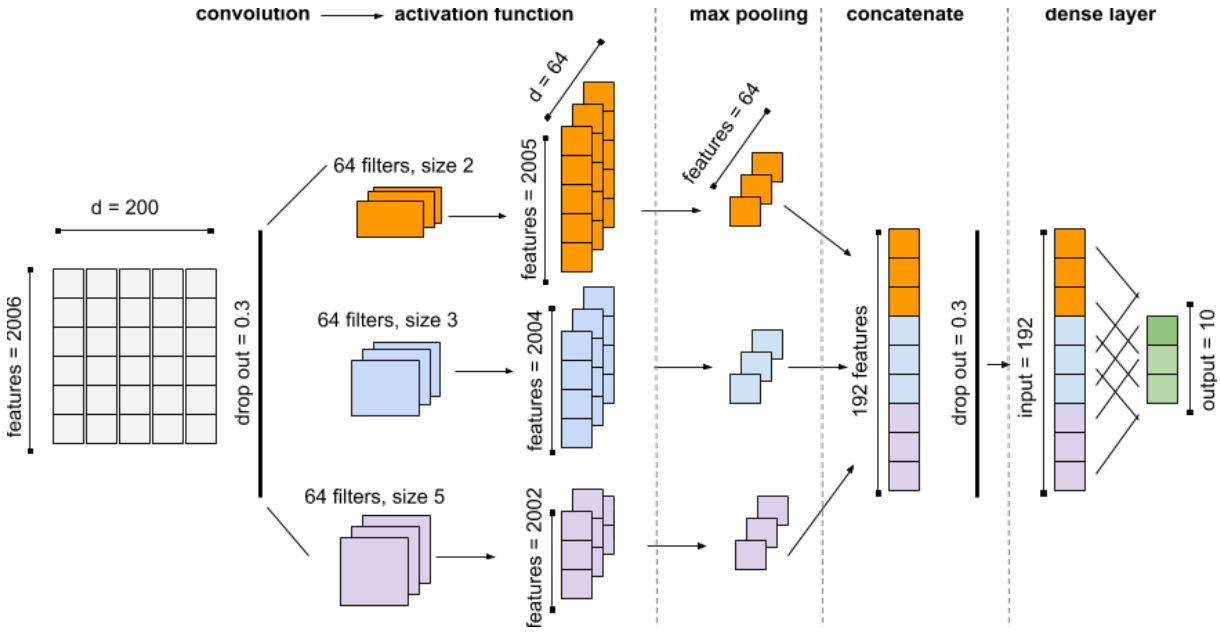


Figura 2: Arquitectura de nuestro sistema

Label	Precision	Recall	F1-score
1	0.5339	0.8223	0.6474
2	0.2690	0.1195	0.1655
3	0.2592	0.1247	0.1684
4	0.2416	0.0820	0.1224
5	0.2473	0.1524	0.1886
6	0.2047	0.0897	0.1247
7	0.2625	0.0984	0.1431
8	0.2815	0.2695	0.2754
9	0.3737	0.2528	0.3016
10	0.5816	0.8205	0.6807
Micro	0.4750	0.4750	0.4750
Macro	0.3255	0.2832	0.2818

Tabla 2: Resultados con CNN utilizando word embeddings estáticos (no entrenables)

F1. Estos resultados pueden deberse a que las características extraídas de la matriz de embeddings por los filtros convolucionales no fueran suficientes a la hora de discriminar entre una clase u otra.

Sin embargo, el modelo CNN con word embeddings no estáticos (tabla 3) mejora significativamente los resultados de los modelos anteriores. En este caso, los resultados obtenidos por las clases mayoritarias continúan siendo superiores al resto clases. No obstante, existe una notable mejoría en lo que respecta a las clases minoritarias, obteniendo finalmente una micro-F1 de 66.72 % y una macro-F1 de 61.81 % sobre el dataset de test.

Label	Precision	Recall	F1-score
1	0.7119	0.7556	0.7331
2	0.6227	0.5771	0.5991
3	0.6254	0.5664	0.5945
4	0.6451	0.5413	0.5887
5	0.6035	0.5572	0.5794
6	0.6444	0.5139	0.5718
7	0.5903	0.5290	0.5579
8	0.5822	0.5770	0.5796
9	0.6026	0.6214	0.6119
10	0.7454	0.7858	0.7651
Micro	0.6672	0.6672	0.6672
Macro	0.6374	0.6025	0.6181

Tabla 3: Resultados con CNN utilizando word embeddings entrenables

	Micro	Macro
SVM(Tf-idf)	0.5197	0.4168
CNN(w2v-estático)	0.4750	0.2818
CNN(w2v-entrenable)	**0.6672**	**0.6181**

Tabla 4: Resultados de micro y macro-F1 de los diferentes modelos. Los valores más altos están resaltados en negrita

5 Discusión

Como se ha mostrado anteriormente, nos encontramos ante un problema de multiclasificación con las clases desbalanceadas. Esto resulta problemático a la hora de clasificar correctamente aquellos comentarios cuya clase está escasamente representada en el dataset de entrenamiento.

La representación de los textos utilizando una bolsa de palabras o tf-idf produce vectores de elevadas dimensiones donde los datos están muy dispersos. Teniendo todo esto en cuenta, emplear un clasificador SVM podría proporcionar buenos resultados (Joachims, 1998).

Cuando comparamos el modelo generado con embeddings estáticos, nos encontramos con que los resultados son incluso inferiores a los proporcionados por el clasificador SVM. A pesar de que el modelo de word embeddings incluya un 82.04 % de las palabras del corpus, las características extraídas por los filtros convolucionales no parecen ser lo suficientemente discriminatorias como para alcanzar los resultados del modelo SVM.

No obstante, al permitir ajustar los word embeddings durante el entrenamiento de la red, podemos comprobar que una arquitectura sencilla de CNN es capaz de alcanzar una mejoría significativa, incluso en las clases menos representadas. En la tabla 4, vemos como el clasificador CNN nos proporciona un 66.72 % de micro-F1 frente a un 51.97 % obtenido por el clasificador SVM. Al realizar la media entre sus F1-score por igual sin ninguna ponderación, obtenemos un 61.81 % de macro-F1 frente al 41.68 % del clasificador SVM, corroborando que el modelo CNN mejora el desempeño a la hora de predecir las clases menos representativas.

6 Conclusión

Debido al creciente volumen de datos disponibles en internet, cada vez un mayor número de pacientes buscan opiniones sobre la eficacia y los posibles efectos de sus tratamientos. Por otro lado, el seguimiento y análisis de este tipo de opiniones sobre los fármacos, aporta una valiosa información complementaria a la hora de detectar posibles efectos adversos, que no fueron detectados durante los ensayos clínicos.

Mientras que la mayoría de los trabajos que han investigado en análisis de sentimiento en el dominio de salud, se han centrado en la clasificación de comentarios en dos o tres clases (positiva, negativa y neutra), en nuestro trabajo abordamos una tarea más complicada al tratar de clasificar los comentarios en 10 clases distintas (cada una de ella refleja el grado de satisfacción del paciente con respecto al fármaco). Además, otro reto importante es que debido al elevado número de clases, muchas clases están poco representadas. Nuestro mejores resultados son obtenidos con el modelo CNN (61.81 % de macro-F1), frente al enfoque tradicional de SVM con bolsa de palabras tf-idf (41.68 % de macro-F1).

Como trabajo futuro, exploraremos otras arquitecturas de aprendizaje profundo, tales como CNN con un mayor número de capas convolucionales, redes recurrentes y arquitecturas híbridas. Además, integraremos información semántica mediante el uso de modelos de embeddings de conceptos. Para abordar el problema de desbalanceo de datos, aplicaremos técnicas de sampling. También nos planteamos reducir el número de clases de 10 a 3 (positivo, negativo y neutro).

Agradecimientos

Este trabajo ha sido financiado por el Programa de Investigación del Ministerio de Economía y Competitividad del Gobierno de España (proyecto DeepEMR TIN2017-87548-C2-1-R).

Bibliografía

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. En *Lrec*, volumen 10, páginas 2200–2204.

Boureau, Y.-L., F. Bach, Y. LeCun, y J. Ponce. 2010. Learning mid-level features for recognition.

Collobert, R. y J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. En *Proceedings of the 25th international conference on Machine learning*, páginas 160–167. ACM.

De Marneffe, M.-C., B. MacCartney, y C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy,

Mayo. European Language Resources Association (ELRA).

Denecke, K. y Y. Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17 – 27.

Er, M. J., Y. Zhang, N. Wang, y M. Pratama. 2016. Attention pooling-based convolutional neural network for sentence modelling. *Information Sciences*, 373:388–403.

Frakes, W. B. y R. Baeza-Yates. 1992. *Information retrieval: Data structures & algorithms*, volumen 331. Prentice Hall Englewood Cliffs, NJ.

Gopalakrishnan, V. y C. Ramaswamy. 2017. Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of applied research and technology*, 15(4):311–319.

Gräßer, F., S. Kallumadi, H. Malberg, y S. Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. En *Proceedings of the 2018 International Conference on Digital Health*, páginas 121–125. ACM.

Ide, H. y T. Kurita. 2017. Improvement of learning for cnn with relu activation by sparse regularization. En *2017 International Joint Conference on Neural Networks (IJCNN)*, páginas 2684–2691. IEEE.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. En *European conference on machine learning*, páginas 137–142. Springer.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. y J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality.

Na, J. y W. Y. M. Kyaing. 2015. Sentiment analysis of user-generated content on drug review websites. *Journal of Information Science Theory and Practice*, 3(1):6–23.

Na, J.-C., W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, y Y.-L. Theng. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. En *International conference on asian digital libraries*, páginas 189–198. Springer.

Nair, V. y G. E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. En *Proceedings of the 27th international conference on machine learning (ICML-10)*, páginas 807–814.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pyysalo, S., F. Ginter, H. Moen, T. Salakoski, y S. Ananiadou. 2013. Distributional semantics resources for biomedical text processing.

Suárez-Paniagua, V. y I. Segura-Bedmar. 2018. Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction. *BMC bioinformatics*, 19(8):209.

Wang, J., Z. Wang, D. Zhang, y J. Yan. 2017. Combining knowledge with deep convolutional neural networks for short text classification. En *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, páginas 2915–2921.

Wilson, T., J. Wiebe, y P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Yadav, S., A. Ekbal, S. Saha, y P. Bhattacharyya. 2018. Medical sentiment analysis using social media: Towards building a patient assisted system. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 23-30 recibido 28-03-2019 revisado 14-05-2019 aceptado 17-05-2019

On the Effect of Word Order on Cross-lingual Sentiment Analysis

Sobre el efecto del orden de las palabras en el análisis de sentimiento crosslingüe

Àlex R. Atrio[1], Toni Badia[2], Jeremy Barnes[3]

[1] HEIG-VD & EPFL
[2]Universitat Pompeu Fabra
[3]University of Oslo
[1]alejandro.ramirezatrio@heig-vd.ch, [2]toni.badia@upf.edu, [3]jeremycb@ifi.uio.no

Abstract: Current state-of-the-art models for sentiment analysis make use of word order either explicitly by pre-training on a language modeling objective or implicitly by using recurrent neural networks (Rnns) or convolutional networks (Cnns). This is a problem for cross-lingual models that use bilingual embeddings as features, as the difference in word order between source and target languages is not resolved. In this work, we explore reordering as a pre-processing step for sentence-level cross-lingual sentiment classification with two language combinations (English-Spanish, English-Catalan). We find that while reordering helps both models, Cnns are more sensitive to local reorderings, while global reordering benefits Rnns.
Keywords: sentiment analysis, cross-lingual, reordering

Resumen: Los modelos de análisis de sentimiento que actualmente representan el estado del arte utilizan el orden de las palabras, ya sea explícitamente al preentrenar con un objetivo de modelización del lenguaje, ya sea implícitamente al recurrir a redes neuronales recurrentes (RNR) o convolucionales (RNC). Esto es un problema para los acercamientos crosslingües que emplean vectores bilingües para entrenar, ya que la diferencia del orden de las palabras entre la lengua de origen y la de destino no se resuelve. En este trabajo, exploramos el reordenamiento de las palabras como etapa de procesamiento previa para la clasificación de sentimiento crosslingüe a nivel de frase, con dos combinaciones de idiomas (Inglés-Castellano, Inglés-Catalán). Descubrimos que aunque el reordenamiento ayuda a los dos modelos, los RNC son más sensibles al reordenamiento local, mientras un reordenamiento global beneficia a los RNR.
Palabras clave: análisis de sentimiento, crosslingüe, reordenamiento

1 *Introduction*

Cross-lingual Sentiment Analysis (CLSA) exploits resources, *e. g.* labeled data of a high-resource language, to train a sentiment classifier for low-resource languages. This approach is useful when a target language lacks plentiful labeled data, particularly for specific domains. Machine Translation (MT) is often used to bridge the gap between languages (Banea et al., 2008; Balahur and Turchi, 2014), but requires abundant parallel data, which may be difficult to find for some low-resource languages. Approaches that use bilingual distributional representations, in contrast, have proven competitive while requiring less parallel data (Chen et al., 2016; Barnes, Klinger, and Schulte im Walde, 2018).

Recently, sentiment classifiers pre-trained on a language modeling task have lead to state-of-the-art results (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2018). This improvement suggests that sentiment analysis benefits from learning word order and fine-grained relationships between tokens, which can be gleaned from unlabeled data. These approaches, however, have only been applied in a monolingual setting and it is not clear how the difference in word orders would affect them in a cross-lingual setup. In this work, we perform an analysis of the effect of word order on cross-lingual sentiment classifiers that use bilingual embeddings as features. We show that these models are sensitive to

ISSN 1135-5948. DOI 10.26342/2019-63-2 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

word order and benefit from pre-reordering the target-language test data so that it resembles the source-language word order.

2 Related Work

Cross-lingual Sentiment Analysis: Cross-lingual approaches to sentiment analysis attempt to leverage available sentiment annotations in a high-resource language for target languages which lack annotated data. This is especially important when the cost of annotating a high-quality sentiment dataset, such as the Stanford Sentiment Treebank (Socher et al., 2013), can be prohibitive (215,154 phrases, each annotated by 3 annotators, at 10 cents an annotation would be 64,546€!). Therefore, it is preferable to make use of those datasets that already exist. Although most approaches to cross-lingual sentiment analysis rely on Machine Translation (Banea et al., 2008; Balahur and Turchi, 2014; Klinger and Cimiano, 2015), this requires large amounts of parallel data, making it less helpful for under-resourced languages.

Bilingual word embeddings have enabled cross-lingual transfer with small amounts of parallel data (Artetxe, Labaka, and Agirre, 2017; Lample et al., 2018b) or even none at all (Lample et al., 2018a; Artetxe, Labaka, and Agirre, 2018a), and are now used as features for state-of-the-art document-level (Chen et al., 2016), sentence-level (Barnes, Klinger, and Schulte im Walde, 2018), and targeted (Hangya et al., 2018) cross-lingual sentiment analysis approaches. The objective of bilingual embeddings is to learn a shared vector space in which translation pairs have similar vector representations. This benefits under-resourced languages as a sentiment classification model trained on the source-language can be applied directly to target-language data, without the need to translate it.

Word Order in Sentiment Analysis: Pre-training sentiment classifiers with a language-modeling task represents a successful transfer learning method. Peters et al. (2018) learn to create contextualized embeddings by training a character-level convolutional network to predict the next word in a sequence. Similarly, Howard and Ruder (2018) introduce techniques that improve the fine-tuning of the base language-model. Likewise, Devlin et al. (2018) introduce a self-attention network and adjust the language modeling task to a cloze task, where they predict missing words in a sentence, rather than the next word given a sequence. They then fine-tune their models on downstream tasks. These models that explicitly learn word order have led to state-of-the-art results on monolingual sentiment tasks.

		EN	ES	CA
Binary	+	1258	1216	682
	−	473	256	467
4-class	++	379	370	256
	+	879	846	426
	−	399	218	409
	−−	74	38	58
	Total	1731	1472	1149

Table 1: Statistics for the OpeNER English (EN) and Spanish (ES) as well as the Multi-Booked Catalan (CA) sentence-level datasets (Agerri et al., 2013; Barnes, Badia, and Lambert, 2018)

Word Reordering: Rule-based pre-reordering has a long tradition in Machine Translation (see Bisazza and Federico (2016) for a survey), where word order directly affects the quality of the final result. Reordering rules can be determined manually (Collins, Koehn, and Kucerova, 2005; Gojun and Fraser, 2012) or with data-driven approaches that either learn POS-tag based (Crego and Mariño, 2006a; Crego and Mariño, 2006b) or tree-based (Neubig, Watanabe, and Mori, 2012; Nakagawa, 2015) reordering rules. The advantage of POS-tag based rules is that they are simple to implement and do not require full parsing of the target-language sentences.

3 Methodology

3.1 Corpora and Datasets

At document-level, bag-of-words models are often expressive enough to give good results without relying on word order (Meng et al., 2012; Iyyer et al., 2015). But because we are interested in word-order effects in cross-lingual sentiment analysis, we require datasets that are annotated at a fine-grained level, *i. e.* sentence- or aspect-level.

For this reason, we use the English and Spanish OpeNER corpora of hotel reviews (Agerri et al., 2013) as well as the Catalan MultiBooked Dataset (Barnes, Badia, and

ORIGINAL	Único punto **negativo** el **ruido** que las ventanas de madera tan típicas de la zona *no* *consiguen* **aislar**
REORDERED	Único **negativo** punto el **ruido** que las ventanas de tan típicas madera de la *no* zona *consiguen* **aislar**
NOUN-ADJ	Único **negativo** punto el **ruido** que las ventanas de madera tan típicas de la zona *no* *consiguen* **aislar**
RANDOM	aislar madera *consiguen* típicas de el de zona las ventanas punto **negativo** Único la *no* **ruido** tan que
ONLY-LEXICON	UNK **negativo** UNK UNK **ruido** UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK **aislar**
NO-LEXICON	Único punto UNK el UNK que las ventanas de madera tan típicas de la zona *no* *consiguen* UNK
Translation	The only **negative** point the **noise** that the typical wooden windows in the area *don't manage* to **block**

Table 2: An example of a negative Spanish sentence (ORIGINAL) with the five reordering transformations applied, as well as its English translation. The **bold tokens** are words found in the sentiment lexicon, and the *italic words* are words that convey sentiment in this instance, but are not in the lexicon

Lambert, 2018). Statistics on the corpora are shown in Table 1. Each sentence is annotated for four classes of sentiment (strong positive, positive, negative, and strong negative). We use the English subset for training our classifiers and the Spanish and Catalan for testing the effects of word order on the target languages. Although these datasets are relatively small, they are all annotated similarly and are in-domain, which avoids problems with mapping labels or domain shifts.

3.2 Bilingual Word Embeddings

VecMap (Artetxe, Labaka, and Agirre, 2016; Artetxe, Labaka, and Agirre, 2017) creates bilingual embeddings by learning an orthogonal projection between two pre-computed monolingual vector spaces and requires only a small bilingual dictionary. We use the publicly available GoogleNews vectors for the English (available at `https://code.google.com/archive/p/word2vec/`), and for Spanish and Catalan we create skip-gram embeddings with 300 dimensions trained on Wikipedia data. The bilingual dictionaries are translated sentiment lexicons (Hu and Liu, 2004) with 5700 pairs for English – Spanish (5271 for English – Catalan).

3.3 Experimental Setup

In order to test whether a sentiment classifier trained on bilingual embeddings is sensitive to word order, we test classifiers on six versions of the target-language sentiment data, which we describe in the following section. An example of these six versions is shown in Table 2.

Original: We test the model on the original data with no changes in word order.

Reordered: A competing hypothesis is that a full pre-reordering of the target-side sequences will be more familiar to the sentiment classifier trained on English and therefore lead to better results. We implement POS-tag based rewrite rules (Crego and Mariño, 2006a; Crego and Mariño, 2006b), which are then applied to the target-language test data before testing.

Noun-Adj: Given that adjectives are important for sentiment analysis, we hypothesize that adjusting the order of nouns and adjectives should be beneficial if the classifier is learning source-language word order. Therefore, we implement a simple reordering which places Spanish and Catalan adjectives before, rather than after, the noun they modify.

Random: We randomly permutate the order of the target-language sentences. If the sentiment classification models take the target language word order into consideration, this should lead to poor results.

Only-Lexicon and No-Lexicon: Finally, we provide two baselines for clarification. The ONLY-LEXICON experiment removes all words which do not appear in the Hu & Liu sentiment lexicon (Hu and Liu, 2004). If our systems take word order into account, they should be affected negatively by this, as the resulting sentence does not resemble the normal word order. If, however, the models are relying on keywords, this will have little effect.

For the NO-LEXICON experiment, we remove all of the words in a phrase which appear in the sentiment lexicon. If the models are attending to sentiment keywords, this approach should lead to the worst performance.

Baselines: We perform additional experiments with monolingual and Machine Translation (MT)-based cross-lingual approaches. For the former, we use the Google API (available at `https://translate.google.com/`)

and translate the target-language data to English.

For both baseline setups, we only test the Random reordering, Only-Lexicon, and No-Lexicon approaches. Additionally, the monolingual setup is not comparable to the MT and cross-lingual versions, as we must use the target-language data for training, development, and testing (70%/10%/20%).

3.4 Models

To test our hypotheses, we compare three different classifiers: a Support Vector Machine (SVM) with Bag-of-Embeddings feature representations, a Convolutional Neural Network (Cnn) (dos Santos and Gatti, 2014; Severyn and Moschitti, 2015), and a Bidirectional Long Short-Term Memory Network (BiLstm) (Luong, Pham, and Manning, 2015). Each of these classifiers theoretically has an increasing reliance on word order. Note that we do not use the bilingual sentiment model (Barnes, Klinger, and Schulte im Walde, 2018), as it jointly learns both projection and classifier and cannot be used as input to the Cnn and BiLstm. Although the SVM does not take into account word order at all, it is a strong baseline for sentiment analysis (Kiritchenko et al., 2014). The Cnn considers only local word order, while the BiLstm relies on both local and long-distance dependencies.

For the neural models, we train five classifiers on five random seeds and report the mean and standard deviation of the macro F_1 score, while we only report the macro F_1 score of a single run for the SVM.

BiLstm We implement a single-layered BiLstm classifier with a 100-dimensional hidden layer, which passes the concatenation of the two final hidden states to a softmax layer for classification. The cross-lingual model is initialized with the pre-trained bilingual embeddings (monolingual embeddings for the monolingual and translation models), use dropout of 0.3 for regularization, and are trained for 30 epochs with a batch size of 32 using Adam as an optimizer. We choose the parameter for training epochs on the source-language development set and test this model on the target-language data.

Cnn The Cnn has a single convolutional layer with filters $\in \{3, 4, 5\}$ followed by a max-pooling layer of length 2. The pooled representation of the sentence is passed to a feed-forward layer and finally a softmax layer of size $\mathbb{R}^{|L|}$ where L is the set of labels. The optimization is the same as the BiLstm, with dropout applied after the feed-forward layer.

SVM Finally, we implement a baseline bag-of-embeddings SVM. For each sentence in the dataset, we create an averaged embedding representation $A = \frac{1}{n}(\sum_{i=1}^{n} e(t_i))$ where $e(t_i)$ is the embedding representation of the ith token in the sentence $S \in \{t_1, t_2, \ldots, t_n\}$ of length n. For the cross-lingual approaches we use the bilingual embeddings (monolingual embeddings for the monolingual and translation approaches) and tune the c parameter on the source-language development set.

4 Results

Table 3 shows the results of all experiments. Firstly, reordering the test data improves the results on all of the eight experiments (we do not consider SVM experiments to calculate improvements as they are invariant to word order). Specifically, the Reordered approach improves the BiLstm results the most on all experiments, while the simpler Noun-Adj flip is the best performing setup with Cnns. This indicates that local word reordering has more of an effect on Cnns, while the global reordering can be more helpful to BiLstms. While the improvements from reordering are often small (0.2 - 1.6 percentage points (ppt)), they are stable.

While it is the case that in both of the target languages Noun-Adj combinations can have a different meaning if the order is switched (for example "el amigo viejo" and "el viejo amigo"), the practical relevance of these order changes is minimal: in the Spanish dataset, of 978 occurrences of Noun-Adj, only 23 (2.35%) occur as well with a Adj-Noun order; in the Catalan dataset, of 745 occurrences of Noun-Adj, only 8 (1.07%) occur as well with a Adj-Noun order.

Random has a more substantial negative effect on monolingual models (an average decrease of 1.6 ppt for BiLstm and 3.0 ppt for Cnn) and MT-based models (1.6/4.3 ppt, respectively) than bilingual embedding models (0.8/1.1). This indicates that noise from the embedding projection renders it more difficult for models to use word order in the cross-lingual setup.

Additionally, Random has a larger effect on the Cnn (an average loss of 1.1 ppt) than on the BiLstm (0.8). This is likely because the Cnn relies on specific combinations of n-

			4-class			Binary		
			BiLstm	Cnn	SVM	BiLstm	Cnn	SVM
BWE	EN-ES	Original	33.3 (1.8)	35.4 (1.1)	34.9	64.9 (0.9)	60.0 (1.4)	66.6
		Reordered	**34.0 (1.6)**	35.6 (1.4)	34.9	**65.1 (1.3)**	60.1 (1.3)	66.6
		N-ADJ	**34.0 (1.8)**	**35.8 (1.2)**	34.9	65.0 (1.2)	**60.2 (1.4)**	66.6
		Random	33.2 (1.3)	35.3 (1.1)	34.9	63.9 (2.3)	58.8 (0.9)	66.6
		Only-Lexicon	28.2 (3.8)	26.9 (2.5)	30.7	57.6 (5.5)	34.2 (5.5)	53.0
		No-Lexicon	31.9 (1.6)	33.2 (1.4)	33.3	61.1	57.1 (2.8)	63.4
	EN-CA	Original	37.0 (1.4)	37.4 (1.5)	33.2	64.0 (1.1)	61.9 (6.8)	68.2
		Reordered	**37.8 (1.2)**	37.9 (1.5)	33.2	**65.6 (1.5)**	62.6 (5.8)	68.2
		N-ADJ	37.7 (1.5)	**38.1 (1.6)**	33.2	65.5 (1.5)	**62.8 (6.3)**	68.2
		Random	35.7 (1.0)	35.6 (1.5)	33.2	63.3 (0.8)	60.8 (5.5)	68.2
		Only-Lexicon	28.2 (1.8)	25.7 (3.2)	23.8	49.9 (4.3)	40.5 (6.7)	39.1
		No-Lexicon	35.9 (1.7)	34.3 (1.8)	31.2	61.7 (1.1)	58.1 (5.3)	63.1
MT	EN-ES	Original	**46.5 (1.2)**	**41.2 (3.7)**	44.6	**71.8 (1.1)**	**64.3 (1.6)**	70.7
		Random	46.0 (1.8)	38.9 (3.9)	44.6	71.0 (1.4)	62.2 (1.5)	70.7
		Only-Lexicon	32.9 (2.5)	28.2 (4.4)	36.2	63.0 (3.8)	44.6 (3.5)	51.9
		No-Lexicon	41.8 (0.7)	37.0 (3.0)	41.6	63.0 (1.1)	54.8 (2.6)	66.2
	EN-CA	Original	**51.5 (3.1)**	**44.1 (4.3)**	46.8	**79.9 (1.5)**	**72.8 (2.0)**	74.2
		Random	49.7 (1.4)	37.7 (3.6)	46.8	76.5 (2.2)	66.4 (1.4)	74.2
		Only-Lexicon	32.0 (2.7)	32.5 (4.0)	36.1	58.4 (7.8)	57.5 (2.9)	43.7
		No-Lexicon	48.4 (2.0)	40.9 (2.9)	46.2	75.6 (1.6)	65.6 (2.8)	70.4

Table 3: Macro F_1 results for all corpora and techniques. We denote the best performing bilingual embedding method per column with a **blue box**, the best MT method with a **pink box**. We do not denote bag-of-words SVM results, as they are invariant to word order

			4-class			Binary		
Monolingual	ES	Original	**43.2 (3.3)**	**36.2 (2.2)**	32.1	**68.5 (3.4)**	**64.8 (2.3)**	52.7
		Random	42.5 (2.6)	32.7 (1.8)	32.1	67.5 (4.2)	63.1 (2.7)	52.7
		Only-Lexicon	27.0 (0.5)	21.2 (4.5)	27.0	45.2 (0.0)	47.9 (3.9)	45.2
		No-Lexicon	37.9 (1.9)	34.3 (2.0)	30.3	64.7 (2.7)	65.0 (0.9)	51.8
	CA	Original	**48.6 (1.6)**	**46.2 (0.8)**	46.8	**77.1 (1.3)**	**76.4 (1.2)**	75.0
		Random	47.4 (1.9)	43.9 (3.0)	46.8	73.6 (1.3)	71.9 (1.9)	75.0
		Only-Lexicon	20.3 (2.8)	27.4 (3.2)	16.7	40.1 (1.5)	56.4 (5.2)	39.6
		No-Lexicon	47.5 (0.6)	45.8 (1.6)	45.8	75.0 (1.6)	74.5 (1.1)	74.8

Table 4: Macro F_1 results for all monolingual models. Although these results are not comparable to BWE or MT, as they are calculated on a smaller dataset, we provide them as a general reference to what results can be expected under monolingual conditions. We denote the best monolingual method with a **green box**

grams in order to correctly classify a sentence. If these are not present, the filters are not effective at classification.

Although they are not comparable (the test datasets have fewer examples), the monolingual models generally perform better than the cross-lingual versions, except for the SVM classifiers. The machine translation approaches perform better than the cross-lingual embedding methods.

The classification models display different trends across the setups. On the monolingual and machine translations setups, the BiLstm is the strongest model, followed by the Cnn and SVM (SVM and Cnn, respectively for machine translation). With bilingual embeddings, however, the SVM outperforms both the BiLstm and Cnn on the Spanish binary setup, while the Cnn is strongest on the multiclass. This shows that BiLstm displays a different behavior with bilingual embeddings.

The machine translation models perform well and surprisingly suffer less than monolingual models (an average decrease of 15.4 ppt for MT BiLstm and Cnn models vs. 20.6 for monolingual) from using only features from the sentiment lexicon (Only-Lexicon). This suggests that MT models rely more on these keywords while ignoring word order effects to a higher degree.

Finally, the No-Lexicon and Only-Lexicon baselines perform poorly, with

model	text	prediction
Original	relación calidad precio muy buena	negative
Reordered	relación muy buena calidad precio	positive
translation	very good quality price relationship	positive
Original	hotel perfecto	negative
Reordered	perfecto hotel	positive
translation	perfect hotel	positive
Original	el desayuno muy bueno .	negative
Reordered	el muy bueno desayuno .	positive
translation	the breakfeast (was) very good	positive
Original	gestión nefasta .	positive
Reordered	nefasta gestión .	negative
translation	terrible management	negative

Table 5: Examples where reordering improves results over original on binary English-Spanish setup with the BiLSTM classifier

Only-Lexicon often more than 20 ppt below the performance of Original. This is due largely to the low coverage of the sentiment lexicon used in this work, as many full sentences were completely unked (38% for Spanish, 43% for Catalan). This also explains the similar performances of Original and No-Lexicon.

5 Analysis

Reordering tends to help both the BiLstm and Cnn models with shorter examples (less than eight tokens long) where adjective order can easily be changed to resemble English word order, such as the examples shown in Table 5. In longer instances (more than ten tokens), however, the reordering either introduces too much noise or does not affect the final prediction. The current reordering models are therefore more adequate for sentiment tasks that deal with shorter texts, such as aspect- or sentence-level, rather than document-level sentiment analysis.

6 Conclusion and Future Work

In this work, we have shown that neural networks that rely on bilingual embeddings as features are sensitive to differences in source- and target-language word order and subsequently benefit from reordering the target language test data. The gains, however, are still relatively small, which suggests that currently bilingual embeddings introduce too much noise for a classifier to generalize well to the target language.

Although our reordering approach does improve the neural models, these more expressive models are still outperformed by the linear SVM with bag-of-embeddings representations. This is likely a side effect of the noise introduced by the bilingual embeddings. At test time, the model receives as input embeddings that are *similar* but not necessarily the same as at training. In the future, it may be helpful to develop models which are more robust to this noise, or alternatively to use low-resource machine translation techniques (Artetxe et al., 2018; Lample et al., 2018a; Artetxe, Labaka, and Agirre, 2018b; Lample et al., 2018c)

Given that language modeling pre-training is beneficial for state-of-the-art results in monolingual sentiment analysis, it is important to realize that cross-lingual models based on bilingual word embeddings do not currently benefit from word order learned in the source language. In the future, we would like to pre-train bilingual language models for cross-lingual sentiment analysis.

References

Agerri, R., M. Cuadros, S. Gaines, and G. Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51(Septiembre):215–218.

Artetxe, M., G. Labaka, and E. Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving mono-

lingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Artetxe, M., G. Labaka, and E. Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Artetxe, M., G. Labaka, and E. Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April.

Balahur, A. and M. Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Banea, C., R. Mihalcea, J. Wiebe, and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135.

Barnes, J., T. Badia, and P. Lambert. 2018. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018.

Barnes, J., R. Klinger, and S. Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493.

Bisazza, A. and M. Federico. 2016. A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena. *Computational Linguistics*, 42:163–205.

Chen, X., B. Athiwaratkun, Y. Sun, K. Q. Weinberger, and C. Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614.

Collins, M., P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540.

Crego, J. M. and J. B. Mariño. 2006a. Improving statistical mt by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Crego, J. M. and J. B. Mariño. 2006b. Integration of pos tag-based source reordering into smt decoding by an extended search graph. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 29–36. Cambridge.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

dos Santos, C. N. and M. Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland, August.

Gojun, A. and A. Fraser. 2012. Determining the placement of german verbs in english–to–german smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 726–735.

Hangya, V., F. Braune, A. Fraser, and H. Schütze. 2018. Two methods for do-

main adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820.

Howard, J. and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Hu, M. and B. Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daume III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China.

Kiritchenko, S., X. Zhu, C. Cherry, and S. M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 437–442.

Klinger, R. and P. Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China, July.

Lample, G., A. Conneau, L. Denoyer, and M. Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November.

Luong, T., H. Pham, and C. D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Meng, X., F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea, July.

Nakagawa, T. 2015. Efficient top-down btg parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 208–218.

Neubig, G., T. Watanabe, and S. Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Severyn, A. and A. Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Traducción Automática

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 33-40 recibido 30-03-2019 revisado 03-05-2019 aceptado 15-05-2019

Adapting NMT to caption translation in Wikimedia Commons for low-resource languages

Adaptando NMT a la traducción de pies de imagen en Wikimedia Commons para idiomas con pocos recursos

Alberto Poncelas[1], Kepa Sarasola[2], Meghan Dowling[1], Andy Way[1], Gorka Labaka[2], Iñaki Alegria[2]
[1]ADAPT Centre, School of Computing, Dublin City University
[2]Ixa Group Faculty of Informatics, (UPV/EHU)
kepa.sarasola@ehu.eus

Abstract: This paper presents a successful domain adaptation of a general neural machine translation (NMT) system using a bilingual corpus created with captions for images in Wikimedia Commons for the Spanish-Basque and English-Irish pairs.
Keywords: Machine Translation, Low-resource languages, Bilingual corpora, Language resources from Wikipedia

Resumen: Este artículo presenta una adaptación a dominio exitosa de un sistema de Traducción automática neuronal (NMT) utilizando un corpus bilingüe creado con los pies de imagen utilizados en Wikimedia Commons para los pares de idiomas español-euskera e inglés-irlandés.
Palabras clave: Traducción automática, Idiomas con recursos limitados, Corpus bilingüe, Recursos lingüísticos extraídos de Wikipedia

1 Introduction

Wikimedia (the umbrella organisation which includes Wikipedia, Wikidata, Commons, etc.) is one of the most valuable sources for the collection of low-resource language corpora. Due to most of the text-based content in Wikimedia being in majority languages such as English and Spanish, the development of a translation tool could help generate new content in these low-resource languages.

In this paper we try to take advantage of existing text in WikiMedia in order to improve translation quality for low-resource languages. Our hypothesis is that given a corpus created with available bilingual captions for images in Wikimedia Commons (in Spanish and Basque), a successful adaptation of a general NMT translation-system to this domain could improve the quality of the translation of new captions. We also hypothesise that additional semantic information extractable from the Wikipedia context in which the images are being used could also be help improve translation quality.

Using new corpora that we generate from Wikimedia Commons image captions, we demonstrate that significant increments in translation quality can be obtained. Moreover, additional small improvements were detected when using additional semantic tags extracted from Wikipedia or Wikidata.

Following on from this, we hypothesise that this process could be applied not only for Basque, but also to other low-resource languages too. To investigate this, we perform similar experiments for the English-Irish pair.

The aim of these experiments is that the newly improved MT systems created will help to increase the number of image captions for low-resource languages in Commons and over time the increased training corpus size will lead to the creation of an even better MT system.

The paper is organised as follows. Section 2 introduces related work on this subject. In Section 3, we describe in detail the process used to create the parallel corpora. Section 4 details the steps performed in this experiment: creation of the initial corpus, design and evaluation of the baseline and the domain-adapted NMT systems, and then the enrichment of the corpus with semantic tags and the final evaluation. Section 5 shows how this work is being continued with other language pairs, detailing a similar experiment that indicates the extensibility of this work to the English-Irish pair. Finally, in Section 6, we draw some conclusions and propose some lines for future work.

2 Related work

NMT is the dominant paradigm in the field as evidenced in particular by large translation

ISSN 1135-5948. DOI 10.26342/2019-63-3 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

providers turning to NMT for their production engines (Wu et al., 2016; Crego et al., 2016) and NMT systems achieving the best results in most cases on standard shared task datasets (Bojar et al., 2017).

2.1 Domain adaptation for NMT systems

Fine tuning is the conventional method of domain adaptation in NMT. This consists of using in-domain data for training the last epoch of a pre-built NMT model. Previous work (Luong and Manning, 2015; Poncelas, de Buy Wenniger, and Way, 2018b) has shown that using in-domain data for the last epochs of an NMT system leads to better results.

2.2 Previous results for the es-eu and en-ga pairs

The only reported results for Spanish-Basque (es-eu) using NMT are those described by Etchegoyhen et al. (2018). They report improvements over other methods using NMT and BLEU scores of around 20 and 23 points for two references of the same corpus. In previous work, Labaka et al. (2014) reported their results using statistical machine translation (SMT) and a hybrid model for the same language pair. Other works involving Basque includes the work of Poncelas, Way, and Sarasola (2018) in which Basque-to-English MT models were improved by fine-tuning with in-domain sentences (retrieved using a approximated target side (Poncelas, de Buy Wenniger, and Way, 2018a)).

In terms of the English-Irish (en-ga) pair, Dowling et al. (2015) have previously presented work on domain-specific SMT, and more recently, preliminary experiments involving en-ga NMT (Dowling et al., 2018).

2.3 Wikipedia and MT

In relation to resources for MT, Otero and López (2010) underline the value of Wikipedia as a multilingual source of comparable corpora. Nothman et al. (2013) generate multilingual named entity recognition by exploiting the text and structure of Wikipedia. This cross-lingual approach achieves up to 95% accuracy.

In Labaka, Alegria, and Sarasola (2016) a SMT system is enriched with extra in-domain parallel corpora compiled from parallel titles in Wikipedia. On the Spanish–English language pair they improve a baseline trained on the Europarl corpus by more than 2 BLEU points when translating texts from the Computer Science domain. In our previous work (Alegria et al., 2013) we described a collaboration framework that enables Wikipedia editors to generate new articles while helping the development of MT systems by providing post-editing logs. This collaboration framework was tested with editors of Basque Wikipedia.

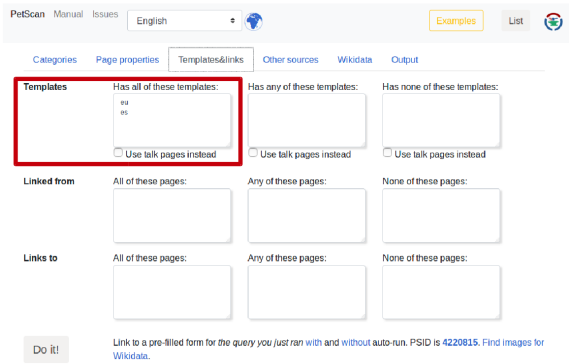


Figure 1: Searching with PetScan for images with templates 'es' (Spanish) and 'eu' (Basque).

3 *Creation of the new corpora*

Our first aim was to build a bilingual corpus, as big as possible, by collecting the captions of images in Wikimedia Commons[1]. We decide to use Petscan[2] for this purpose. The method we have developed can be used for any other pair of the 278 languages present in Wikipedia[3].

3.1 *CommonsCaptions* corpus

Petscan is a useful tool for searching for information in the three main Wikimedia services: Wikipedia (articles), Wikidata (linked data) and Wikimedia Commons (images). Besides a lot of other search capabilities, it also allows the user to look for images in Commons that contain a text description written in a given language, e.g.: searching the {{es}} template provides captions in Spanish. Figures 1 and 2 show two screenshots of the Petscan query that in April 2018 provided us with a list of the 6,876 images with captions in both Spanish and Basque. The list of files with both these templates had some false positives (since the templates defined to introduce text in Basque and Spanish may be used for other fields, since their use is not exclusive to captions).

As a first step our bilingual corpus was created solely through the collection of image cap-

[1] https://commons.wikimedia.org
[2] https://petscan.wmflabs.org
[3] https://stats.wikimedia.org/EN/Sitemap.htm

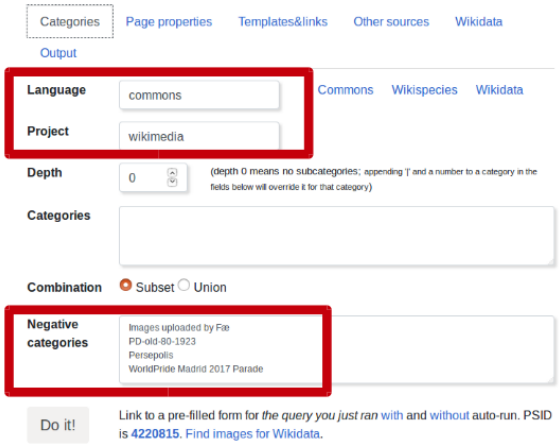


Figure 2: Searching in PetScan for images in Wikimedia Commons (the tab *Categories*).

tions, but eventually extra sentences could be obtained from text used in Commons categories and transclusions (the inclusion of an image into another Wikimedia document by cross-reference).

The caption descriptions detailed here are usually not full sentences. They accompany usages on projects such as Basque and Spanish Wikipedia and, for example, may only contain a single surname if in a biography. It is difficult to collect enough text for a low-resource language like Basque to be able to train an MT system.

However, many of the descriptions generated contained the same general short note which was not an accurate caption. We discarded such images by adding them to a *negative categories* list (see Figure 2). The final query[4] is represented with PSID=4220815, and provided us with 6,876 results.

Finally, after discarding 3,316 images with problematic captions (non-alphabetical captions, non-parallel captions, etc.), the final parallel corpus was composed of a total number of 3,560 captions. This *CommonsCaptions* corpus is publicly available online[5].

3.2 TaggedCommonsCaptions corpus

After building the *CommonsCaptions* corpus collect semantic information inherent to the Wikipedia context to test whether the inclusion of this knowledge could bring an improvement in the quality of the translations.

We used a simple approach: adding a tag to each of the images to distinguish 11

[4] https://petscan.wmflabs.org/?psid=4220815

[5] http://ixa.si.ehu.eus/node/11513?language=en

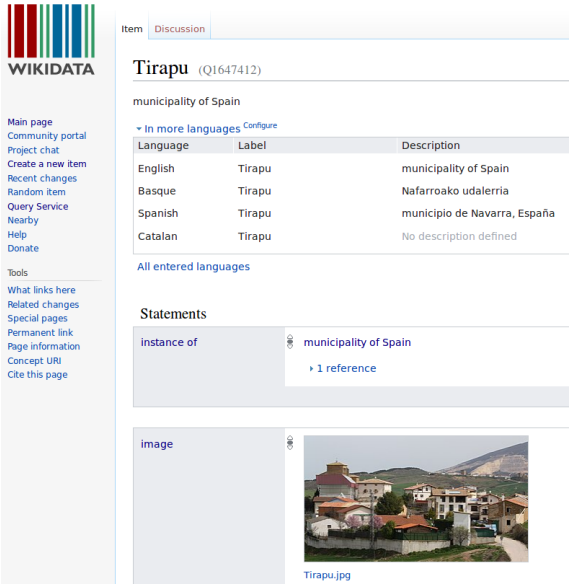


Figure 3: *Tirapu* concept in Wikidata.

different kinds of image (Person, Human-Group, Place/Location, Institution, Building, Animal/Plant, Event/Sport, History, Map/Icon, Culture, and Others). Our hypothesis was that these tags could be almost entirely automatically inferred from their Wikimedia context. A priori, extracting this new information from Wikimedia seemed easy: inferring the tag from the category assigned to the image in Wikimedia Commons. However, the task ultimately proved to be more difficult than expected. The problem is that Commons category system is a folk taxonomy. Not every relation between a category and its parent is an "is-a" relationship. For example, to know whether an image corresponds to a person a query to Petscan asking for those images whose category in Commons (with a maximum depth of 5) could be "People" ([Query PSID 4346397]), but unfortunately many of the 1,492 results are not those of persons. Using a maximum depth of 3 ([Query PSID 4346796]) deviation is not so wide, the results get only 248 results, but again many of them are not persons.

The alternative was using Wikidata instead, as its taxonomy is more reliable. We defined two reliable automatic ways to extract that information: (1) directly from Wikidata, and (2) indirectly via the Wikipedia pages where the image is used.

Directly from Wikidata: If the image is the one used to illustrate a concept in Wikidata, the desired semantic tag will be the value of the property "instance-of" for that concept in Wikidata. The coverage is 5.5% (196 images are tagged over 3,560 captions). For example, in Figure 3

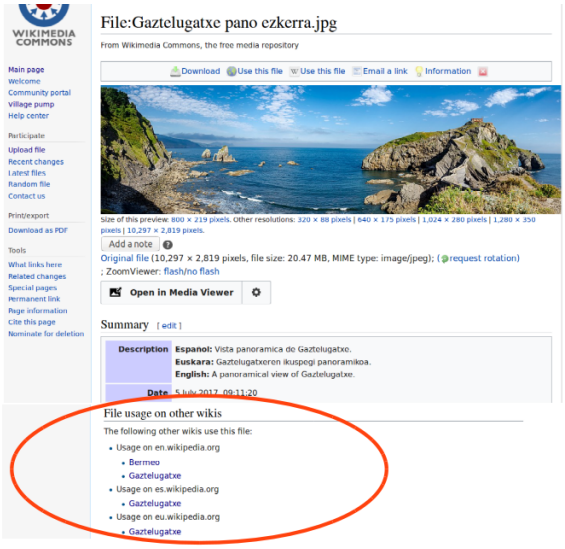


Figure 4: Image usage on other wikis

the caption of the image Tirapu.jpg is tagged as *Institution* as it is the image for the concept Q1647412, defined in Wikidata as an instance of "municipality of Spain" (institution).

Indirectly via the Wikipedia pages where the image is used: Firstly, given one image, we obtain all the titles of Wikipedia pages in Basque, Spanish and English where the image is used (Wikimedia Commons offers this information). Secondly, we extract the Wikidata identifier for each of those Wikipedia pages. Thirdly, among those concepts we select the one repeated in most languages (Basque, Spanish and/or English). Finally, we extract the "instance-of" value in Wikidata. For example, the image *Gaztelugatxe_pano_ezkerra.jpg* is not directly related to any concept in Wikidata, but looking at its "*File usage on other wikis*" information (see Figure 3.2), we can see that it is used in *Gaztelugatxe* and *Bermeo* articles on the English Wikipedia (Q1496690 and Q695444 concepts in Wikidata).

As the concept Q1496690 is the most representative, using "island" (its value for the "instance-of" property) we assign the tag "Place/Loc". The coverage of this second way of tagging is 45.6% of the images (1,623 over 3,560).

Implicit information in the name of the image: Complementarily, 33.4% of the tags could be obtained looking for 40 words inside the name of the image. For example, as the word 'ermita' ('hermitage' in English) is in the name of the image named *San_Esteban_ermita.jpg*, the image is tagged as 'Building'. The following are other alternative words to tag buildings: hotel, palace, alcázar, museo (museum), tomb, teatro (theatre), zubi (bridge), puente (bridge), architecture, eliza (church), ermita (hermitage) and geltokia (station). After accumulating the explained automatic criteria, a total of 64.9% of the images were tagged automatically, and therefore only 35.1% had to be selected by hand. Our *Tagged-CommonsCaptions* corpus is also publicly available online[6].

4 *Design and evaluation of the MT system*

Following the description of the parallel corpora related to the domain, and the current state-of-the-art MT technology, in this section we describe the method used to build the MT engines (baseline, domain-adapted system and semantically enriched system) and their evaluation.

4.1 Baseline system

The system created to be the state-of-the-art baseline for our experiments uses the optimal configuration for generic MT in Basque-Spanish, recently determined by Etchegoyhen et al. (2018). The NMT systems used in the experiments have been built using the OpenNMT-py toolkit (Klein et al., 2017). The models follow the attention-based encoder-decoder approach (Bahdanau, Cho, and Bengio, 2014). The encoder and the decoder consist of a 4-layer Recurrent Neural Network (RNN) with 800 LSTM (Hochreiter and Schmidhuber, 1997) hidden units, using a vocabulary size of 50,000 in each language. The models were trained for 13 epochs using Stochastic Gradient Descent with an initial learning rate of 1 and applying a learning decay of 0.7 after epoch 10.

The results of Etchegoyhen et al. (2018) show that the use of morphological processing is time-consuming and does not achieve a large improvement. Taking that into account, we decided not to perform any linguistically-motivated approach to segmentation. Hence, the pre-processing of the data is language-independent. The sentences were tokenised and truecased (proper capitalization recovered), and we applied Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016) (trained in both languages using 30,000 merge operations).

In Table 1 we find a summary of the datasets. To build representative translation models for the Basque-Spanish language pair the Elhuyar Corpus was used. It contains 2 million sentences

[6] http://ixa.si.ehu.eus/node/11513?language=en

	Elhuyar	Captions
train	2,061,863	3,000
dev	3,482	–
test	3,405	560

Table 1: Number of parallel sentences in the eu-es Elhuyar and *CommonsCaptions* datasets.

for both languages, collected from professional translations in different domains, and bilingual web pages (*train* row in Table 1). This is a reduced version of the corpus used by Etchegoyhen et al. (2018), where they also included an extra set of 807,222 sentences collected from comparable data in the news domain.

For building the models adapted for the image captions, we use the *CommonsCaptions* corpus (explained in Section 3.1). We use 3,000 sentences for training and the remaining 560 sentences as our *test* set. The development (*dev*) set used for the models was the same as the Elhuyar *dev* set.

		Elhuyar test	Captions test
es → eu	BLEU	17.61	19.68
	NIST	5.80	5.17
	TER	65.74	69.09
eu → es	BLEU	27.58	23.91
	NIST	7.52	4.94
	TER	56.69	60.27

Table 2: Results of the BLM models

In Table 2 we provide the results of the baselines, (*BLM models*) using several evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and TER (Snover et al., 2006). These scores indicate how good the outputs of the NMT systems are compared to a human-translated reference. For BLEU and NIST, the higher the score, the better the translation is estimated to be, while for TER goes, as it is an error rate, the lower the better.

4.2 Domain-adapted NMT system

The domain adaptation using the *CommonsCaptions* corpus was designed following the fine-tuning approach (Luong and Manning, 2015; Poncelas, de Buy Wenniger, and Way, 2018b).

In these experiments, the models explained in Section 4.1 are fine-tuned with the *CommonsCaptions* corpus. The models were trained for 12 epochs with the general Elhuyar corpus, with the last iteration trained only with the sentences from the domain corpus (*CommonsCaptions*). In total there were 13 epochs, the same as for the baseline models.

Table 3 summarises the results of the evaluation of domain-adapted models (*ADPM models*). We can see that the in-domain system shows a big improvement with respect to the baseline *BLM* system in terms of BLEU and TER metrics. The improvement in BLEU of ADPM models with respect to BLM (Table 2) when evaluated in *CommonsCaptions* test ranges from 19.68 to 23.01 (relative improvement of 18%) for the es-eu pair and from 23.91 to 29.59 (relative improvement of 23%) for the alternative direction. TER also decreases (meaning an increment of quality) from 69.09 to 62.61 (relative improvement of 9%) for the es-eu pair and 60.27 to 54.08 (relative improvement of 10%) for the eu-es pair.

When using the NIST metric, we see a small drop in performance. The proposed reason for that is that *n*-gram statistics substantially differ in the general corpus and in the domain of captions. While BLEU simply calculates *n*-gram precision adding equal weight to each one, NIST also calculates how informative a particular *n*-gram is. That is to say, when a correct *n*-gram is found, the rarer that *n*-gram is, the more weight it will be given.

		Elhuyar test	Captions test
es → eu	BLEU	15.06	23.01
	NIST	4.94	4.92
	TER	67.92	62.61
eu → es	BLEU	25.41	29.59
	NIST	7.10	5.40
	TER	59.22	54.08

Table 3: Results of the ADPM models

4.3 Domain-adapted and "semantic sensitive" NMT system

A new experiment was designed to examine whether the rich semantic knowledge inherent to the Wikipedia context can improve the results.

The new system is exactly the same as the previous system described in Section 4.2, but the corpus is replaced with a new version that includes a "semantic tag" as complementary and separated information in both languages. For example, the parallel texts for the image Basurto_ospitalea_geltokia_1.jpg in the *CommonsCaptions* corpus were the following:
eu: *Basurtu-Ospitalea tren geltokia*
es: *Estación de tren Basurto-Hospital*

		Captions test (categorised)
es → eu	BLEU	23.26
	NIST	5.12
	TER	61.63
eu → es	BLEU	29.71
	NIST	5.58
	TER	53.65

Table 4: Results of the SMM models

		BLM	ADPM	SMM
es → eu	BLEU	19.68	23.01	23.26
	NIST	5.17	4.92	5.12
	TER	69.09	62.61	61.63
eu → es	BLEU	23.91	29.59	29.71
	NIST	4.94	5.40	5.58
	TER	60.27	54.04	53.65

Table 5: Results of the 3 models (BLM, ADPM and SMM) using Captions test for the es-eu pair.

and now the parallel texts in the *TaggedCommonsCaptions* corpus are the following:
eu: *Building* ||| *Basurtu-Ospitalea tren geltokia*
es: *Building* ||| *Estación de tren Basurto-Hospital*. As in Section 4.2, the new semantic sensitive models (*SMM models*) is trained for 12 epochs with the general Elhuyar corpus, and then once more with the tagged domain corpus in the 13th iteration.

Table 4 summarises the results of the evaluation of the eu-es and es-eu systems. Figures show an improvement in the three metrics, but unfortunately none of them are statistically significant (using Bootstrap Resampling (Koehn, 2004) at level p=0.01). Table 5 shows a summary of the results of the three systems when they are tested on the *CommonsCaptions* test. The improvement in BLEU ranges from 19.68 to 23.26 (relative improvement of 18%) for the es-eu pair and from 23.91 to 29.71 (relative improvement of 24%) for the alternative direction.

5 Experiments with other language pairs

As the methodology is language-independent we repeat the experiment for the English-Irish (en-ga) pair. We knew that the obtained corpus would be smaller (7,554 words), but taking into account the poor performance in this domain of the baseline en-ga bilingual corpus we wanted to test the contribution of the supplementary corpus.

We present a summary of the data used in Table 6. As baseline data we use a collection of parallel sentences in the public administration domain from diverse sources crawled from the Web [7] and provided by different organizations: Department of Culture Heritage and the Gaeltacht (DCHG), Digital Corpus of the European Parliament (DCEP), Directorate General for Translation (DGT-TM) and Conradh na Gaeilge (CnaG).

Using Petscan to retrieve images with captions in both English and Irish we obtained 2,738 images captions [8]. After performing a cleaning process similar to that described in Section 3.1 for the es-eu pair, the remaining 434 captions were included in the final corpus: 350 were used for training and 84 for testing as shown in Table 6.

	Irish data	Captions
train	108,796	350
dev	655	–
test	1,516	84

Table 6: Number of parallel sentences in the Irish and *CommonsCaptions* datasets.

In order to ensure that this work would be comparable to other studies on en-ga MT, we chose to use the same corpus used by Dowling et al. (2018) in their NMT experiments mentioned in Section 2. This data contains a mixture of crawled data, data from an Irish government department, data from EU institutions, as well as a small amount of data from other sources.

The configuration of the NMT models trained in the en-ga language pairs is the same as that used in en-eu models. In Table 7 we present the results of the models trained on the baseline data and the fine-tuned model. The results show that significant improvements (with $p = 0.01$) are also possible for this pair when using the caption corpus [9], even when the number of sentences is 10 times smaller.

6 Conclusions and Future Work

We have created three corpora (Spanish-Basque *CommonsCaptions*, Spanish-Basque *TaggedCommonsCaptions* and English–Irish *CommonsCaptions*) that are considerable contributions for two low-resource languages. These corpora have been extracted from open data in

[7] `http://www.citizensinformation.ie` and `https://www.teagasc.ie/` websites

[8] `https://petscan.wmflabs.org/?psid=4414745`

[9] `http://ixa.si.ehu.eus/node/11513?language=en`

		BLM	ADPM
en → ga	BLEU	15.24	19.61
	NIST	2.80	2.71
	TER	72.97	68.74
ga → en	BLEU	10.18	14.84
	NIST	2.36	3.05
	TER	74.02	66.86

Table 7: Results of the baseline and the adapted systems evaluated with *CommonsCaptions* test for the en-ga language pair.

Wikimedia. We have also demonstrated that a simple adaptation of a general NMT system to this domain produces significant improvements in translation quality. Supplementary small improvements were detected when using additional semantic tags extracted from the Wikipedia context where the images are being used. The new method to create bilingual corpora and to use it to adapt a general NMT system is language-independent and could be useful to supply Wikimedia Commons content in more languages in the future. We have tested it for the Spanish-Basque language pair and confirmed its utility in the English-Irish pair.

There is broad scope for improving these results, especially when taking into account that when we created our *CommonsCaptions* corpus we collected captions for 3,560 images, but 19,971 images with captions in Basque but not in Spanish were left out. In addition to this there are more than one million images with captions in Spanish but not in Basque.

Another possible improvement is the creation of bigger corpora with additional text information from Wikipedia and Wikidata: (1) the list of the titles in Basque and Spanish for all the articles in Wikipedia [10]; (2) descriptions in Wikidata for both languages; or (3) the first sentence in the contents in Basque and Spanish for all the articles in Wikipedia (if the number of words is not very different).

Acknowledgments

We are indebted to Elhuyar Foundation for providing the corpus to be used in this research. The research leading to these results was carried out as part of the TADEEP project (Spanish Ministry of Economy and Competitiveness TIN2015-70214-P, with FEDER funding). This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[10] Wikipedia_titles_es_eu_2018 Corpus: `http://ixa.si.ehu.eus/node/11500?language=en`

References

Alegria, I., U. Cabezon, U. F. de Betono, G. Labaka, A. Mayor, K. Sarasola, and A. Zubiaga. 2013. Reciprocal enrichment between basque wikipedia and machine translation. In *The People's Web Meets NLP*. Springer, pages 101–118.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Bojar, O., R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark.

Crego, J., J. Kim, G. Klein, A. Rebollo, K. Yang, J. Senellart, E. Akhanov, P. Brunelle, A. Coquard, Y. Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Diego, CA.

Dowling, M., L. Cassidy, E. Maguire, T. Lynn, A. Srivastava, and J. Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings of the The Fourth LRL Workshop: Language Technologies in support of Less-Resourced Languages*, pages 314–318, Poznan, Poland.

Dowling, M., T. Lynn, A. Poncelas, and A. Way. 2018. Smt versus NMT: Preliminary comparisons for Irish. In *Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, USA.

Etchegoyhen, T., E. M. Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. C. Etxabe, A. J. Carrera, I. E. Santos, and M. M. eta Eusebi Calonge. 2018. Neural machine translation

of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*, pages 139–148, Alicante, Spain.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Labaka, G., I. Alegria, and K. Sarasola. 2016. Domain adaptation in MT using titles in wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2209–2213, Portorož, Slovenia.

Labaka, G., C. España-Bonet, L. Màrquez, and K. Sarasola. 2014. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125.

Luong, M.-T. and C. D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam.

Nothman, J., N. Ringland, W. Radford, T. Murphy, and J. R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Otero, P. G. and I. G. López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25, Valletta, Malta.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Poncelas, A., G. M. de Buy Wenniger, and A. Way. 2018a. Data selection with feature decay algorithms using an approximated target side. In *15th International Workshop on Spoken Language Translation*, pages 173–180, Bruges, Belgium.

Poncelas, A., G. M. de Buy Wenniger, and A. Way. 2018b. Feature decay algorithms for neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alacant, Spain.

Poncelas, A., A. Way, and K. Sarasola. 2018. The adapt system description for the iwslt 2018 basque to english translation task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 76–82, Bruges, Belgium.

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725, Berlin, Germany,.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 41-48 recibido 31-03-2019 revisado 07-05-2019 aceptado 15-05-2019

Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon

Segmentación morfológica para extraer un lexicón bilingüe español-náhuatl

Ximena Gutierrez-Vasques[1], **Alfonso Medina-Urrea**[2], **Gerardo Sierra**[3]
[1]Universidad Nacional Autónoma de México
xim@unam.mx
[2] El Colegio de México
amedinau@colmex.mx
[3]Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Abstract: The aim of this work is to extract word translation pairs from a small parallel corpus and to measure the impact of dealing with morphology for improving this task. We focus on the language pair Spanish-Nahuatl, both languages are morphologically rich and distant from each other. We generate semi-supervised morphological segementation models and we compare two approaches (estimation, association) for extracting bilingual correspondences. We show that taking into account typological properties of the languages, such as the morphology, helps to counteract the negative effect of working with a low-resource language.
Keywords: Morphology, bilingual, translation, Nahuatl, Spanish

Resumen: El objetivo de este trabajo es extraer pares de traducción a partir de un corpus paralelo pequeño, así como medir el impacto de lidiar con la morfología para mejorar esta tarea. Nos enfocamos en el par de lenguas español-náhuatl, las dos lenguas son morfológicamente ricas y tipológicamente distantes. Generamos modelos semisupervisados de segmentación morfológica y comparamos dos enfoques (estimativo, asociativo) para extraer pares bilingües de palabras. Mostramos que tomar en cuenta las propiedades tipológicas de la lengua, como la morfología, ayuda a contrarrestar el efecto negativo de trabajar con una lengua de bajos recursos.
Palabras clave: Morfología, bilingüe, traducción, nahuatl, español

1 *Introduction*

In natural language processing (NLP), bilingual lexicon extraction is the task of obtaining a list of word pairs deemed to be word-level translations (Haghighi et al., 2008). This is an important task, since bilingual dictionaries are expensive resources that are not always available for all language pairs, specially for low-resource languages. Moreover, extracting lexical translations is an important step for building statistical machine translation (SMT) models. There is a wide variety of approaches to perform this task. However, most of them assume that there are large amounts of clean parallel corpora readily available.

The performance of downstream bilingual lexicon extraction methods tends to drop when they face small amount of data or a distant language pair. Working with low-resource languages in NLP implies several challenges due to the lack of language technologies and digitally available corpora. Traditional methods need to be adapted in order to deal with the scarcity.

Our case of study is the language pair Spanish-Nahuatl, which is spoken in the same country. These languages are typologically distant (Indo-European and Uto-Aztecan language families). In order to perform bilingual lexicon extraction for this language pair, there are several issues to address, i.e., it is difficult to obtain corpora, these languages have different morphological and syntactical phenomena, and it is not easy to find standardized annotated resources for Nahuatl.

Due to this, we explore the performance

ISSN 1135-5948. DOI 10.26342/2019-63-4 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

of widely used bilingual lexicon extraction methods when applied to a small parallel corpus of Spanish-Nahuatl. We conjecture that in order to find bilingual correspondences in a low-resource setting, it is not only important to pay attention to the methods but also to the morphology of the languages.

We apply different types of morphological processing to measure the impact of morphology in the quality of translation pairs. We trained our own semi-supervised morphological segmentation models.

Our work tries to be as unsupervised as possible, since it is difficult to rely on Nahuatl resources, due to the lack of orthographic norm, big dialectal variation and scarcity of resources.

The corpus used for our experiments is small in terms of the amount of text required by popular NLP models. However, we would like to highlight the challenges that arise when working with this language pair, and also to establish a first step in the development of automatic translation technology which is currently not available for this language pair.

The structure of the paper is as follows. Section 2 contains a brief overview regarding bilingual lexicon extraction methods and morphological phenomena of languages. Section 3 describes the dataset and methods used in our work. Finally, section 4 and 5 contain the discussion and conclusions based on the results.

2 Related work

2.1 Bilingual lexicon extraction

Bilingual lexicon extraction has been an active area of research for several years, especially with the availability of big amounts of parallel corpora that allow to model the relations between lexical units in one corpus and lexical units of the translated texts. This task became very important in terms of SMT systems, where word and phrase alignments are a fundamental step to translate a whole sentence.

Some of the most popular methods for word alignment are the IBM-models (Brown et al., 1993). They are based on an estimation approach that involves the use of probabilistic models that estimate parameters through a maximization process and produces probabilistic tables of lexical translations.

There are also methods that are based on an association approach, where word association or similarity measures are taken into account to find the word translations in a parallel corpus (Tufiş and Barbu, 2002; Ahrenberg, Andersson, and Merkel, 1998; Fung and Yee, 1998; Moore, 2005; Lardilleux, Lepage, and Yvon, 2011).

However, the quality of word alignment methods is heavily dependant on the amount of parallel data. There are alternative approaches, e.g., some works assume that if there is not enough parallel corpora for a language pair, there is enough comparable corpora or monolingual corpora for each of the languages. In these approaches bilingual lexicons are induced by taking into account several features, e.g, orthographic and temporal similarity (Schafer and Yarowsky, 2002), association measures, topical information (Mimno et al., 2009) and contextual features (Harris, 1954; Firth, 1957).

There are many works focused on the latter, the most recent ones use distributional and distributed vector representations. The idea is to build multilingual representations or to map monolingual vectors to the same space in order to find the closest translations (Lauly, Boulanger, and Larochelle, 2014; Hermann and Blunsom, 2014; Mikolov, Le, and Sutskever, 2013). These state of the art methods may not require parallel corpora but they are still based on huge amounts of monolingual or comparable corpora in order to work properly. It has been shown that when they face a low resource setting they can have even worst performance than less sophisticated methods (Levy et al., 2016).

Another alternative is to use pivot languages as an intermediary language to extract bilingual lexicon (Tanaka and Umemura, 1994; Wu and Wang, 2007; Kwon, Seo, and Kim, 2013).

2.2 Morphology

Morphology deals with the internal structure of words. Languages of the world have different word production processes. This morphological richness vary from language to language, depending on their linguistic typology.

In NLP, morphology is usually tackled by building morphological analysis/taggers tools. Commonly, lemmatization and stemming methods are used to reduce the morphological variation by converting words

forms to a standard form, i.e., a lemma or a stem. However, most of these technologies are focused in a reduced set of languages. For languages like English, with plenty of resources and relatively poor morphology, morphological processing may be considered solved.

However, this is not the case for all the languages. Specially for those with complex morphological phenomena where it is not enough to remove inflectional endings in order to obtain a stem.

Taking into account the morphological characteristics of languages is important for bilingual lexicon extraction tasks, since the alignment complexity between typologically different languages is far from the alignment complexity between similar languages (Cakmak, Acar, and Eryigit, 2012).

Moreover, when the morphological variation in a text is reduced, it can lead to an improvement of the performance of several NLP tasks, specially for low resource settings. For instance, Nießen and Ney (2004) incorporated the morphology of a low-resource language pair. By doing this, they significantly reduced the amount of parallel corpora needed to train a machine translation system.

Recently there has been a renewed interest in morphology from the NLP perspective, e.g., building vector representations of the morphs in order to improve the word representations (Lazaridou et al., 2013; Botha and Blunsom, 2014; Soricut and Och, 2015).

3 Methodology

3.1 The parallel corpus

Nahuatl is an indigenous language of Mexico with around 1.5M speakers. It is mostly spoken in central Mexico. Nahuatl does not have a web presence or text production comparable to Spanish. We decided to work with a parallel corpus. Since most of the documents that can be easily found in Nahuatl are translations, it seems easier to obtain parallel corpora than monolingual.

We used an existent digital parallel corpus that was created for this language pair (Gutierrez-Vasques, Sierra, and Pompa, 2016) and that is freely available through a search interface[1]. It is important to mention that this corpus was originally extracted from non digital books, therefore, it was digitized using an optical character recognition software (OCR) that could not properly identify Nahuatl words and made several types of mistakes (the authors performed a manual correction).

The documents gathered in this parallel corpus are not homogeneous, they come from different domains and there is dialectal, diachronic and orthographic variation. The Nahuatl language does not have a standardized orthography. The lack of an orthographic norm is an issue that has a negative impact in NLP tasks, since there can be many different word forms corresponding to the same word.

Although this corpus has around 1,186,662 tokens (taking into account the documents of both languages), we only used a subset (Table 1). We chose the documents that had more or less systematic writing, i.e., similar orthography regardless of the domain.

Language	**Tokens**	**Types**	**Sentences**
Spanish (ES)	118364	13233	5852
Nahuatl (NA)	81850	21207	5852

Table 1: Size of the parallel corpus

3.2 Morphological normalization and segmentation

We have mentioned before that taking into account typological properties of the morphology of the languages can help to improve the estimation of bilingual correspondences. In this sense, it is not only important to pay attention in the word alignment methods but also in the morphological text representation.

On one hand, reducing the morphological variation in a text can counteract the negative effect of working in a low-resource setting. The more productive the morphological inflection system of a language, the greater the number of different word forms in a text (Kelih, 2010), i.e., for highly productive languages it is less likely to find repeated words within a text (specially if there is corpora scarcity). Few repetitions of words can be problematic for extracting bilingual correspondences using statistical NLP methods, e.g., not enough contexts to model a word.

On the other hand, in order to find lexical

[1] Axolotl corpus http://www.corpus.unam.mx/axolotl

correspondences between Spanish and Nahuatl, it is important to notice their morphological phenomena. Nahuatl can agglutinate many different prefixes and suffixes to build complex words. Spanish uses mainly suffixes and it has a fusional behavior, where morphemes can be fused or overlaid into a single one. Table 2 shows an example[2].

ti-nech-caqui-z-nequi 2.SG.S-1.S.O-'hear'-FUT-'want' "Tú me quieres oír" (Spanish) *you want to hear me*
Lexical correspondence: oir-caqui

Table 2: Example of Nahuatl-Spanish

In order to normalize the texts and obtain morphological representations, we used different types of processing. For languages like Spanish, it may be enough to use lemmatization or stemming to reduce morphological variation. There are plenty of available tools, so we performed lemmatization and stemming of the Spanish texts using the FreeLing tool (Padró and Stanilovsky, 2012).

However, for Nahuatl, it is necessary to perform morphological segmentation. Since Nahuatl is a language that faces scarcity of data and of language technologies, we trained our own semi-supervised segmentation models using Morfessor 2.0 (Virpioja et al., 2013).

The development and test sets were built with the help of linguists that morphologically segmented a set of words. Morfessor works in a unsupervised way, i.e., it does not necessarily require a tagged training corpus to build a model. However, a tagged development set, with gold-standard segmentations, can be added in order to improve the generated model. For Nahuatl, we used small development and test sets, therefore, we used a semi-supervised setting.

Although it is not common to perform morphological segmentation of Spanish, we also trained a semisupervised segmentation model using Morfessor 2.0 for Spanish. In this case, there were already available development and test sets (Méndez-Cruz, Medina-Urrea, and Sierra, 2016). Table 3 shows the size of the datasets that were used for training morphological segmentation models.

[2]Leipzig Glossing Rules were used for interlinear morpheme-by-morpheme glosses

Spanish	**Tokens**	**Types**
Training	2175533	99564
Development	800	800
Test	792	792
Nahuatl	**Tokens**	**Types**
Training	83229	22174
Development	1379	1379
Test	288	288

Table 3: Datasets used for morphological segmentation

Since we aimed to achieve the best segmentation possible under these conditions, we optimized the main parameters involved in the segmentation models trained with Morfessor 2.0. We tried several values (from 0.1 to 10) of the unannotated corpus likelihood weight (α) that controls the oversegmentation or undersegmentation of the model (Smit et al., 2014). We also tried several ways of calculating the counts of words during the training phase (token based training, log of tokens, type based training).

Tables 4 and 5 show several of our segmentation models and their evaluation. BPR metric (Virpioja et al., 2011) was used, this is a popular metric that evaluates how correctly the morpheme boundaries are placed within the words, compared to a gold-standard reference.

Finally, we selected the best evaluated models. For Nahuatl we used the model obtained with α =0.8 and for Spanish the one with α =0.4.

3.3 Spanish-Nahuatl lexicon extraction

Since we have parallel corpus available, one natural step would be to train a SMT system. However, since we are facing a low-resource setting of distant languages, we decided to take a "first-things-first" approach (Monson et al., 2004), i.e., first explore how difficult would be to find word to word correspondences and the impact that different morphological representations can have for improving this task.

In order to automatically extract bilingual word pairs, we used two different approaches. We used an estimation approach, in particular, we used IBM model 1 (Brown et al., 1993). This model has proven to be still a

α	Pre	Rec	F
0.1	67.7%	89.3%	77%
0.2	69.6%	87.7%	77.5%
0.3	70.1%	84.9%	76.8%
0.4	70.9%	83.6%	76.7%
0.5	71.2%	82.3%	76.4%
0.6	72.5%	81.6%	76.8%
0.7	73.5%	79.6%	76.4%
0.8	**75.1**%	**80**%	**77.5**%
0.9	75.5%	77.4%	76.4%
1	76.6%	77.4%	77%
3	84.5%	49.6%	62.5%
10	97.6%	22.9%	37.1%

Table 4: Morphological segmentation of Nahuatl using Morfessor 2.0 (token based training)

α	Pre	Rec	F
0.1	77.1%	82.6%	79.7%
0.2	80.8%	82.1%	81.4%
0.3	82.2%	81.6%	81.9%
0.4	**84**%	**80.6**%	**82.3**%
0.5	83.9%	78.4%	81.1%
0.6	85%	77.8%	81.3%
0.7	86.3%	77%	81.4%
0.8	86%	75.8%	80.6%
0.9	86.4%	75.5%	80.6%
1	87.2%	75.5%	80.9%
3	89.8%	66.3%	76.2%
10	97.9%	24.4%	39.1%

Table 5: Morphological segmentation of Spanish using Morfessor 2.0 (logarithm of token frequency training)

strong baseline for bilingual lexicon extraction tasks (Levy et al., 2016).

On the other hand, we used an association based approach (Lardilleux and Lepage, 2009), where only those words that appear exactly in the same sentences are considered for alignment. The idea of this method is to produce more candidates, artificially, by creating many subcorpora of small sizes (subsampling). This method is usually known as sampling-based or Anymalign.

We chose these methods because they have similar performance buy they may not make the same mistakes, since they are based in different approaches.

It is not common to test these downstream methods in a low-resource setting. However, sample-based method (anymalign) has shown high accuracy for extracting translation candidates of low-frequency words (Kwon et al., 2014; Lardilleux, Lepage, and Yvon, 2011). On the other hand, IBM models have been used to extract bilingual lexicon from small quantities of parallel sentences of phonemic transcriptions (Adams et al., 2015).

We used different types of morphological text representations: Without any processing (ES-NA), Spanish lemmatized, Nahuatl segmented (ES_{lem}-NA_{morph}) and both languages morphologically segmented (ES_{morph}-NA_{morph}). An evaluation set was built by random sampling 150 Spanish words with frequency greater than 2 in the corpus, stopwords and grammatical words were excluded, the evaluation set mostly contains verbs, nouns and adjectives. For each of these words, several translation candidates were annotated, i.e., an expert annotated possible translations using the parallel corpus as a reference[3]. This evaluation, or test, lexicon required the help of a human annotator since it is difficult to rely in a single standardized Spanish-Nahuatl dictionary to extract evaluation word pairs (due to the orthographic and dialectal variation).

In order to make the results more comparable, we used the same set of words across the different morphological analyses, i.e., we took the random sample of words (without any morphological analysis) and we lemmatized and segmented them to build the different test sets.

Precision at 1 (p@1) and precision at 5 (p@5) were used for the evaluation, i.e., it was evaluated if a correct translation was within the top 1 or the top 5 candidates (Table 6).

It is important to mention that the translations were evaluated out of context. The evaluation, or test, lexicon was done by a human annotator since it is difficult to rely in a single standardized dictionary to extract evaluation word pairs.

Additionally, we were interested in performing a deeper quantitative and qualitative analysis. We focused on the analysis of

[3]The evaluation set and translation candidates can be found at https://github.com/ElotlMX/nahuatl

nouns and verbs. Table 7 shows, from the total amount of nouns and verbs in the evaluation dataset, what proportion was correctly translated.

In a more qualitative analysis, we noticed that there are some verbs in Spanish that, in spite of being very frequent, they were difficult to translate by the methods, e.g., copular or linking verbs (to be), auxiliar verbs (have). We conjecture that some copular and auxiliar verbs were difficult to align, since in Spanish these types of verbs are expressed in the syntax, while in Nahuatl they may be expressed in the morphology (agglutinated to another verb or a noun), or not appear at all.

	IBM %		ANYM %	
	p@1	p@5	p@1	p@5
ES-NA	48.9	73.1	43.8	61.3
ES_{lem}-NA_{morph}	54.6	78.6	**66.6**	**89.3**
ES_{morph}-NA_{morph}	49	73.9	57.4.8	79.9

Table 6: Bilingual lexicon extraction evaluation

	Verbs %		Nouns %	
ES-NA	p@1	p@5	p@1	p@5
IBM	41	66	50.9	75.4
ANYM	33.9	55.3	52.8	67.9
ES_{lem}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	58.3	79.1	50	67.2
ANYM	70.8	**89.5**	53.4	74.1
ES_{morph}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	38.2	61.7	54.3	**78.2**
ANYM	47	79.4	60.8	73.9

Table 7: Proportion of nouns and verbs in the dataset that were correctly translated

4 Discussion

Table 6 shows that all the types of morphological processing that we applied, helped to improve the bilingual lexicon extraction in our low-resource setting.

The most suitable text representation for obtaining translation pairs seems to be ES_{lem}-NA_{morph}. This type of morphological representation not only achieves better results compared to the methods applied to texts without any processing *(ES-NA)*, but the obtained pairs are closer to a Spanish-Nahuatl dictionary entry (word forms with few or none inflections).

Moreover, this setting is the one that is able to obtain the translation of more verbs (Table 7). This is important since verbs constitute the morphologically most complex word class in Nahuatl. Therefore, we would expect these translations were the more challenging to obtain.

The association based method ($ANYM$) seems to greatly benefit from the morphological analysis.

The evaluation of the ES_{morph}-NA_{morph} setting was not so straightforward. The precision in this setting was calculated taking into account less translation pairs, since we had to discard several problematic cases, e.g., Spanish single morphs that correspond to several grammatical functions. It is not always possible to find a morph to morph correspondence between the two languages, due to the differences between their morphological phenomena.

5 Conclusions

In this work, we present morphological segmentation models for Nahuatl and Spanish and we performed Spanish-Nahuatl bilingual lexicon extraction in a low-resource setting. Our conjecture was that morphology plays an important role to improve the performance of this task, specially when we deal with two morphologically-rich distant languages. We applied different types of morphological processing to the texts (stemming, lematization, segmentation) and we extracted bilingual correspondences using two popular approaches.

Using lemmas for Spanish and morphs for Nahuatl, combined with a sampling-based method, achieved the best performance. We showed that developing morphological segmentation tools is specially important for Nahuatl, this is an agglutinative and polysynthetic language that benefits from morphological segmentation, however, since Nahuatl is a low-resource language, there is scarcity of language technologies to process it.

As future work, we would like to extend and automatize the evaluation since relying on human annotators is expensive. We hope that this work could be useful for building translation technologies for this language pair.

Acknowledgements

This work was supported by the Mexican Council of Science and Technology (CONACYT), fund 2016- 01-2225.

References

Adams, O., G. Neubig, T. Cohn, and S. Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*.

Ahrenberg, L., M. Andersson, and M. Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 17th international conference on Computational linguistics*, pages 29–35. Association for Computational Linguistics.

Botha, J. and P. Blunsom. 2014. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Cakmak, M. T., S. Acar, and G. Eryigit. 2012. Word alignment for english-turkish language pair. In *LREC*, pages 2177–2180.

Firth, J. 1957. Papers in linguistics, 0xford: Oxford university press.

Fung, P. and L. Y. Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, pages 414–420. Association for Computational Linguistics.

Gutierrez-Vasques, X., G. Sierra, and I. H. Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

Harris, Z. S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hermann, K. M. and P. Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Kelih, E. 2010. The type-token relationship in slavic parallel texts. *Glottometrics*, 20:1–11.

Kwon, H.-S., H.-W. Seo, M. Cheon, and J.-H. Kim. 2014. Iterative bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm. *Journal of Contemporary Engineering Sciences*, 7(24):1335–1343.

Kwon, H.-s., H.-w. Seo, and J.-h. Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 11–15.

Lardilleux, A. and Y. Lepage. 2009. Sampling-based multilingual alignment. In *Recent Advances in Natural Language Processing*, pages 214–218.

Lardilleux, A., Y. Lepage, and F. Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217.

Lauly, S., A. Boulanger, and H. Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Lazaridou, A., M. Marelli, R. Zamparelli, and M. Baroni. 2013. Compositionally derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

Levy, O., A. Søgaard, Y. Goldberg, and I. Ramat-Gan. 2016. A strong baseline for learning cross-lingual word embeddings from sentence alignments. *arXiv preprint arXiv:1608.05426*.

Méndez-Cruz, C.-F., A. Medina-Urrea, and G. Sierra. 2016. Unsupervised morphological segmentation based on affixality measurements. *Pattern Recognition Letters*, 84:127–133.

Mikolov, T., Q. V. Le, and I. Sutskever. 2013. Exploiting similarities among languages

for machine translation. *arXiv preprint arXiv:1309.4168*.

Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889. Association for Computational Linguistics.

Monson, C., A. Lavie, J. Carbonell, and L. Levin. 2004. Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, pages 52–61. Association for Computational Linguistics.

Moore, R. C. 2005. Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 1–8. Association for Computational Linguistics.

Nießen, S. and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Schafer, C. and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning*, pages 1–7. Association for Computational Linguistics.

Smit, P., S. Virpioja, S.-A. Grönroos, M. Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Soricut, R. and F. J. Och. 2015. Unsupervised morphology induction using word embeddings. In *HLT-NAACL*, pages 1627–1637.

Tanaka, K. and K. Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics*, pages 297–303. Association for Computational Linguistics.

Tufiş, D. and A.-M. Barbu. 2002. Lexical token alignment: Experiments, results and applications. In *Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas*, pages 458–465.

Virpioja, S., P. Smit, S.-A. Grönroos, M. Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Virpioja, S., V. T. Turunen, S. Spiegler, O. Kohonen, and M. Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *TAL*, 52(2):45–90.

Wu, H. and H. Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Análisis Textual

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 51-58 recibido 01-04-2019 revisado 09-05-2019 aceptado 14-05-2019

Evaluation of Distributional Semantic Models for the Extraction of Semantic Relations for Named Rivers from a Small Specialized Corpus

Evaluación de Modelos Semánticos Distribucionales para la Extracción de Relaciones Semánticas Activadas por Ríos con Nombre Propio de un Corpus Especializado de Pequeño Tamaño

Juan Rojas Garcia, Pamela Faber
Universidad de Granada
{juanrojas, pfaber}@ugr.es

Abstract: EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of landform concepts such as named rivers (e.g., *Nile River*), distributional semantic models (DSMs) were applied to a small-sized, specialized corpus to extract the terms related to each named river mentioned in it and their semantic relations. Since the construction of DSMs is highly parameterized and their evaluation in small specialized corpora has received little attention, this paper identified parameter combinations in DSMs suitable for the extraction of the semantic relations *takes_place_in*, *affects*, and *located_at*, frequently held by named rivers in the corpus. The models were thus evaluated using three gold standard datasets. The results showed that, for a small-sized corpus, count-based models outperformed prediction-based ones with the log-likelihood association measure, and the detection of a specific relation depended largely on the context window size.
Keywords: Named river, terminology, knowledge representation, distributional semantic model, text mining

Resumen: EcoLexicon (http://ecolexicon.ugr.es) es una base de conocimiento terminológica sobre el medioambiente, cuyo diseño permite la contextualización geográfica de los ríos con nombre propio (RNP) (v.gr., *Río Nilo*). Se aplicaron modelos semánticos distribucionales (MSD) a un corpus especializado de pequeño tamaño para extraer los términos relacionados con los RNP y sus relaciones semánticas. Puesto que el funcionamiento de los MSD depende de la configuración de sus parámetros, y su evaluación en corpus especializados de pequeño tamaño ha sido menos explorada, en este artículo se identifica la combinación de parámetros adecuada para extraer las relaciones semánticas *tiene_lugar_en*, *afecta* y *localizado_en*, activadas frecuentemente por los RNP. Los MSD se evalúan con tres conjuntos de datos anotados manualmente. Los resultados indican que, para un corpus de pequeño tamaño: los modelos basados en recuentos con la medida de asociación *log-likelihood* superan a los modelos predictivos; y la representación de una relación específica depende del tamaño de la ventana contextual.
Palabras clave: Río con nombre propio, terminología, representación del conocimiento, modelo semántico distribucional, minería de textos.

1 Introduction

EcoLexicon (http://ecolexicon.ugr.es) is a multilingual, terminological knowledge base on environmental science that is the practical application of Frame-based Terminology (Faber, 2012). The flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas. However, the representation of geographically contextualized LANDFORM concepts, such as named rivers (e.g., *Mississippi River*), depends on knowing which terms are semantically related to each named

ISSN 1135-5948. DOI 10.26342/2019-63-5 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

landform, and how these terms are related to each other.

With the aim of representing in EcoLexicon the conceptual structures underlying the usage of named landforms mentioned in a small-sized, English specialized corpus on Costal Engineering (7 million tokens), the terms related to each named landform and their semantic relations were extracted with distributional semantic models (DSMs) and other statistical techniques (Rojas Garcia and Faber, 2019a and 2019b). In this paper, we focus on named rivers.

DSMs represent the meaning of a term as a vector by considering the statistics of its cooccurrence with other terms in the corpus. Although a DSM can help identify semantic relations between terms, the construction of a suitable DSM for a particular task is highly parameterised. Even though numerous studies have addressed the evaluation and optimization of DSMs in very large, general corpora (Baroni et al., 2014; Kiela and Clark, 2014), the ability of DSMs to capture different semantic relations in smaller specialized corpora has received little attention.

The objective of this paper was to identify parameter combinations in DSMs suitable for the extraction of three semantic relations, held by named rivers, in the small specialized corpus mentioned above. Hence, the models were evaluated using evaluation data that contained pairs of semantically related terms, manually extracted from the same corpus. One of the terms was always a named river, and the other one was an entity or process. The semantic relations that linked the terms were those frequently activated by named rivers in the corpus, namely: (1) *takes_place_in*; (2) *affects*; and (3) *located_at*. Three gold standard datasets were thus built.

The rest of this paper is organized as follows. Section 2 provides background on DSMs, as well as a literature review on their application and evaluation. Section 3 explains the materials, methods, and DSMs evaluation applied in this study, and the construction of the gold standard datasets. Section 4 shows the results obtained. Finally, Section 5 discusses the results, and presents the conclusions derived from this work as well as plans for future research.

2 Background and Literature Review

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller. and Charles, 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance or cosine similarity, *inter alia*.

Depending on the language model (Baroni et al., 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde et al., 2006) is an example of this type of model.

Prediction-based models exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include continuous bag-of-words (CBOW) and skip-gram models (Mikolov et al., 2013).

Count-based DSMs have been extensively studied (Kiela and Clark, 2014; Lapesa et al., 2014; Sahlgren and Lenci, 2016). Research shows that parameters, such as the context window size, influence the semantic relations that are captured, either syntagmatic relations or paradigmatic relations (i.e., synonymy, antonymy, hyponymy and meronymy). The syntagmatic relations examined in much research are either phrasal associates (e.g., *help - wanted*) (Lapesa et al., 2014) or syntagmatic predicate preferences (Erk et al., 2010) in general language. In this study, we focused on three specific syntagmatic relations, namely, *takes_place_in*, *located_at*, and *affects*, which were activated by named rivers in the specialized language of Coastal Engineering. As far as we know, this framework has not been studied in the context of DSM evaluation, which constitutes an original aspect of this work.

The ability of count-based models and prediction-based models (CBOW and skip-gram) to detect the three syntagmatic relations is described in this paper. Both types of DSM have also been recently compared. Baroni et al. (2014) compared them on several datasets and found that the prediction-based models provided better results. In contrast, Ferret (2015) found that count-based models performed better. In another work that compared the ability of both DSMs to capture paradigmatic relations (synonymy, antonymy, and hyponymy) and syntactic derivatives, Bernier-Colborne and Drouin (2016) observed that not only the semantic relations detected by the DSMs

depended on the window size, but also the values of this parameter mostly coincided in both DSMs.

Levy et al. (2015) yielded valuable insights, showing the following: (1) when the parameters of the models were tuned correctly, count-based and prediction-based models obtained similar accuracy; and (2) the best model depended on the task to be carried out. Nevertheless, Asr et al. (2016), and Sahlgren and Lenci (2016) reported that count-based models outperformed prediction-based ones on small-sized corpora of under 10 million tokens.

Work in lexical semantics and DSMs includes, *inter alia*, the identification of semantic relations (Bertels and Speelman, 2014), classification of verbs into semantic groups (Gries and Stefanowitsch, 2010), and the use of word vectors as features for automatic recognition of named entities in text corpora (El Bazi and Laachfoubi, 2016).

3 Materials and Methods

3.1 Materials

3.1.1 Corpus Data

The named rivers and related terms were extracted from a subcorpus of English texts on Coastal Engineering, on which the DSMs were also built. This subcorpus, comprising roughly 7 million tokens, is composed of specialized and semi-specialized texts, and is an integral part of the EcoLexicon English Corpus (23.1 million tokens).

3.1.2 GeoNames Geographic Database

The automatic detection of the named rivers in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) has over 10 million proper names for 645 different geographic entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored.

3.1.3 Gold Standard Datasets

The DSMs, built on our corpus, were evaluated on gold standard data, manually extracted from the same corpus. The gold standard datasets contained pairs of semantically related terms, in which the semantic relations were those frequently activated by named rivers in the corpus, namely, *takes_place_in*, *affects*, and *located_at*. Three gold standard datasets[1] were thus built, one for each of the semantic relations. It is necessary to note that the designations and meaning of these relations are those used in EcoLexicon (Faber et al., 2009).

The three semantic relations always linked a named river to an entity or process expressed by a nominal term. Specifically, the *takes_place_in* relation holds between a process (e.g., *runoff*) and a named river where the process occurs (see Table 1). The *affects* relation relates a named river to an entity or process (e.g., *soft mud*, *supply of sediment*) on which the named river performs a causative action (see Table 2). Finally, the *located_at* relation indicates the location of an entity (e.g., *jetty*) in a named river (see Table 3).

The three datasets contain 100 instances for each semantic relation, which were all used for the evaluation.

process	*takes_place_in*	named river
consolidation of the land	*takes_place_in*	Mississippi river mouth
runoff	*takes_place_in*	Mississippi river basin
sea level rise	*takes_place_in*	Mississippi river delta

Example from the corpus:
(1) ***Consolidation of the land*** *is occurring, as noted before, at the* ***mouth of the Mississippi River****, where the...*

Table 1: Extract from the gold standard dataset for the *takes_place_in* relation

named river	*affects*	entity / process
Mississippi river	*affects*	Saint Bernard river delta
Mississippi river	*affects*	soft mud
Seine River	*affects*	fresh water

Example from the corpus:
(1) *The Chandeleurs Islands are remnants of the* ***Saint Bernard River delta****, formed by the* ***Mississippi River****.*

Table 2: Extract from the gold standard dataset for the *affects* relation

entity	*located_at*	named river
jetty	*located_at*	Mississippi river mouth
soft mud	*located_at*	Mississippi river mouth
barrier island	*located_at*	Mississippi river mouth

Example from the corpus:
(1) *Natural and anthropogenic effects combine to result in the maximum erosional stress on* ***barrier islands****, located near the* ***mouth of the Mississippi River****.*

Table 3: Extract from the gold standard dataset for the *located_at* relation

The annotation of the pair of terms extracted from the corpus was carried out by three terminologists from our research group. *Cohen's*

[1] The datasets will be available on the website of the LexiCon Research Group at the University of Granada (http://lexicon.ugr.es/).

kappa coefficient was used as the statistical measure of inter-annotation agreement, and the scores for all the annotator pairs were over 90%.

3.2 Methodology

3.2.1 Pre-processing

The corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased with the Stanford *CoreNLP* package (Manning et al., 2014) for R programming language (R Core Team 2019). The multiword terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

In the DSMs, only terms larger than two characters were considered. Numbers, symbols, and punctuation marks were removed. Since closed-class words are often considered too uninformative to be suitable context words, stopwords were not used (i.e., determiners, conjunctions, relative adverbs, and prepositions). Additionally, the minimal occurrence frequency was set to 5 so that the co-occurrences were statistically reliable (Evert S., 2008).

3.2.2 Named River Recognition

Both normalized and alternate names of the rivers in GeoNames were searched in the lemmatized corpus. The recognized designations were normalized and automatically joined with underscores. Most rivers cited in the corpus were in GeoNames (97%), while others were identified by manual inspection (3%). Anaphoric elements referring to a river were replaced by the corresponding named river in the lemmatized corpus. For this task, the automatic anaphora resolution function from *CoreNLP* package was used, and other cases were manually replaced. The 250 rivers with the highest number of mentions in the corpus are shown on the map in Figure 1.

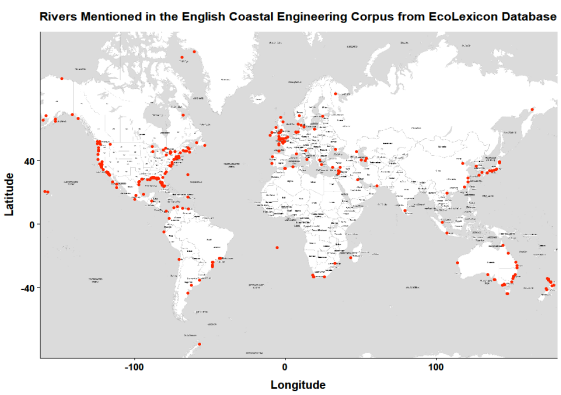


Figure 1: Location of the named rivers mentioned in the corpus

3.2.3 Construction of the DSMs

The experiment we carried out involved a comparative evaluation of two types of DSMs for a small-sized, specialized corpus, namely, count-based and prediction-based models. Both model types produced the vector representation of a term based on the contexts in which it appeared in our corpus. For this paper, the contexts of a target term (i.e., a named river) were the terms that co-occurred with it inside a sliding context window, which spanned a certain number of terms on either side of the target term.

These DSMs have various parameters that must be set to build the model. The parameters impinge on both the term representations produced, and the accuracy of the similarity scores between term vectors when the models are compared (Baroni et al., 2014).

Therefore, to assess the influence of the parameters of both DSMs on their ability to capture the three semantic relations targeted in this paper, various settings for each parameter were tried, and the combinations of these parameter settings were evaluated.

3.2.4 Parameter Setting of the Count-based Model

The first model type evaluated was a count-based model, also called bag-of-words (BOW) model. The BOW model was built with the R package *quanteda* (Benoit et al., 2018) for text mining.

To build a BOW model, a term-term matrix of co-occurrence frequencies was first computed, according to a specific size for the sliding context window. Then, the matrix was subjected to a specific weighting scheme, namely, an association measure that increases the importance of the context terms that are more indicative of the meaning of the target term. A dimensionality reduction technique could also transform this weighted matrix, but for this work, it was not applied. Therefore, the 3,000 most frequent words were used, which included all the named rivers and terms stored in the three evaluation datasets.

Regarding the context window, we tested size values ranging from 1 to 10 words on either side of the target term, and the context window was allowed to span sentence boundaries. The context window shape was always rectangular (i.e., the increment added to the co-occurrence frequency of a pair of terms was always 1, regardless of the distance between the two terms inside the context window). The frequencies observed on the left and right of a target term were added.

With respect to the weighting schemes, three association measures, defined in Evert S.'s (2008) work on collocation, were tested: (1) statistical log-likelihood; (2) positive pointwise mutual information (PPMI); (3) *t*-score. Log-likelihood and PPMI are widely used in computational linguistics, whereas *t*-score is popular in computational lexicography (Evert S. et al., 2017).

Research in computational linguistics reveals that log-likelihood is able to capture syntagmatic and paradigmatic relations (Lapesa et al., 2014), and to achieve better performance for medium-to-low-frequency data than other association measures (Alrabia et al., 2014). PPMI and *t*-score, on the other hand, have been found to work adequately for different applications in previous research when compared to other association measures (Baroni et al. 2014; Kiela and Clark, 2014).

Finally, following Lapesa et al. (2014), the association scores were transformed to reduce skewness in this way: log-likelihood and PPMI scores were both transformed by adding 1 and calculating then the natural logarithm (ln), whereas *t*-scores were transformed by calculating the square root (sqrt).

The settings tested for each of the two parameters were:

1. Size of the context window: 1-10 words.
2. Weighting scheme: ln(log-likelihood + 1), ln(PPMI + 1), sqrt(*t*-score).

3.2.5 Parameter Setting of the Prediction-based Model

The second model type evaluated was a prediction-based model, specifically, the model known as *word2vec* (W2V) (Mikolov et al., 2013). These term vectors are learned by training a neural network on a corpus according to two different architectures. The architecture continuous bag-of-words (CBOW) aims to predict the target term based on its context terms, while the architecture skip-gram aims to predict the context terms of a target term. The W2V model was built with the R package *wordVectors* (Schmidt and Li, 2017).

For W2V, five hyperparameters were examined, the same as those tested by Bernier-Colborne G. and Drouin (2016) for paradigmatic relations and syntactic derivatives. The first one was the architecture used to learn the term vectors. The second one was the training algorithm, either using a hierarchical softmax function, or by sampling negative examples, in which case the number of negative samples must be selected. The third hyperparameter was the subsampling threshold for frequent terms, namely, some occurrences of those terms whose relative frequency in the corpus is greater than a threshold, are randomly deleted before the model is trained. Finally, the dimensionality of the term vectors, and the size of the context window are the other hyperparameters.

The settings tested for each of the five hyperparameters were:

1. Architecture: CBOW or skip-gram.
2. Negative samples: 5, 10 or none (in this case, hierarchical softmax is used).
3. Subsampling threshold: low (10^{-5}), high (10^{-3}) or none.
4. Dimensionality of term embeddings: 100 or 300.
5. Size of context window: 1-10 words.

3.2.6 Evaluation of the DSMs

The DSMs were compared and described. First, for each named river, a sorted list of neighbours was obtained by computing the cosine similarity between the named river's vector and the vectors of all other context terms, and by sorting then these context terms in descending order of magnitude.

Subsequently, the sorted list of neighbours was evaluated on the whole gold standard dataset for each of the three semantic relations. The measure used to evaluate the models was mean average precision (MAP) (Manning et al., 1998). This measure tells us how accurate the sorted list of neighbours we get for a named river is, based on the rank of its related terms according to the gold standard. The nearer the related terms are to the top of this list on average for each named river, the higher the MAP.

4 Results

BOW and W2V models were compared by observing the MAP of each model on the three datasets. The maximum MAP achieved by each model is shown in Table 4.

The results indicated, on the one hand, that the BOW model reached a higher MAP than W2V on the three semantic relations when its parameters were correctly tuned. On the other hand, the *takes_place_in* relation was the most accurately captured by both models when they were tuned for this relation, followed by the *located_at* and *affects* relations.

Dataset	BOW model		
	Max. MAP	Weighting scheme	Window size
takes_place_in	**0.544** (0.347 ± 0.118)	LL	4
located_at	**0.418** (0.321 ± 0.056)	LL	2
affects	**0.351** (0.201 ± 0.053)	LL	3
Dataset	**W2V model**		
	Max. MAP		**Window size**
takes_place_in	0.346 (0.298 ± 0.042)		4
located_at	0.221 (0.196 ± 0.013)		2
affects	0.182 (0.141 ± 0.021)		3

Table 4: Maximum MAP (with average and standard deviation in brackets) of BOW and W2W models on each dataset. LL stands for the log-likelihood weighting scheme

The higher accuracy of *takes_place_in* may be due to the large number of instances in specialized texts in Coastal Engineering which express the processes that occur in named rivers. As for the *located_at* relation, it is also frequent that the texts mention the entities in named rivers. However, it seems that the number of instances of this semantic relation in the whole corpus is not large enough for the DSMs to represent them as accurately as *takes_place_in* instances.

We hypothesize that the low accuracy of both models to capture the *affects* relation could be caused by the lack of semantic expressivity of this relation, and by its high combinatorial potential to relate named rivers to both processes and entities. For instance, the conceptual proposition MISSISSIPPI RIVER *affects* SAINT BERNARD RIVER DELTA would be more meaningful if the relation were *creates* instead of *affects* (see sentence (1) in Table 2). The expressiveness of the *affects* relation could be increased by splitting it up into a number of different relations, as proposed by Reimerink A. and León-Araúz, (2017). In doing so, possibly, each of those specific relations could be more accurately represented by the DSMs.

Refining the *affects* relation not only would enhance the performance of DSMs in our opinion, but also would greatly improve the knowledge representation of named rivers, in the same way that the expressiveness of meronymy has been increased by splitting it up into six different relations in EcoLexicon (Reimerink and León-Araúz P, 2017).

Interestingly, in each dataset, the maximum MAP of W2V model was reached when the window size was the same as that of the BOW model. For that reason, to assess the impact of the window size on the accuracy of both DSMs, the average MAP for each setting of this parameter (i.e., for each window size between 1 and 10 words) is shown in Figure 2. The average MAP was used, instead of the maximum, because it allowed us to determine which window size settings consistently produced satisfactory results, regardless of the settings used for the other parameters.

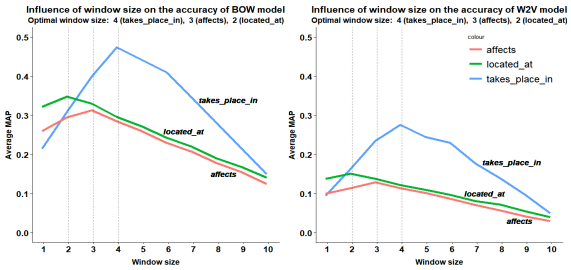


Figure 2: Average MAP of BOW model (left) and W2V model (right) with regard to window size

Figure 2 points out that, in both DSMs, the optimal size was 4 words for the *takes_place_in* relation, 3 words for *affects*, and 2 words for *located_at*.

Since the count-based model BOW notably outperformed the prediction-based W2V on the three datasets, for the sake of simplicity, the setting influence of the other four hyperparameters of W2V are succinctly reported because they did not lead to substantial accuracy improvements on either dataset. The neural network architecture skip-gram worked, on average, better than CBOW, and a negative sampling of 10 samples reached a larger MAP than the hierarchical softmax. The subsampling threshold was not conducive significant gains, and the optimal setting for the dimensionality of the term embeddings was 300 dimensions.

5 Conclusions

To extract knowledge for the representation in EcoLexicon of the conceptual structures (Faber P., 2012) that underlie the usage of named rivers in a small-sized, English Coastal Engineering corpus, count-based and prediction-based DSMs were applied to the corpus to extract the terms related to each named river. Since the construction of DSMs is highly parameterized, and their evaluation in small specialized corpora has received little attention, this paper identified parameter combinations in DSMs suitable for the extraction of the semantic relations *takes_place_in*, *affects*, and *located_at*, frequently held by named rivers in the corpus. The models were thus evaluated using three gold standard datasets.

Count-based models, with the log-likelihood association measure, showed the best performance for the three semantic relations. These results reinforce the findings of previous research which states, on the one hand, that count-based DSMs surpass prediction-based ones on small-sized corpora of under 10 million tokens (Asr et al., 2016; Sahlgren and Lenci, 2016), and on the other hand, that log-likelihood achieves greater accuracy for medium-to-low-frequency data than other association measures (Alrabia et al., 2014).

For both DSMs, the optimal window size depended on the semantic relation that was to be captured, and the specific values coincided in both DSMs, namely, a window size of 4 words for the *takes_place_in* relation, 3 words for *affects*, and 2 words for *located_at*. The dependence of the window size on the specific semantic relation, and the coincidence of the values in both DSMs are findings in line with those reported by Bernier-Colborne and Drouin (2016), upon comparing the ability of both DSMs to capture paradigmatic relations (synonymy, antonymy, and hyponymy), and syntactic derivatives.

It was also found that the *takes_place_in* relation was the most accurately represented by both DSMs, followed by *located_at*. The *affects* relation showed the lowest accuracy values, maybe due to the lack of semantic expressivity of this relation, and its high combinatorial potential to relate named rivers to both processes and entities (Faber et al., 2009). In future research, the *affects* relation will be split up into different relations, specific for named rivers, by following the methodology proposed by Reimerink A. and Léon-Araúz (2017), and by classifying the verbs that collocate with named rivers into the lexical domains and subdomains developed by Faber and Mairal (2009). In doing so, it will be analysed whether the DSMs reach higher accuracy values for each of the specific relations than for the *affects* relation.

Extensions of this work will include testing: other DSMs, such as GloVe; other parameters, such as the shape of the context window; and the application of the dimensionality reduction technique for texts Topic Modeling (Blei et al., 2003). Furthermore, the DSMs will be evaluated on gold standard datasets for named bays and beaches, and the three datasets for named rivers will be increased with more annotated data.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author.

References

Alrabia, M., N. Alhelewh, A. Al-Salman, and E. Atwell. 2014. An empirical study on the Holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics*, 5(1): 1-13.

Asr, F., J. Willits, and M. Jones. 2016. Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In A. Papafragou, D. Grodner, D. Mirman and J. Trueswell (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci)*, Philadelphia, 1092-1097.

Baroni, M., G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL*, vol. 1, 238-247.

Benoit K, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, & A. Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774.

Bernier-Colborne, G., and P. Drouin. 2016. Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm)*, Osaka (Japan), 52-61.

Bertels, A., and D. Speelman. 2014. Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2): 279-303.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

El Bazi, I., and N. Laachfoubi. 2016. Arabic named entity recognition using word representations. *International Journal of Computer Science and Information Security*, 14(8): 956-965.

Erk, K., S. Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse

selectional preferences. *Computational Linguistics*, 36(4): 723-763.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter, 1212-1248.

Evert, S., P. Uhrig, S. Bartsch, and T. Proisl. 2017. E-VIEW-alation – A large-scale evaluation study of association measures for collocation identification. In *Proceedings of the eLex 2017 Conference*, Leiden, 531-549.

Faber, P. (ed.). 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., and R. Mairal. 2009. *Constructing a Lexicon of English Verbs*. Berlin/New York: Mouton de Gruyter.

Faber, P., P. León-Araúz, and J.A. Prieto. 2009. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies* 1: 1-23.

Ferret, O. 2015. Réordonnancer des thésaurus distributionnels en combinant différents critères. *TAL*, 56(2): 21-49.

Gries, S., and A. Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In S. Rice, and J. Newman (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford (California): CSLI, 73-90.

Kiela, D., and S. Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, 21-30.

Lapesa, G., S. Evert, and S. Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, Dublin, 160-170.

Levy, O., Y. Goldberg, and I. Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, (3): 211-225.

Manning, C.D., P. Raghavan, and H. Schütze. 1998. *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.

Manning, C.D., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, 55-60.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale.

Miller, G.A., and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1): 1-28.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reimerink, A., P. and León-Araúz. 2017. Predicate-argument analysis to build a phraseology module and to increase conceptual relation expressiveness. In R. Mitkov (ed.), *Computational and Corpus-Based Phraseology*. Cham (Switzerland): Springer, 176-190.

Rohde, D., L. Gonnerman, and D. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8: 627-633.

Rojas-Garcia J., and P. Faber. 2019a. Extraction of terms for the construction of semantic frames for named bays. *Argentinian Journal of Applied Linguistics* 7(1): 27-57.

Rojas-Garcia J., and P. Faber. 2019b. Extraction of Terms Related to Named Rivers. *Languages* 4(3): 46.

Sahlgren, M., and A. Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (Texas), 975-980.

Schmidt, B., and J. Li. 2017. *wordVectors. Tools for creating and analyzing vector-space models of models of texts.* R package version 2.0.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 59-66 recibido 31-03-2019 revisado 07-05-2019 aceptado 15-05-2019

Spelling Normalisation of Basque Historical Texts

Normalización de Textos Históricos Vascos

Ainara Estarrona1[1], Izaskun Etxeberria[2], Ander Soraluze[1], Manuel Padilla[3]
[1]Centre Nationale de la Recherche (CNRS)-IKER (UMR 5478)
[2]University of the Basque Country (UPV/EHU)-Ixa group
[3]University of Pau and Pays de l'Adour (UPPA)-IKER (UMR 5478)
{ainara.estarrona, izaskun.etxeberria, ander.soraluze, manuel.padilla}@ehu.eus

Abstract: This paper presents a computational method and its evaluation in a real scenario with the aim of normalising Basque historical texts in order to be analysed using standard Natural Language Processing tools (NLP). This normalisation work is part of a more general ongoing project called *Basque in the Making (BIM): A Historical Look at a European Language Isolate*, whose main objective is the systematic and diachronic study of a number of grammatical features of the Basque language.
Keywords: Text normalisation, Digital Humanities, Finite-state technology, Historical Corpus, Diachronic Syntax

Resumen: En este artículo se presenta y evalua en un entorno real un método computacional con el objetivo de normalizar textos históricos vascos para que, una vez normalizados, puedan ser analizados con herramientas estándar de Procesamiento del Lenguaje Natural (PLN). Este trabajo de normalización forma parte de un proyecto en curso más general llamado *Basque in the Making (BIM): A Historical Look at a European Language Isolate*, cuyo objetivo principal es el estudio sistemático y diacrónico de ciertas características gramaticales de la lengua vasca.
Palabras clave: Normalización de textos, Humanidades Digitales, Tecnología de estado finito, Corpus Histórico, Sintaxis diacrónica

1 Introduction

In this paper we present a computational method and its evaluation in a real scenario with the aim of normalising Basque historical texts so that, once normalised, they can be analysed with standard NLP tools. This work is part of a more general ongoing project called *Basque in the Making (BIM): A Historical Look at a European Language Isolate*[1].

The main objective of this project consists of a systematic diachronic study of a number of grammatical features of Basque as we know it today in order to know the evolution of certain grammatical structures and to compare them with the theoretical discoveries of the current diachronic synthesis and theoretical linguistic. Hence, previous research work in an academic environment is applied and tested on a real digital library where more dialects, periods and phenomena have been analysed and evaluated.

The second stated goal is to create a comprehensive annotated historical corpus of Basque that will comprise both part-of-speech and syntactic annotation as well as a rich set of metadata structure. This way, it will be possible to search the annotated corpus by words, lemmas, grammatical categories, by sequences of grammatical categories, and by specific structural configurations (such as relatives, correlatives, finite subordination and others) as well as by publication year, author or place of publication, thanks to the aforementioned metadata. In a previous project carried out between the IKER centre[2] based in Bayonne (France) and the Ixa group[3] of the University of the Basque Country, a search interface was created to analyse the syntactic variability of the Basque language in different dialects (Uria and Etxepare, 2011). This interface is called *Basyque (Basque Syntactic Database)*[4] and we plan to use this experience now to create a search interface to analyse syntactic variability in a diachronic way. Thus, the preparation of resources for the syntactic exploration

[1]http://ixa2.si.ehu.es/bim/en

[2]http://www.iker.cnrs.fr/?lang=fr

[3]http://ixa.si.ehu.es/

[4]http://ixa2.si.ehu.es/atlas2/index.php?lang=en

ISSN 1135-5948. DOI 10.26342/2019-63-6
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

of the historical corpus in Basque is essential to perform systematic diachronic examinations on the main characteristics of Basque grammar. Taking all this account, *BIM* is an interdisciplinary project, where experts on Linguistics and Natural Language Processing take part.

As said before, in this article we will focus on the implementation and evaluation of computational techniques for corpus normalisation. We base our work on previous research carried out on the automatic normalisation of both historical and dialectal Basque texts (Etxeberria et al., 2014; Etxeberria et al., 2016).

Consequently, the methodology to be followed is already partially defined. It must be stressed that the method has been previously tested and evaluated always, however, in a experimental scenario. An important contribution of this paper is testing the method in a real application where periods and dialects that have never been analysed before are also considered, adding complexity to the normalisation task.

After having presented the project briefly, in Section 2 of the paper we will make an exposition of the related work. In Section 3 we will present the corpus with which we are going to work. Section 4 will be dedicated to the normalisation process including manual annotation and automatic normalisation. The main experimental results will be outlined and discussed in Section 5. Finally, we will review the main conclusions and preview future work in Section 6.

2 Related work

Several techniques have been used for the normalisation of historical texts and they can be roughly divided into three groups:

- Rule-based methods (hand-written phonological grammars) were the first approach, but they require a big amount of manual work.
- Unsupervised techniques: systems that work without supervision. Applying edit-distance (Levenshtein distance) or phonetic distance (by i.e., the *Soundex* algorithm) are popular solutions. Such approaches are often used as a baseline for testing new systems (Jurish, 2010).
- Machine-learning based techniques: systems that learn from examples of standard-variant pairs. These techniques are the most popular nowadays and our approach is within them.

Kestemont, Daelemans, and Pauw (2010) carry out lemmatisation on a Middle Dutch literary corpus, presenting a language-independent system that can 'learn' intra-lemma spelling variation. This work employs a novel string distance metric to better detect spelling variants. The semi-supervised system attempts to re-rank candidates suggested by the classic Levenshtein distance, leading to substantial gains in lemmatisation accuracy.

Pettersson, Megyesi, and Tiedemann (2013) treat the normalisation task as a translation problem, using character-based SMT techniques (CSMT) in the spelling normalisation process. Then, in Pettersson, Megyesi, and Nivre (2014) they evaluate and compare their approach with a memory-based filtering and a Levenshtein-based approaches considering five languages. The SMT-based approach generally works best.

Using the same approach (CSMT) Scherrer and Erjavec (2016) develop a language-independent word normalisation method and test it on a task of modernising historical Slovene words. They perform two sets of experiments: supervised and unsupervised. In the first one, they use the lexicon of word pairs as training data to build a CSMT system. In the second one, they simulate a scenario in which word pairs are not available. They show that both methods produce significantly better results than the baselines.

In our previous work (Etxeberria, Alegria, and Uria, 2019) a different approach is presented. The method learns to map phonological changes using a noisy channel model that combines weighted finite-state transducers (WFST) and language models. During evaluation our approach is compared with the CSMT methods explained in Pettersson, Megyesi, and Nivre (2014) and in Scherrer and Erjavec (2016) using same historical corpora. The results show that the WFST approach produces similar or better scores for the six languages, and it can be considered within the state-of-the-art research.

Lately, some authors have begun applying NMT models in order to analyse whether they perform better than SMT models for the historical spelling normalisation task. Thus,

in (Bollmann and Søgaard, 2016) they explore the suitability of a deep bi-LSTM neural network architecture applied on a character level, and show that it outperforms the baselines (conditional random fields and Norma tool by Bollmann et al. (2012)).

In a similar way, eight different NMT models are applied in (Tang et al., 2018) to the spelling normalisation task in five languages, English, German, Hungarian, Icelandic, and Swedish. Using word accuracy as the evaluation metric, NMT models perform better than SMT models for four languages, but when CER is used, all the NMT models perform better than SMT models for all the languages. The authors carry out particular experimentation on Swedish by increasing the size of the training set and they conclude that the performance of NMT models is highly related to this size.

In any case, for the moment we have ruled out neural methods due to the features mentioned of the BIM corpus: several historical periods, several dialects and short texts. Further research is necessary for a robust solution based on neural methods.

3 The corpus

The corpus covers the most representative written production between the 15th and the 18th century, a time span within which all historical dialects of Basque are represented. Texts have been selected on the basis of: 1) their representativeness; 2) the existence of reliable editions; and 3) their social context. To compile such a corpus, we have made use of some ancient texts that are available in the *Klasikoen Gordailua* website[5], which covers almost all the classic Basque texts over the last centuries. It is a corpus comprising more than 12 million words and therefore, it is therefore a solid base for the analysis of the development of the Basque dialects. The texts belong to different literary genres: narrative, poetry and verses. But we will focus on prose texts, considering that the syntactic structures we are interested in may be different from one genre to another. As the texts are classified according to time, language and literary genre, it will possible to take into account such parameters when searching for data. Besides, the corpus is transcribed and digitised in .pdf and .rtf formats, which makes it easy to access them for work.

However, as we know that some errors have occurred in the transcription process of such corpora, we consider the revision of the corpus is essential, and we have therefore carried out an important work of philological revision of the texts comparing them with their facsimiles. In this way, we have updated the spelling of the texts, while maintaining the phonological characteristics of each of them.

In terms of the corpus size, our goal is to analyse as many texts as possible, as a robust annotated corpus is necessary to achieve significant results. For the moment, we foresee starting with a 750.000 word reference corpus, expanding from the 15th century to the mid 18th century, the period encompassing Archaic and Old Basque. We have already finished the selection of those critical editions we will work on. When working with texts from different time periods and places, we are compiling a very rich corpus in respect to its quality. However, the diversity of dialects and varieties adds complexity to the normalisation task.

In this first step, relevant old works from different dialects and literary genres have been selected in order to be able to identify from the outset the challenges we will have to face during the process of labeling the entire corpus.

1. Joanes Leizarraga, *Jesus Krist Gure Jaunaren Testamentu Berria*, 1571.
2. Bernard Etxepare, *Linguae Vasconum Primitiae*, 1545.
3. *Refranes y Sentencias*, 1596 (anonymous).

The first work, Leizarraga, is a translation of the New Testament written basically in Lapurdian dialect. It consists of around 74,000 words and is undoubtedly the most extensive work documented in Basque until the 17th century.

The second work, Etxepare, is considered to be the first documented book in the Basque language. These are poems written in the Lower Navarrese dialect and consist of around 7,000 words.

Finally, the third work chosen, RS, is a compendium of sayings from the end of the 16th century written in an archaic Biscayan dialect and contains around 3,000 words.

[5] http://klasikoak.armiarma.eus/

4 *Text normalisation*

Text normalisation is a necessary step in this project. Having developed a lot of NLP based technology for the analysis of standard Basque, we have now to tackle the text normalisation, considering that once ancient or dialectal texts are standardised, the developed NLP tools could be applied for the linguistic analysis of the corpora.

The process of text normalisation will be carried out in 2 phases. First, part of the corpus will be annotated manually, and then, the rest of the corpus will be normalised automatically.

4.1 Manual annotation

In previous work on the standardisation of historical Basque texts (Etxeberria, 2016), it has been demonstrated that manual annotation of 10% of the text is sufficient to obtain satisfactory results in automatic normalisation. Therefore, we have randomly collected and processed 10% of each of the 3 texts selected to proceed to the manual normalisation.

The processing of the text consists of 3 steps: tokenization, named entity recognition and lexical recognition. After these 3 steps, each word in the text will be marked with the corresponding labels (see Figure 1):

- ENT-Zuz: 'Correct entity', when the named entity recognition (Alegria et al., 2006) identifies a proper name (green colour).
- STD-Zuz: 'Correct standard', when the morphological analyser of standard Basque (Alegria et al., 1996) identifies the word or lemma (fuchsia colour).
- OOV: 'Out of vocabulary', when the morphological analyser of standard Basque does not identify the word or lemma (blue color).

The manual annotation is done using the *BRAT* tool[6] and the task is carried out by a single linguistic annotator with training in historical texts. Figure (1) shows the interface for the manual annotation.

The task of the annotator is to assign to each word of the text its corresponding standard form according to the dictionary of the

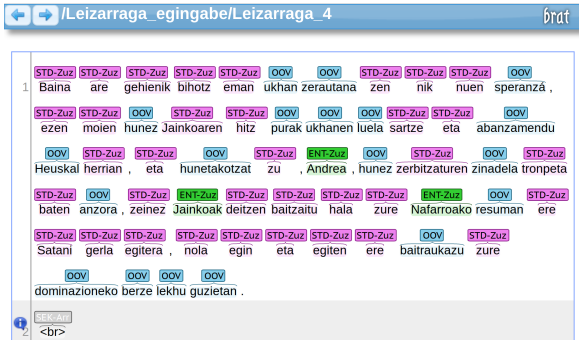


Figure 1: Interface for the manual annotation of the corpus in which a fragment of Leizarraga's work can be seen

Royal Academy of the Basque Language[7]. Thus, the goal of manual standardisation is to remove all OOV tags (in blue) based on a classification which is used to specify the type of variation. The classification of variants is represented using the following labels:

- OOV-Ald: 'Variant', when the annotator assigns a standard variant to the non-standard form (yellow colour).
- OOV-Zuz: 'Correct', when the morphological analyser does not recognise the word because it is archaic or exclusive of historical texts, but the annotator wants to point out that it is correct (fuchsia colour).
- STD-Ald: 'Variant', when an old word coincidentally looks like a modern Basque word ('false friends', yellow colour)[8].

In addition to this classification of variants, we also have the SEK-Ber label which means 'special section' (grey colour). This label is used to mark a paragraph where the annotator has found some interesting historical morphosyntactic phenomenon for subsequent morphonsyntactic analysis.

It should be stressed that in previous research work (Etxeberria et al., 2016) only Out-of-Vocabulary (OOV) words were revised and normalised. However, taking into

[6] http://brat.nlplab.org/index.html

[7] https://www.euskaltzaindia.eus/hizkuntza-baliabideak

[8] We have also added a label to differentiate variants related to verbs or verbal auxiliaries (OOV-AldADIJ) because we believe it may be interesting to have these types of variants identified for future steps during the normalisation and morphosyntactic analysis process.

account that in this project we work for a real scenario, the annotator will have to check all the words one by one, also those that have the STD-Zuz label, since sometimes old words coincidentally look like modern Basque words ('false friends'), which adds an obvious complexity to the task.

One way to assess the complexity of the task in each work may be to measure the number of standard words and variants in each of them. Table 1 presents the total number of words of each of the works treated in this paper and described in Section 3, as well as the number of words that have the labels STD-Zuz, OOV and ENT-zuz in each of them (with their percentages). Finally, it shows the number of words manually annotated in each work.

	Leizarra	Etxepare	RS
Words	73,610	6,860	3,083
STD-Zuz	50,321	4,416	1,709
%	68.36%	64.37%	55.43%
OOV	20,924	2,403	1,352
%	**28.42%**	35.03%	**43.85%**
ENT-Zuz	2,365	41	22
%	3.21%	0.60%	0.71%
Manually	7,548	694	3,083

Table 1: Total number of words for each work, total number of words with STD-Zuz, OOV and ENT-Zuz labels and their percentages, and total number of words manually annotated

Thus, we can foresee that normalising *RS* (55% of standard forms) will be a much more difficult task than normalising Leizarraga's work (68%), as we will see in Section 5.

Finally, the time needed to manually annotate the whole corpus has been estimated and it can be seen that about 143 words per hour are labeled. Therefore, a person would need about 3 years to label the entire corpus manually and this is obviously why we plan to implement computational techniques to normalise the corpus. Once 10% of the text is manually normalised, the rest will be automatically normalised using statistical learning methods, which we will explain in detail in the following Section.

4.2 Automatic normalisation

All the word pairs in the learning set are taken into account for the learning process: some of them are labeled as *Zuz* (correct) and other as *Ald* (variant), but all of them are going to be used in the learning process. However, we have prepared two different lists of word pairs for the test set: the first one contains all the word pairs in the test set (correct or variant) but the second one contains only the word pairs labeled as variants.

We have reapplied the method previously cited in (Etxeberria et al., 2014). This method uses *Phonetisaurus*[9], a Weighted Finite State Transducer (WFST) driven phonology tool (Novak, Minematsu, and Hirose, 2012).

After collecting the word pairs (historical–standard) in the learning set into a dictionary, the application of the tool includes three major steps:

1. Sequence alignment. In order to align historical words and their equivalent standard words (not words and pronunciations), the results of the alignment process are joint grapheme/grapheme chunks that we use in the next step.
2. Model training. A joint n-gram language model is trained using the aligned data and then converted into a WFST (after some tuning, a joint 7-gram model was generated).
3. Decoding. The default decoder used in the WFST-based approach finds the best hypothesis for the input words, given the WFST obtained in the previous step. It is also possible to extract the k-best output hypotheses for each word.

Phonetisaurus tool has been used to learn the changes that occur within the word pairs in the learning set, which by itself produces a grapheme-to-grapheme system.

Once this model is trained and converted to a WFST format, it can be used to generate correspondences between previously unseen words and modern standard forms. When multiple possibilities for a corresponding historical variant exist, some filtering becomes necessary and the quality of this filtering becomes very important to improve the results. Based on the tuning process carried out in our previous work (Etxeberria et al., 2016), the number of candidates is set to 5.

[9] `https://github.com/AdolfVonKleist/Phonetisaurus`

Regarding the filtering process of the generated candidates, the first filter is obvious: the transductions that do not correspond to any accepted standard word form are eliminated. For selecting standard words, a morphological analyzer of Basque is used (Alegria et al., 2009). From the remaining candidates, if there are any, the most probable transduction according to *Phonetisaurus*'s weight model is selected. If there is no standard candidate for a given input, we have tried different options: 1) giving the input word (historical) as output (standard); 2) the most probable transduction according to *Phonetisaurus*'s weight model is selected.

Another important option could be taking into account the memorised word pairs. There are exceptions, but a word is usually always standardised in the same form, therefore it could be a good idea to take advantage of the learned word pairs.

4.3 Evaluation

Once we have part of the text manually annotated and the method applied, we evaluate the quality of the automatic normalisation method we propose. We have carried out 10-fold cross-validation experiments on the manually annotated text.

Results are given based upon precision, recall and F-score IR measures.

5 Results and discussion

We have started with Leizarraga's work and we have used 10% of the text manually annotated for evaluation. Different experiments have been carried out based on the filtering process explained in Table 2 (Section 4.2). In the first one, we have evaluated the results when *Phonetisaurus* does not give an answer if it does not find a standard option among the first 5 options it offers (silence). In this case, the precision was very high (0.9630). Using information about previously seen examples (memory) recall rises 3 points and the F-score almost 2.

In any case, the goal for the project is that the system always offers an answer, therefore we have carried out two other experiments. When the system does not find a standard form among the first 5 options, it gives as output the same input (out=input) or it gives as an output the first of them (first WFST).

The last experiment consists of combining memory and the first proposal of the WFST

Filter	P	R	F
Silence	0.9630	0.8784	0.9187
Memory	**0.9695**	0.9070	0.9372
Out=Input	0.9245	0.9245	0.9245
First WFST	0.9368	0.9368	0.9368
Memory+First	0.9424	**0.9424**	**0.9424**

Table 2: Results for Leizarraga's work. When *Phonetisaurus* does not find a standard option, four approaches are tested: silence (no output), memory (output if it has been seen in the training set), echo of the input and first proposal of the normalizator.

(2nd. and 4th. options) obtaining the best recall and F-score (memory + first). This option has been chosen for the all the texts of the project and results obtained for the three texts can be seen in Table 3. P, R and F have the same value when always one option is proved, so, only one figure is shown for each one. We have added in the table the results obtained if we only consider the variants with the intention of really seeing how the system behaves in the most difficult cases and pointing out the difficulty of the task[10].

After fixing the filter, the next work was that of Etxepare. Taking into account that the dialects of Leizarraga and Etxepare are close, we tried to use the same system trained with Leizarraga to normalise Etxepare and evaluated with the manually annotated sample. The results were not as expected with an F-Score of 0.7516 taking into account all the words, and only an F-score of 0.4617 for the variants. Using only the sample manually annotated the results improved (F-Score 0.5958 for the variants) but were still clearly insufficient, therefore we proposed 2 possible solutions: i) to annotate manually more text, and ii) to use for learning both Leizarraga and Etxepare. We decided to combine both sources and the results improved considerably as can be seen in Table 3 (2nd. column).

This suggests that in the future, instead of having a system for each work, we could implement a system for each dialect or dialects that are linguistically proximate, including for learning what is learned in all the works that belong to that dialect or group of dialects.

[10]The results decrease when evaluating only the variants, but it is necessary to bear in mind that the real scenario is the one in which the system normalises all the text, variants and standard forms included.

The last work was *RS*. This is a special work because it reflects a variety of language very far from the standard and very archaic. The first results using the manually annotated sample of 10% were clearly insufficient with an F-score less than 0.50. The work has about 3,000 words, therefore we thought it was convenient to make an effort and manually annotate it in its entirely. The results obtained using cross-validation can be seen in Table 3 (3rd. column).

	Leizar.	Etxep.	RS
Test-variants	0.8768	0.7038	0.6683
Test-all	0.9424	0.8646	0.8066

Table 3: Accuracy for *Leizarraga*, *Etxepare* and *RS* taking into account all words or only variants

The results clearly show the difficulty of the task in the case of works so distant from the standard variety of the language. As we have shown in Table 1 (Section 4.1), Leizarraga's work is the one that has more standard words (68%) and in which we obtain better results, followed by Etxepare's work (64%) and, finally, the worst results are for *RS*, which is the work farthest from the standard variety, with only 55% of standard words.

Results for *RS* are worse than expected in spite of annotating all the text. This is an important finding for the project and has led us to decide that sometimes the whole text should be manually annotated (luckily only for short texts).

6 Conclusions and future work

We have presented and evaluated in a real scenario a normalisation method for Basque historical texts for several historical periods and several dialects within the framework of a more general project of creating a syntactically annotated historical corpus of the Basque language.

Our phonological induction inspired method has been evaluated on three works of the 16th century selected from the corpus. The method achieves F-scores between 80% and 95% in the task. We believe our method is a good solution using a limited amount of supervision to achieve acceptable results without significant manual annotation effort.

We are aware of the limitations of our corpus, in terms of its extension, in order to achieve adequate results using methods based on neural networks. Nevertheless, Moeller et al. (2019) have tried different strategies to deal with these types of limitations and we believe that it would be an interesting line of investigation in the case of the Basque language.

It is important to underline that the experiments have been carried out on a real historical digital library, overcoming the limits of some academic repositories. Combining historical periods, dialects and scarcity of data, we have concluded that the results can be, in some cases, worse than expected (*RS* corpus).

In addition to the quantitative evaluation, we intend to carry out an error analysis in order to detect the weak points of our system and be able to search for strategies aimed at improving the results. A preliminary error analysis in the Leizarraga's work have been carried out concluding that an important part of the errors correspond to the system of auxiliary verbs. Something that we already foresaw, since the system of auxiliary verbs in historical Basque and its different dialects is especially complicated.

After normalising the corpus we will proceed to the automatic morphosyntactic analysis. This automatic analysis will be revised manually to detect errors and to proceed to the labeling of interesting morphosyntactic phenomena from the point of view of diachronic syntax. This syntactically annotated corpus will facilitate the systematic study of a number of grammatical features of the Basque in a diachronic way by means of a search interface, on which we are already working, and it will be the only tool of these characteristics existing for the Basque language. The annotated corpus and the search interface will be public and freely available to the research community.

Acknowledgments

We are indebted to Josef Novak for his useful help concerning Phonetisaurus, and to Jordi Porta, Eva Petterson, Yves Scherrer and Tomaž Erjavec for providing their corpora and for addressing our inquiries about their experiments. Grateful thanks to the three anonymous referees for helping us improving the paper. The research leading

to these results was carried out as part of the *BIM* project (Agence Nationale de la Recherche, France) and the *BERBAOLA* project (Basque Government funding, Elkartek KK-2017/00043).

References

Alegria, I., O. Arregi, N. Ezeiza, and I. Fernández. 2006. Lessons from the development of a named entity recognizer for basque. *Procesamiento del lenguaje natural*, 36.

Alegria, I., X. Artola, K. Sarasola, and M. Urkia. 1996. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203.

Alegria, I., I. Etxeberria, M. Hulden, and M. Maritxalar. 2009. Porting basque morphological grammars to foma, an open-source tool. In *Finite-State Methods and Natural Language Processing*. Springer, pages 105–113.

Bollmann, M., S. Dipper, J. Krasselt, and F. Petran. 2012. Manual and semi-automatic normalization of historical spelling-case studies from early new high german. In *KONVENS*, pages 342–350.

Bollmann, M. and A. Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. *arXiv preprint arXiv:1610.07844*.

Etxeberria, I. 2016. *Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz*. Ph.D. thesis, Universidad del Pais Vasco / Euskal Herriko Unibertsiatea.

Etxeberria, I., I. Alegria, M. Hulden, and L. Uria. 2014. Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52:13–20.

Etxeberria, I., I. Alegria, and L. Uria. 2019. Weighted finite-state transducers for normalization of historical texts. *Natural Language Engineering*, 25(2):307–321.

Etxeberria, I., I. Alegria, L. Uria, and M. Hulden. 2016. Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1064–1069.

Jurish, B. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77.

Kestemont, M., W. Daelemans, and G. D. Pauw. 2010. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301.

Moeller, S., G. Kazeminejad, A. Cowell, and M. Hulden. 2019. Improving low-resource morphological learning with intermediate forms from finite state transducers. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 81–86.

Novak, J. R., N. Minematsu, and K. Hirose. 2012. WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49.

Pettersson, E., B. Megyesi, and J. Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. *Proceedings of LaTeCH*, pages 32–41.

Pettersson, E., B. Megyesi, and J. Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, volume 18, pages 54–69.

Scherrer, Y. and T. Erjavec. 2016. Modernising historical Slovene words. *Natural Language Engineering*, 22(6):881–905.

Tang, G., F. Cap, E. Pettersson, and J. Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. *arXiv preprint arXiv:1806.05210*.

Uria, L. and R. Etxepare. 2011. Basyque: Aplicación para el estudio de la variación sintáctica. *Linguamática*, 3(1):35–44.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 67-74 recibido 31-03-2019 revisado 14-05-2019 aceptado 17-05-2019

Automatic proficiency classification in L2 Portuguese

Clasificación automática del nivel de proficiencia en Portugués Segunda Lengua

Iria del Río
University of Lisbon, Center of Linguistics-CLUL
igayo@letras.ulisboa.pt

Abstract: We present the first experiments on automatic proficiency classification for L2 Portuguese. For the experiments, we take advantage of a new version of the NLI-PT dataset, a compilation of L2 Portuguese texts written by learners. We use supervised learning and we approach the task as a classification problem, using the CEFR scale. Different linguistic features are tested, combined with different algorithms. With the best model, we get an accuracy of 72%, a result in line with previous experiments with other languages.
Keywords: Proficiency level, CEFR, L2 Portuguese, Supervised Learning

Resumen: Este trabajo presenta los primeros experimentos en clasificación automática del nivel de proficiencia en Portugués Segunda Lengua (L2). En los experimentos se usa una nueva versión del dataset NLI-PT, una compilación de textos escritos por estudiantes de Portugués L2. La tarea se aborda con aprendizaje supervisado, y se concibe como un problema de clasificación, usando la escala del MCER. Diferentes características lingüísticas son analizadas, así como diferentes algoritmos. Con el mejor modelo hemos obtenido una exactitud del 72%, un resultado en línea con previos experimentos realizados con otras lenguas.
Palabras clave: Nivel de proeficiencia, MCER, Portugués Segunda Lengua, Aprendizaje Supervisado

1 *Introduction*

This work has two main contributions. First, we present a larger and better version of the NLI-PT dataset[1], a compilation of L2 Portuguese texts with different types of linguistic annotations. Secondly, we describe the first experiments in automatic proficiency classification for L2 Portuguese, where we got similar results to previous works on the field.

The availability of data with linguistic annotations benefits different types of research, from theoretical analysis to statistical approaches like Machine Learning. Learner data is particularly difficult to gather, because of the specific context where this data is produced. For the English language there are big collections of learner data available, like the Cambridge Learner Corpus (16 millions of words) (Nicholls, 1999), but such type of collections are not common for other languages. The NLI-PT dataset aims to solve this gap for European Portuguese. We present a bigger and improved version, with more texts, better annotations and a different and more intuitive organization of the data.

As an example of the usefulness of the dataset, we present the first experiment for automatic proficiency classification of L2 Portuguese. Proficiency classification is a common task in second language learning. The development of the learner is usually defined in relation to a specific scale with different levels of linguistic complexity. One of the most common scales is the one described in the Common European Framework of Reference for Languages (CEFR) (Europe et al., 2009). The CEFR defines 3 broad divisions: A, basic user; B, independent user; C, proficient user, which are subdivided into 6 devel-

[1] http://www.clul.ulisboa.pt/en/resources-en/11-resources/894-nli-pt-a-portuguese-native-language-identification-dataset

ISSN 1135-5948. DOI 10.26342/2019-63-7
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

opment levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Each level is related to specific linguistic features and skills, establishing a progression from a very rudimentary language to a performance close to a native production. In this context, it is common that learners of a second language perform placement tests that define their proficiency level. The interest of an automatic system that can perform this task is, therefore, evident.

Automatic proficiency classification is commonly considered as a type of Automatic Essay Scoring (AES) task. AES systems are primarily developed for English (Burstein, 2003; Burstein and Chodorow, 2012; Yannakoudakis and Loo 2013), but in recent years systems for other languages have began to emerge (Vajjala and Loo, 2013). AES has been modeled in different ways, as a regres-sion (Yannakoudakis, Briscoe, and Medlock, 2011), ranking (Taghipour and Ng, 2016) or a classification problem (Pil´an, Vajjala, and Volodina, 2016). The features used are di-verse, from Bag-of-words (BOW) to more ab-stract representations that use higher levels of linguistic information (morphological, syn-tactic or even discursive). It is also very com-mon the use of descriptive metrics of the text related to word or sentence length, like av-erage syllable length, which have been con-nected to proficiency development in the area of Second Language Acquisition (SLA)(Lu, 2012). In general, AES is seen as a monolin-gual task, but recent works like (Vajjala and Rama, 2018) have explored multi and cross-lingual approaches.

Usually, the term AES is used as a general term for referring to different tasks: from proficiency classification of learner texts to readability assessment of teaching materials (Pilán, Vajjala, and Volodina, 2016). In fact, for the Portuguese language, and to the best of our knowledge, AES works have focused only on readability assessment (Branco et al., 2014) and (Curto, Mamede, and Baptista., 2015). We would like to differentiate the nature of these tasks, that is, readability assessment of input materials for the student, and proficiency classification of learners' texts because the linguistic parameters they involve are different. Readability assessment tasks usually focus on evaluating the linguistic complexity of potential input materials for the students. The goal is to automatically select materials that are appropriate for the learner. Therefore, systems working on readability assessment are trained with materials designed for learners and not written by them. Those materials use linguistic variants that are close to the target-native language, in some cases (for students with an elementary knowledge of the L2), they are simplified versions of this target-native language. However, these texts do not present the linguistic distinctive features that we can find in learner productions like, for example, orthographic or morphological errors or anomalous lexical or syntactic constructions influenced by the L1. These distinctive linguistic features constitute a challenge for automatically processing L2 texts since NLP tools are commonly built using native language models.

In our experiments, we apply the main three levels of the CEFR scale, A, B, and C, to automatically classify L2 Portuguese texts. Moreover, we try to answer the following question: What does it define the proficiency level of an L2 Portuguese text? To find an answer, we applied supervised machine learning techniques to build a classification model; we tested different algorithms and representations of the texts, from BOW to models that use descriptive features commonly used in the area and in SLA research (like average word length). With the best model, we got an accuracy of 72%, a result similar to those obtained in previous works. But, what is more relevant, we gained meaningful insights about the relation of certain linguistic features and automatic proficiency classification in L2 Portuguese.

The paper has the following structure: in section two we present related work; the NLI-PT dataset and the features used in the experiments are presented in section three; section four focuses on the description of our methodology and the results of the experiments; finally, in section five, we present our main conclusions and some future work.

2 Related Work

In this section, we present two types of previous work: SLA studies that have analyzed the relation between certain features and proficiency levels, and approaches that have used machine learning techniques to predict learner proficiency using the CEFR scale.

(Lu, 2012) analyses in detail the relation-

ship between proficiency in L2 English and several lexical dimensions, concluding that the features linked to lexical variation (like Type-Token ratio) are the most correlated to the quality of an L2 essay. Several features identified as relevant in this work have been used by automatic approaches afterward. (Kyle and A. Crossley, 2014) explored lexical features too and showed that 47.5% of the variance in holistic scores of lexical proficiency in English as second language can be explained using a range of lexical sophistication indices. Other characteristics like syntactic complexity or error patterns have been studied too, mainly for English (Tono, 2000),(Lu, 2012), (Vyatkina, 2012), but also for other languages (Gyllstad et al., 2014).

Concerning automatic proficiency classification, (Yannakoudakis et al., 2018) is one of the most recent works for the English language. The authors used a subset of the Cambridge Learner Corpus with human proficiency annotations (levels A1 to C2), containing a total of 2,312 texts. They model the task as a ranking function and evaluate the quality of the predicted score by calculating Pearson's product-moment and Spearman's rank correlation coefficient against the scores assigned by a human expert. The features used include character sequences, POS, hybrid word, and POS sequences, phrase structure rules and error rates. The best model gets a Pearson r of 0.765 and a Spearman ρ of 0.773, with a κ of 0.738 (the standard error is 0.026) that indicates high agreement between the predicted CEFR scores and those assigned by humans.

In another recent study, (Vajjala and Rama, 2018) present the first multi and cross-lingual approach for proficiency classification. The authors use 2,286 manually graded texts (five levels, A1 to C1) from the MERLIN learner corpus (Boyd et al., 2014). It is an unbalanced dataset, with the following distribution of learner texts: German, 1,029 texts; Italian, 803 texts, and Czech, 434 texts. The authors compare different algorithms: logistic regression, random forests, multi-layer perceptron, and support vector machines for experiments with non-embedding features, and Neural Network models trained on task-specific embedding representations for other experiments. For non-embedding features, the best algorithm is Random Forests in most of the scenarios. They use a wide range of features: word and POS n-grams; task-specific word and character embeddings trained through a softmax layer; dependency n-grams (not used before); domain features mainly linked to lexical aspects (Lu, 2012); and error features. In their experiments, monolingual and multilingual models achieve similar performance, and cross-lingual classification yields lower, but comparable results to monolingual classification. For monolingual experiments, the best result (F1-score) is achieved with word n-grams plus domain features (German=0.686; Italian 0.837; Czech= 0.734). For multilingual tests, the best result is 0.726 with POS n-grams and information of the L1 as a feature. In cross-lingual experiments, they use German texts for training, and they get a 0.653 F1 score for Czech using dependency n-grams, and a 0.758 for Italian using POS n-grams. We can see that the features that allow for the best results vary in the experiments: words n-grams in monolingual; POS n-grams in multilingual; dependency n-grams for Czech and POS for Italian in cross-lingual experiments.

(Vajjala and Lõo, 2014) perform proficiency classification for Estonian. They use a corpus with 879 texts, belonging to four proficiency levels (A2 to C1) and also a balanced version of this dataset with 92 texts per category. They compare classification and regression models and use a rich feature set (78 features) that considers the morphological complexity of Estonian, as well as lexical richness features inspired by (Lu, 2012). Interestingly, POS models achieved poor performance and were not considered in the feature set. The best model is classification, with an accuracy of 79% in the whole dataset and 76.9% in the balanced one. For both datasets, the category with the poorest performance is B2. The authors perform a feature analysis and show that with the 27 best features they achieve a performance of 78.3%. The 10 best features in this group of 27 are lexical (Corrected Type Token Ration, Squared Verb Variation) and morphological (2nd person inflected verbs, distinct cases used in the document).

3 Dataset

3.1 Corpus

For our experiments, we use an updated version of the NLI-PT dataset (del Río,

Zampieri, and Malmasi, 2018). The goal of this resource is to make available annotated data produced by L2 Portuguese learners. NLI-PT was originally created for running Native Language Identification (NLI) experiments, and it contains written texts compiled from different learner corpora of L2 Portuguese. Those texts are presented in a clean TXT version, together with versions annotated at two linguistic levels: morphological (POS) and syntactic. The annotation of the dataset was performed with freely available tools. For POS there is a simple POS representation, that is, only type of word, and a fine-grained POS, which is the type of word plus its morphological features. The annotations were performed using the LX Parser (Silva et al., 2010) for the simple POS and the Portuguese morphological module of Freeling (Padró and Stanilovsky, 2012) for detailed POS. Concerning syntactic annotations, NLI-PT includes constituency (from LX Parser) and dependency (DepPattern toolkit (Otero and González, 2012)) annotations.

The new version of the dataset is bigger and contains several improvements. We have corrected some tokenization issues and improved the constituency annotations. Besides this, we have enlarged the dataset with texts from the CAL2 learner corpus [2] (930 new texts). Additionally, we have modified the structure of the dataset. In the new version, the name of the files contains three types of information: the source corpus, the L1 and the proficiency level of the text. For example, for the file "ara_A_008CVETF_cop.txt", the prefix *ara* corresponds to the native language, Arabic; the *A* corresponds to the CFR proficiency level and the suffix *cop* refers to the source of the file, the COPLE2 corpus (Mendes et al., 2016). The CEFR proficiency levels considered in the original learner corpora are not the same: two corpora consider five levels, A1 to C1, while other two consider only three major levels, A, B and C. For this reason, in NLI-PT we have homogenized the levels to three: A, B, and C. The final dataset contains a total of 3,069 texts, corresponding to 15 different native languages. The distribution by proficiency level is presented in Table 1.

As we can see, the distribution of texts by proficiency level is not balanced. For this reason, in our experiments we have used two different datasets: one containing the whole dataset and one with a balanced distribution of 466 texts per class. For the experiments, we split both corpora in training (80%) and testing (20%) sets.

Proficiency	**Number of Texts**
A - Beginner	1,388
B- Intermediate	1,215
C- Advanced	466
Total	**3,069**

Table 1: Distribution of texts by CEFR proficiency level in the NLI-PT dataset

3.2 Features

We were interested in investigating the impact of different linguistic features in the classification task. For this reason, we have tested different types of features extracted from NLI-PT:

1. **Bag of words**, with different variations: using the word form, tokens and lemmas. We performed some initial experiments with the training set to check which representation produced the best results. We got similar results for word forms and tokens, being the word form representation slightly better. For this reason, for further experiments we kept only the word form representation.

2. **POS n-grams**: we used the fine-grained POS representation of NLI-PT, which contains the main POS and also morphological information, like gender or number. We consider that this information could be especially interesting because Portuguese has a rich morphology, and this feature is problematic for certain learners, especially at the initial stages. Agreement errors like *aréia branco* (*white*-MasculineSingular *sand*-FemininesSingular) can be captured with a POS n-gram representation, and we wanted to measure the impact of this feature. We evaluated n-grams of different sizes in the experiments.

3. **Dependency triplets n-grams**: we extracted dependency triplets with the form *head, relation, dependent* generated with DepPattern. Dependency re-

[2] http://clunl.fcsh.unl.pt/en/online-resources/corpus-de-aquisicao-de-l2/

lations are not common in proficiency classification, and we were interested in checking their impact. We also evaluated different types of sizes for the dependency n-grams.

4. **Descriptive and lexical features of the text**: set of 39 features that the studies of SLA have proved as linked with proficiency. Those features are not present in NLI-PT, and therefore we extracted them using the software Pylinguistics (Woloszyn et al., 2016). The features include different types of measures:
   - **Lexical features**: number of nouns, number of verbs, number of connectives, lexical diversity, content diversity...
   - **Descriptive measures**: average syllables per word, syllable count, word count, etc. We also used the Portuguese adaptation of the Flesch reading index (Martins et al., 1996).

4 Experiments

As we have seen, the task of proficiency evaluation can be considered as a classification or a regression problem, depending on the way we consider the proficiency levels, that is, as discrete or continuous scales. For this first attempt, we explore the task as classification, considering that this model obtained better results in previous works (Vajjala and Lõo, 2014).

We used the scikit-learn package (Pedregosa et al., 2011) for training and testing the models and for feature selection. We divided both datasets into training and test sets. We performed some initial tests for feature selection (see above) and for evaluating different algorithms. In these previous experiments, we performed 10-fold cross-validation with the training set and the different sets of features, and we trained a different classifier for each type of features to support a comparison of them. We evaluated Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Random Forests, and LogitBoost. In general, we had the best results with three algorithms: Logistic Regression (LR), Random Forests (RF) and LogitBoost (LB). For this reason, we only used the models generated with these three algorithms against the test set.

We employed accuracy as the main measure to evaluate the performance of our trained models. We also report weighted-F1 score because the whole dataset is unbalanced. Weighted-F1 score is computed as the weighted average of the F1 score for each label, taking label support (i.e., number of instances for each label in the data) into account. As a baseline, we used text length, extracted with Pylinguistics.

4.1 Results and Discussion

Due to space restrictions, we report only the best-performing systems for each combination of features.[3]

Features	Accuracy	F1
Baseline_LR	0.58	0.54
BOW_LB	**0.70**	**0.7**
POS_LB	**0.66**	**0.65**
Dep_RF	0.64	0.59
Desc_RF	0.63	0.59
ALL(noBOW)_RF	**0.67**	**0.66**
ALL_LR	**0.72**	**0.71**

Table 2: Results for the whole dataset

Features	Accuracy	F1
Baseline_RF	0.48	0.46
BOW_LB	**0.66**	**0.67**
POS_RF	**0.63**	**0.63**
Dep_RF	0.54	0.53
Descriptive_RF	0.57	0.57
ALL(noBOW)_RF	**0.59**	**0.58**
ALL_RF	**0.65**	**0.64**

Table 3: Results for the balanced dataset

Features	A-F1	B-F1	C-F1
Baseline_LR	0.67	0.58	0
BOW_LB	**0.8**	**0.70**	**0.43**
POS_LB	**0.77**	**0.66**	**0.28**
Dep_RF	0.74	0.64	0.02
Desc_RF	0.72	0.63	0.13
ALL(noBOW)_LR	**0.77**	**0.67**	**0.3**
ALL_LR	**0.8**	**0.72**	**0.42**

Table 4: Results per class for the whole dataset

[3]For each set of features, the abbreviation after the underscore indicates the name of the algorithm employed: LR for Logistic Regression; RF for Random Forests; LB for LogitBoost.

Features	A-F1	B-F1	C-F1
Baseline_RF	0.62	0.45	0.31
BOW_LB	**0.75**	**0.62**	**0.63**
POS_RF	**0.73**	**0.57**	**0.58**
Dep_RF	0.7	0.47	0.43
Desc_RF	0.64	0.52	0.53
ALL(noBOW)_RF	**0.7**	**0.52**	**0.53**
ALL_RF	**0.71**	**0.6**	**0.59**

Table 5: Results per class for the balanced dataset

For all the models, the results obtained are better in the whole dataset than in the balanced one. The best result we got is **0.72** accuracy using an ensemble combination of all the features (ALL) with LR, although this value is very close to the one using a BOW representation, 0.7. Interestingly, in the balanced dataset the ensemble combination with all the features had worse results than the best model, BOW_LB, which uses only one feature. For the ensemble combination that does not use lexical information (it does not include the word forms), POS+Dep+Desc., the results are slightly better than the POS n-gram representation for the whole dataset and worse for the balanced dataset. Both results seem to indicate that adding more linguistic features to the best single-feature models (BOW and POS n-grams) implies only a small gain for the whole dataset and a drop in accuracy for the balanced dataset. In this case, simpler models work generally better.

If we compare the models that use only one type of feature, the best results are for the BOW representation in both datasets, followed by the POS n-gram representation. One of the reasons why the BOW representation captures better the proficiency can be the fact that it keeps the information concerning orthographic problems. A comparison between the results for each type of feature shows similar behavior for both datasets, with the only difference that the Descriptive feature set performs better than the Dep one for the balanced dataset. The algorithms with the best results differ among datasets: LR for the whole dataset; RF for the balanced one.

Concerning the results by class, we can see clear differences between datasets. In the whole corpus, the C class (the one with fewer texts) performs clearly worse than the other two, being the best result 0.43 (BOW_LB). In the balanced dataset, the F1 score is more equalized between classes, being the results for the B and the C classes pretty similar. In fact, in general, the C class gets better results than the B class. For both datasets and for all models, the best results are always for the A class. This pattern suggests that A texts exhibit certain specific features that make them easy to identify, in comparison with the other two levels.

Due to the lack of space, we cannot include the confusion matrix for each model, therefore, we include table 6 as a reference. False negatives for A are more frequent in the adjacent class, B, and the same happens with the B class, where more texts are classified as C than as A. For the C class, false negatives are more frequent in the previous class, B. This picture seems to show the expected progression of the learners as the proficiency level increases. Interestingly, B texts are more often confused with the adjacent class, instead of with the previous one.

	A	B	C
A	66	18	10
B	12	60	21
C	12	29	52

Table 6: Confusion matrix for the best model (Words_LB) in the balanced dataset

5 *Conclusions and future work*

We present a new improved version of the NLI-PT dataset, and we use it to perform the first experiments on proficiency classification for L2 Portuguese, using the CEFR scale. We modeled the task as classification, and, with the best model, we obtained an accuracy of 72%.

We were interested in answering the question: What defines the proficiency level of an L2 Portuguese text? With this goal, we tested the contribution of different linguistic features combined with different algorithms to the classification task. Additionally, we wanted to test the influence of the distribution of texts by class, and therefore we used two datasets: the whole NLI-PT corpus and a balanced set. We found that an ensemble model, with all the features combined, had the best accuracy (for the whole dataset), but

also that a BOW model achieves a very similar performance (70% accuracy) in both corpora. A POS n-gram model, that does not use lexical information, gets a close result, 66%. This finding is particularly interesting because a POS n-gram model is a more abstract representation that can be less biased by topic or task variables and that can be applied to other L2 Portuguese corpora or even to other similar languages, like Spanish. The two ensemble combinations of features, one with all the features and the other without BOW, get slightly better results than the single-feature models in the whole dataset, but not in the balanced one. This result seems to indicate that simpler models work better in our datasets. A hypothesis that can explain the dominance of the BOW model is the fact that it may capture the orthographic particularities of the learners' writing, but further analyses are needed to prove this. For both datasets and all models, the class with the best F1 score is the basic user level. This fact seems to indicate that this is the proficiency level with the most characteristic traits.

We would like to investigate several aspects in future work. Since variables like task or textual genre have been proved to influence the linguistic complexity and accuracy of L2 texts (Alexopoulou et al., 2017), we would like to test our best models against L2 Portuguese texts from different sources with different topics and tasks, to check the influence of these variables. We also would like to run a cross-lingual experiment with a close language, like Spanish, following the approach of (Vajjala and Rama, 2018). Concerning the machine learning techniques we employed, we are curious about changing the approach and conceive the task as regression, as in (Yannakoudakis et al., 2018). Finally, we would like to test word embeddings and a neural network model, as in (Vajjala and Rama, 2018).

Aknowledgements

This work was supported by national funds through FCT - Fundacão para a Ciência e a Tecnologia postdoctoral program.

References

Alexopoulou, D., M. Michel, A. Murakami, and D. Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques: Task effects in a large-scale learner corpus. *Language Learning*, 03.

Boyd, A., J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, and C. Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC'14*, Reykjavik, Iceland, may.

Branco, A., J. A. Rodrigues, F. Costa, J. R. Silva, and R. Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. In *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR 2014*, pages 256–261.

Burstein, J. 2003. The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. Shermis and J. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, chapter 9, pages 113–121.

Burstein, J. and M. Chodorow. 2012. Progress and New Directions in Technology for Automated Essay Evaluation. *The Oxford Handbook of Applied Linguistics*, pages 487–497, 01.

Curto, P., N. Mamede, and J. Baptista. 2015. Assisting European Portuguese Teaching: Linguistic features extraction and automatic readability classifier. In *Computer Supported Education. Selected Papers from CSEDU2015*. Springer-Verlag.

del Río, I., M. Zampieri, and S. Malmasi. 2018. A Portuguese Native Language Identification Dataset. In *Proceedings of the 13th BEA Workshop*, pages 291–296, New Orleans, Louisiana, June. Association for Computational Linguistics.

Europe, C., C. Cultural Co-operation, E. Committee, and M. Languages Division. 2009. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.*

Gyllstad, H., J. Granfeldt, P. Bernardini, and M. Källkvist. 2014. Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity

in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, 14(1):1–30.

Kyle, K. and S. A. Crossley. 2014. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49, 09.

Lu, X. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2):190–208.

Martins, T. B. F., C. M. Ghiraldelo, M. d. G. V. Nunes, and O. N. d. Oliveira Junior. 1996. *Readability formulas applied to textbooks in Brazilian Portuguese*. Icmsc-Usp.

Mendes, A., S. Antunes, M. Janssen, and A. Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings LREC*.

Nicholls, D. 1999. The Cambridge Learner Corpus-Error coding and analysis.

Otero, P. G. and I. González. 2012. DepPattern: a Multilingual Dependency Parser. In *Proceedings of PROPOR*.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings LREC*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pilán, I., S. Vajjala, and E. Volodina. 2016. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *CoRR*, abs/1603.08868.

Silva, J. R., A. Branco, S. Castro, and R. Reis. 2010. Out-of-the-Box Robust Parsing of Portuguese. In *Proceedings of PROPOR*, pages 75–85.

Taghipour, K. and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891. Association for Computational Linguistics.

Tono, Y. 2000. A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora*, Bern, Switzerland. Peter Lang.

Vajjala, S. and K. Loo. 2013. Role of Morpho-Syntactic Features in Estonian Proficiency Classification. In *Proceedings of the Eighth BEA Workshop*, pages 63–72, Atlanta, Georgia, June. ACL.

Vajjala, S. and K. Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden, November. LiU Electronic Press.

Vajjala, S. and T. Rama. 2018. Experiments with universal CEFR classification. *CoRR*, abs/1804.06636.

Vyatkina, N. 2012. The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. *The Modern Language Journal*.

Woloszyn, V., S. Castilhos, D. Barone, and L. Wives. 2016. Pylinguistics: an open source library for readability assessment of texts written in Portuguese. 04.

Yannakoudakis, H. 2013. Automated assessment of English-learner writing. Technical Report UCAM-CL-TR-842, University of Cambridge, Computer Laboratory, October.

Yannakoudakis, H., T. Briscoe, and B. Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *HLT '11*, pages 180–189, Portland, Oregon, USA, June. ACL.

Yannakoudakis, H., Øistein E Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Procesamiento Multilingüe

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 77-84 recibido 22-03-2019 revisado 05-05-2019 aceptado 15-05-2019

Cross-lingual Diachronic Distance: Application to Portuguese and Spanish

Distancia diacrónica interlingüística: aplicación al portugués y el castellano

José Ramom Pichel Campos[1], **Pablo Gamallo Otero**[2], **Iñaki Alegria Loinaz**[3]
[1]imaxin|software
[2]CITIUS-Universidade de Santiago de Compostela
[3]IXA Taldea-UPV/EHU
[1]jramompichel@imaxin.com
[2]pablo.gamallo@usc.es
[3]i.alegria@ehu.eus

Abstract: The aim of this paper is to establish a corpus-based methodology for automatically measuring the cross-lingual distance between historical periods of two languages using *perplexity*. The corpus of both has been constructed adhoc with the closest spelling to the original representing chronologically and in a balanced way fiction and non-fiction. The methodology has been applied to two related languages, Portuguese and Spanish, and measured their diachronic distances both in original orthography and in an automatically transcribed spelling.
Keywords: Corpus linguistics, Historical Linguistics, Language distance, Development of linguistic resources and tools

Resumen: El objetivo de este trabajo es establecer una metodología basada en corpus para medir automáticamente la distancia interlingüística entre períodos históricos de dos lenguas mediante *perplexity*. El corpus de los dos idiomas ha sido construído adhoc con ortografía lo más próxima a la original representando cronológicamente y de forma balanceada ficción y no ficción. Se ha aplicado la metodología a dos lenguas relacionadas, Portugués y Español, y medido sus distancias diacrónicas tanto en ortografía original como en una ortografía transcrita automáticamente.
Palabras clave: Lingüística de Corpus, Lingüística Histórica, Distancia entre Lenguas, Desarrollo de recursos lingüísticos y herramientas

1 *Introduction*

Languages are constantly changing throughout their history (Millar and Trask, 2015) in such a way that it is as challenging to measure the diachronic distance between periods of the same language as it is to measure the cross-lingual distance between related languages. It is also a challenge to reduce this automatic distance to a single metric to validate the hypotheses of language historians.

There have been different approaches to obtain language distance measures, namely in phylogenetic studies within historical linguistics (Petroni and Serva, 2010), in dialectology (Nerbonne and Heeringa, 1997), in language identification (Malmasi et al., 2016), and in the field of second language acquisition (Chiswick and Miller, 2004). However, to the best of our knowledge, there is no work on how to measure cross-lingual diachronic distance of two different languages. This article proposes a corpus-driven methodology for automatically measuring a cross-lingual diachronic distance between two languages from a historical corpus.

For this general purpose, we consider that the concept of language distance is closely related to the process of language identification. In fact, the more difficult the identification of differences between two languages or language varieties is, the shorter the distance between them. The best language identification systems are based on n-gram models

ISSN 1135-5948. DOI 10.26342/2019-63-8 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

of characters extracted from textual corpora (Malmasi et al., 2016). As a result, character n-grams not only encode lexical and morphological information but also phonological features since phonographic written systems are related to the way languages were pronounced in the past.

The specific objective of the present article is to apply this perplexity-based measure to study and compare the cross-lingual diachronic distance among historical periods of two close-related languages: European Portuguese and European Spanish, from 12th to 20th century. To achieve this goal, we have carried out two different experiments: one applying the methodology of cross-lingual diachronic distance calculation based on perplexity to historical corpus whose texts are written with a spelling very close to the original source; and another applying the same method to the same corpus but automatically transcribed to a common orthography that approximates the two compared languages.

The results show that the two languages are not separated from the Middle Ages in a linear way, but that approximations and divergences occur along the time axis.

Finally, an additional objective of the article is to verify whether the proposed cross-lingual diachronic distance fits the opinion and analysis of philological experts.

The article is organized as follows. Some related work is introduced in Section 2. Then, the method and the corpus are described in sections 3 and 4, respectively. Section 5 introduces the experiments along with a discussion on the results. Finally, conclusions and future work are addressed in Section 6.

2 Related work

Language distance has been defined from different perspectives using different methods. We will explore two different approaches: phylogenetics and corpus based strategies.

2.1 Linguistic Phylogenetics

The objective of linguistic phylogenetics, a sub-field of historical and comparative linguistics, is to classify the languages by building a rooted tree that describes the evolutionary history of a set of related languages or varieties. In order to automatically build phylogenetic trees, many researchers made use of a specific technique called *lexicostatistics*, which is an approach of comparative linguistics that involves quantitative comparison of lexical cognates, which are words with a common historical origin (Nakhleh, Ringe, and Warnow, 2005; Holman et al., 2008; Bakker et al., 2009; Petroni and Serva, 2010; Barbançon et al., 2013). More precisely, lexicostatistics is based on cross-lingual word lists, e.g. Swadesh list (Swadesh, 1952) or ASJP database (Brown et al., 2008), in order to automatically measure distances using the percentage of shared cognates.

Levenshtein distance among words (Yujian and Bo, 2007) in a cross-lingual list is one the most common metrics used in this field (Petroni and Serva, 2010). Ellison et al., (2006), present a method to build language taxonomies comparing lexical forms. The method only compares words language-internally and never cross-linguistically. Finally, Satterthwaite (2011) and Rama and Singh (2009) test four techniques to construct phylogenetic trees from corpora: cross-entropy, cognate coverage distance, phonetic distance of cognates and feature N-grams. They conclude that these measures can be very useful for languages which do not have linguistically hand-crafted lists. Finally, using perplexity-based distance, Gamallo et al., (2017), built a network that represents the current map of similarities and divergences among the main languages of Europe.

2.2 Language distance

To measure language distances, complex language models have been built from large cross-lingual and parallel corpora to obtain metrics to measure language distances. In these works, models are mainly built with distributional information on words, i.e., they are based on co-occurrences of words, and therefore languages are compared by computing cross-lingual similarity on the basis of word co-occurrences (Liu and Cong, 2013; Gao et al., 2014; Asgari and Mofrad, 2016).

Degaetano-Ortlieb et al., (2016) present an information-theoretic approach based on entropy to investigate diachronic change in scientific English. Rama et al., (2015) use cross-entropy to measure distances, while Singh (2007) uses phonetic distances. These studies can be seen as the most related to our work, which is corpus-driven and has been previously applied to the diachronic varieties of the same language (Pichel, Gamallo, and

Alegria, 2018).

3 Methodology

3.1 Perplexity-Based Measure

The distance measure of our method is based on *perplexity*, which is a widely-used evaluation metric for language models. It has been used as a quality measure for language models built with n-grams extracted from text corpora (Chen and Goodman, 1996; Sennrich, 2012). It has also been used in very specific tasks, such as to classify formal and colloquial tweets (González, 2015), and to identify close-related languages (Gamallo et al., 2016). In Gamallo et al., (2017), a specific perplexity-based distance, called PLD, has been defined and applied to compute the distance of different European languages. In a previous work (Pichel, Gamallo, and Alegria, 2018), we applied PLD to measure the diachronic distance between different historical periods of the same language. In the current work, our aim is to apply PLD to measure cross-lingual diachronic distance between two different languages in the same historical periods. In order to be able to compare the perplexity distances we have obtained with those reported in Gamallo et al., (2017), we use the same PLD configuration: namely, 7-gram language models, smoothing technique based on linear interpolation, and train/test corpora with 1,25M/250K words, respectively.

3.2 Task Description

Our methodology requires a representative and balanced historical corpus for each language. The corpus, divided into different historical periods, consists of two versions: texts with original spelling (or as close as possible to the original), and texts automatically transcribed to a common orthography that phonetically approximates the compared languages. In the current work, we will apply this methodology to two close-related languages: Portuguese (Portugal) and Spanish (Spain). Our method is divided into the following specific sub-tasks:

1. First, we search for textual sources to create our diachronic corpus containing texts with a spelling as close as possible to the original for each language. Once the textual sources have been selected, we eliminate noise from the documents, specially excerpts in other languages.

2. Second, we define linguistic and literary equivalent periods for each language. In the definition of periods, we take into account dates of orthographic changes to better observe the possible variations concerning the distance between languages through the time axis. In the current experiments, we have selected six historical periods for the two compared languages.

3. Third, once we have decided on the common historical periods for all languages, we select a representative and balanced historical corpus with an acceptable size for each language. We try to design a corpus that is representative according to Biber's criteria (1993): For this purpose, texts from several genres and topics were retrieved. Both non-fiction and fiction texts for each period have been collected, including fiction subgenres such as narrative, poetry, theater, religious texts for the medieval period, etc. Concerning non-fiction texts, essays were mostly used.

4. Once the textual sources of our corpus have been selected and the periods have been established, two subcorpora are created for each period: train and test. In the train partition, we include for each period texts in original spelling in fiction and non-fiction. In order to facilitate a better representation of the language for each period, the fiction and non-fiction texts in both the train and the test were balanced at approximately 50% (the test and train texts are distinct sets). It is worth mentioning that the train and test partitions are not manually annotated as our method is fully unsupervised.

5. A spelling normalization is applied to all the texts and a transcribed version is obtained for each corpus. The common alphabet consists of 34 symbols, representing 10 vowels (including accents) and 24 consonants, designed to cover most of the commonly occurring sounds, including several consonant palatalizations and a variety of vowel articulation. The encoding is thus close to a phonological one and, then, makes it possible to simplify and homogenize cases in which similar sounds (generally palatalizations) are transcribed differently in different languages. For instance, the palatalized nasal sound is transcribed by our normalizer as "ny", thus unifying the Portuguese spelling "nh" and the Spanish "ñ". Similarly, the palatalized lateral is transcribed as "ly", simplifying the

OS	TS	Edited
Com seu meneio hipócrita, calando. Na alma lodosa da blasfémia o grito. Então exultarão os bons, e o ímpio, (...)	com seu meneio hipocrita calando na alma lodosa da blasfemia o grito então exultarão os bons e o impio (...)	Com seu meneio hipócrita, calando. Na alma lodosa da blasfémia o grito. Então exultarão os bons, e o ímpio, (...)

Table 1: Portuguese excerpt in three versions: original spelling (OS), transcribed (TS), and edited text.

two different spellings "lh" in Portuguese and "ll" in Spanish.

6. Finally, we perform the PLD calculations between pairs of cross-lingual diachronic periods in both original spelling and in automatic transcription, so as to obtain the corresponding distances. The results are evaluated and analyzed later.

In order to allow researches to apply the methodology to any language, we have developed a pipeline architecture in Perl, which is freely available[1]. With this implementation, we have built train partitions giving rise to six different 7-gram diachronic language models per language. Then, we have analyzed all test documents so as to generate six 7-gram files per language.

4 Corpus

The Corpus that we have built and used in our experiments, called *Carvalho*, is freely available and contains the diachronic corpus for the two languages: Carvalho-PT-PT (European Portuguese) and Carvalho-ES-ES (European Castillan, also known as Spanish of Spain).

Our initial aim was to classify the corpus for both languages into historical periods with three fundamental stages: medieval period (XII-XV), modern age (XVI-XVIII), and contemporary age (XIX-XX), following the classification provided by Corpus Helsinki (Rissanen and others, 1993).

However, as Portuguese and Spanish have a large volume of texts and different orthographic standards in the 19th and 20th centuries, we have decided to divide these two centuries into two subperiods (XIX-1, XIX-2, XX-1 and XX-2).

Regarding the different orthographic standards in Portuguese, there was a first orthographic standard in 1779 promoted by the *Academia das Ciências de Lisboa*, which was later reformed in the years: 1885, 1911, 1945, 1973 and 1990. In the case of Spanish, the orthographic standard of 1741 promoted by the *Real Academia Española* was consolidated in the two successive centuries.

We have chosen to use documents with a spelling as close as possible to the original text. This decision makes it possible to compute the cross-lingual diachronic distance between texts in both original and transcribed spelling. Table 1 shows three excerpts of the same text, belonging to the book *A Harpa do crente* by Alexandre Herculano (1810-1877). On the left, we show the original spelling (OS) of the document we have selected to be part of our corpus. In the middle, the same text has been transcribed to a common spelling (TS), including lower-case transformation. On the right, we show an edited version adapted to the current Portuguese. Only OS and TS versions have been selected. No edited version has been introduced in our corpus.

To create the Portuguese Carvalho-PT-PT corpus, we identified and selected documents from the following repositories: Tycho Brahe corpus[2] (Galves and Faria, 2010), Colonia[3] (Zampieri, 2017), *Corpus Informatizado do Português Medieval* (Digited Corpus of Medieval Corpus) (Xavier, Brocardo, and Vincente, 1994), Project Gutenberg, specially for the XIX century[4], Wiki source[5], OpenLibrary[6], Arquivo Pessoa[7],

[1] https://github.com/gamallo/Perplexity

[2] http://www.tycho.iel.unicamp.br/corpus/index.html

[3] http://corporavm.uni-koeln.de/colonia/

[4] https://www.gutenberg.org/browse/languages/pt

[5] https://en.wikisource.org/wiki/Category:Portuguese_authors

[6] https://openlibrary.org/

[7] http://arquivopessoa.net/textos/

Carvalho PT/ES	Train-pt	Test-pt	Train-es	Test-es
XII-XV	1.509M	305K	1.317M	314k
XVI-XVIII	1.449M	289K	1.302M	314K
XIX-1	1.262M	253K	1.368M	311K
XIX-2	1.464M	312K	1.315M	257K
XX-1	1.325M	336K	1.252M	253K
XX-2	1.688M	363K	1.231M	250K

Table 2: Size of Train and Test corpora in six historical periods of Portuguese and Spanish

Linguateca[8], *Corpus de Textos antigos* (Corpus of old texts)[9], *Domínio Público*[10]

Concerning Spanish, Carvalho-ES-ES was built from the following repositories: Project Gutenberg, specially for the XIX century[11], OpenLibrary[12], Wiki source[13].

Finally, the two corpora were partitioned into train and test parts so as to compute the perplexity-based measure (PLD). Table 2 shows the size of both Train and Test corpora across the 6 periods of each language.

5 *Experiments*

The experiments we have carried out consist of measuring the cross-lingual diachronic distance between the different historical periods of Portuguese and Spanish. First, we applied the PLD distance to Carvalho-PT-PT / Carvalho ES-ES in original spelling (OS). Then, PLD was applied to the same corpus but transcribed into a common spelling (TS).

5.1 Results

Table 3 shows the results of applying PLD to OS and TS versions of the Portuguese and Spanish corpora period by period. More precisely, we compared each period cross-lingually: for instance, the PLD distance between the Spanish and Portuguese Medieval periods (XII-XV) in OS is 11,49, but in TS is, as expected, lower: 8,9. And we did the same with the rest of the periods. Figure 1 depicts the same information in a plot so as to better observe how the two languages behave in relation to each other throughout history.

Periods	PLD (OS)	PLD (TS)
XII-XV	11.48	8.9
XVI-XVIII	12.12	8.59
XIX-1	11.54	8.72
XIX-2	9.78	7.49
XX-1	13.20	9.34
XX-2	11.99	9.04

Table 3: Cross-lingual diachronic distance (PLD) between Spanish and Portuguese across six historical periods in original spelling (OS) and transcribed (OS).

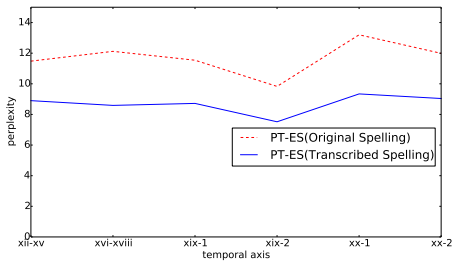


Figure 1: Cross-lingual diachronic distance between Spanish and Portuguese through time axis in OS and TS.

5.2 Discussion

The maximum PLD distance in OS is 13.2, which was reached in the first half of the 20th century (XX-1), while the minimum PLD distance is 9.83, obtained in the second half of the 19th century (XIX-2). In TS, the maximum distance is 9.34 in XX-1, while the smallest one is 7.52 in XIX-2. According to the results reported in Gamallo et al., (2017), the PLD scores of close-related languages of the same family range from 7 (e.g., Croatian

[8] https://www.linguateca.pt/
[9] http://alfclul.clul.ul.pt/teitok/cta/index.php?action=textos
[10] http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=16090
[11] https://www.gutenberg.org/browse/languages/es
[12] https://openlibrary.org/
[13] https://en.wikisource.org/wiki/Category:Spanish_authors

and Bosnian) to 9 (e.g., Czech and Slovak). Those values were obtained from transcribed spelling (TS). Therefore, the distance between all the historical periods of Portuguese and Spanish is always framed in a typical distance of very close languages if they were using a common transcribed spelling.

Another important finding is the following. In all historical periods, the rate of decrease in the distance between the OS and TS varies between 3.86 in XX-1 and 2.31 in XIX-1. This significant drop in PLD seems to suggest that spelling is an important factor in making the difference between the two languages. With a common orthography, Portuguese and Spanish have a very small distance, similar to that of two variants of the same language. By contrast, with two well-differentiated orthographies (as they currently have), the distance widens to more than 13 PLD and resembles that of two clearly different (even if closely related) languages, such as Spanish and Catalan, which have a PLD distance of 14 according to Gamallo et al., (2017).

Yet, The most important observation that can be extracted from the results is the following. The two languages do not separate linearly along the time axis, as might be expected from two languages that start from the same root tongue and standardize independently. On the contrary, their evolution takes place with convergences and divergences not necessarily related to the chronological order. In the first half of the 19th century (XIX-1), both languages diverge with a similar distance to the medieval distance (XII-XV), whereas in the second half of the 19th century (XIX-2) is when their distance converge the most. Later, in the following period (XX-1), their distance increases again reaching the maximum distance but immediately decreases until it reaches values in XX-2 close to those of the Middle Ages.

There may be socio-political motives explaining the consecutive approaches/separations between the two languages. The rapprochement in the second post-Renaissance period (XVI-XVIII) could be explained for the political and cultural hegemony that Castile had in that period that influenced the Portuguese elites, in addition to Portugal's political dependence during the seventeenth century which also influenced cultural and supposedly linguistic issues. Because of this, Spanish words were taken in with ease, as if they were not truly foreign words, but family words (Venâncio, 2014). Also, the promoters of vernacular Portuguese in the Modern Age accentuated and made symbolic use of the difference against the competing language (Spanish). And orthography, above all, served for such a delimiting process (Corredoira, 1998).

The following period of rapprochement between the two languages, in the second half of the 19th century (XX-2), could be due, in part, to the global effects of French and its influence on Roman languages after the Enlightenment period (Curell, 2006). The subsequent distancing between Portuguese and Spanish at the beginning of the 20th century (XX-1) would be partially explained, in addition to the new orthographic rules for Portuguese approved in those years, by the influence of Romanticism, the concept of nation-state and the linguistic *casticism* that derives from this national sentiment.

6 Conclusion and Further work

The present work consists of the automatic calculation of the cross-lingual diachronic distance from two historical corpus of different languages in original orthography. This perplexity-based measure, PLD, was previously used to calculate language distance (Gamallo, Pichel, and Alegria, 2017) and diachronic language distance between different historical periods of the same language (Pichel, Gamallo, and Alegria, 2018).

The experiments we carried out led us to conclude that orthography is an important factor in the distance between Portuguese and Spanish. We also observed that the their distance does not increase chronologically but that historical periods of divergence are followed by periods of convergence and the other way around.

In addition to all these observations, one of the main contributions of this work is the compilation of a freely available diachronic corpus for two languages in closer original spelling: Carvalho-PT-PT and Carvalho-ES-ES[14]. This corpus has been collected from different open historical corpora and texts repositories.

Based on these results, we are planning to use PLD to measure the distance between di-

[14] `https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho`

atopic varieties such as European and Brazilian Portuguese or Latin American Spanish and European Spanish.

Acknowledgments

The authors thanks the referees for thoughtful comments and helpful suggestions. We are very grateful to Fernando Venâncio from the University of Amsterdam, José António Souto Cabo and Carlos Quiroga from the University of Santiago de Compostela for his expertise in Portuguese and Spanish Language history. This work has received financial support from the DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), and the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Asgari, E. and M. R. K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 65–74, San Diego, California.

Bakker, D., A. Muller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, and E. W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181.

Barbançon, F., S. Evans, L. Nakhleh, D. Ringe, and T. Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, 30:143–170.

Biber, D. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Brown, C. H., E. W. Holman, S. Wichmann, and V. Velupilla. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4).

Chen, S. F. and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chiswick, B. and P. Miller. 2004. *Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages*. Discussion papers. IZA.

Corredoira, F. V. 1998. *A construção da língua portuguesa frente ao castelhano: o galego como exemplo a contrario*.

Curell, C. 2006. La influencia del francés en el español contemporáneo. In *La cultura del otro: español en Francia, francés en España*, pages 785–792. Universidad de Sevilla.

Degaetano-Ortlieb, S., H. Kermes, A. Khamis, and E. Teich. 2016. An information-theoretic approach to modeling diachronic change in scientific english. *Selected Papers from Varieng-From Data to Evidence (d2e)*.

Ellison, T. M. and S. Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 273–280.

Galves, C. and P. Faria. 2010. Tycho Brahe parsed corpus of historical Portuguese. *URL: http://www. tycho. iel. unicamp. br/˜ tycho/corpus/en/index. html*.

Gamallo, P., I. Alegria, J. R. Pichel, and M. Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177.

Gamallo, P., J. R. Pichel, and I. Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Gao, Y., W. Liang, Y. Shi, and Q. Huang. 2014. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393(C):579–589.

González, M. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7.

Holman, E., S. Wichmann, C. Brown, V. Velupillai, A. Muller, and D. Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica*, 42(2):331–354.

Liu, H. and J. Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Malmasi, S., M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, and J. Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL Shared Task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, pages 1–14, Osaka, Japan.

Millar, R. M. and L. Trask. 2015. *Trask's historical linguistics*. Routledge.

Nakhleh, L., D. A. Ringe, and T. Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420.

Nerbonne, J. and W. Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pages 11–18.

Petroni, F. and M. Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, 389(11):2280–2283.

Pichel, J. R., P. Gamallo, and I. Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to european portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155.

Rama, T., L. Borin, G. Mikros, and J. Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification.

Rama, T. and A. K. Singh. 2009. From bag of languages to family trees from noisy corpus. In *Proceedings of the International Conference RANLP-2009*, pages 355–359.

Rissanen, M. et al. 1993. The helsinki corpus of english texts. *Kyttö et. al*, pages 73–81.

Satterthwaite-Phillips, D. 2011. *Phylogenetic Inference of the Tibeto-Burman Languages Or on the Usefulness of Lexicostatistics (and" megalo"-comparison) for the Subgrouping of Tibeto-Burman*. Stanford University.

Sennrich, R. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Singh, A. K. and H. Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, pages 40–47. Association for Computational Linguistics.

Swadesh, M. 1952. Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society 96*, pages 452–463.

Venâncio, F. 2014. O castelhano como vernáculo do português.

Xavier, M. F., M. T. Brocardo, and M. Vincente. 1994. Cipm–um corpus informatizado do português medieval. *Actas do X Encontro da Associação Portuguesa de Linguística*, 2:599–612.

Yujian, L. and L. Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Zampieri, M. 2017. Compiling and processing historical and contemporary portuguese corpora. *arXiv preprint arXiv:1710.00803*.

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 85-92 recibido 03-04-2019 revisado 14-05-2019 aceptado 17-05-2019

Action Type induction from multilingual lexical features

Tipos de acciones inducidas a partir de de características léxicas multilingües

Lorenzo Gregori, Rossella Varvara, Andrea Amelio Ravelli
University of Florence
{lorenzo.gregori, rossella.varvara, andreaamelio.ravelli}@unifi.it

Abstract: This paper presents a vector representation and a clustering of action concepts based on lexical features extracted from IMAGACT, a multilingual and multimodal ontology of actions in which concepts are represented through video prototypes. We computed vectors for 1,010 action concepts, where the dimensions correspond to verbs in 10 languages. Finally, an unsupervised clustering method has been applied on these data in order to discover action classes based on typological closeness. Those clusters are not language-specific or language-biased, and thus constitute an inter-linguistic classification of action domain.
Keywords: Action verbs, Vector Space Model, Multilingual Semantics, Action clustering

Resumen: Este artículo presenta una representación vectorial y un clúster de conceptos de acción basados en características léxicas extraídas de IMAGACT, una ontología de acciones multilingüe y multimodal en la que los conceptos se representan a través de prototipos de video. Calculamos vectores para 1.010 conceptos de acción, donde las dimensiones corresponden a verbos en 10 idiomas. Finalmente, se ha aplicado un método de agrupación no supervisada en estos datos para descubrir clases de acción basadas en la proximidad tipológica. Esos clústers no son específicos del idioma ni están sesgados por él, y por lo tanto constituyen una clasificación interlinguística del dominio de acción.
Palabras clave: Verbos de acción, clasificación interlinguística, agrupación de acción

1 Introduction

In the mind of a speaker, the linking between real world entities (i.e. objects and events) and their mental representation is expressed through symbols (i.e. lexical items), which are part of his own mother-tongue language[1]. This picture becomes complicated due to cognitive economy constraints (Rosch, 1978), because one lexical item productively applies to a set of objects or events, resulting in a one-to-many relations (Moneglia, 1996). Consequently, the mental representation of the world consists in a complex network of connections between entities, thoughts and lexicon. To give an idea of this complexity, we can take into account all the possible pragmatic actions that an English speaker can correctly refer to with the verb *put*. In fact, this verb activates in the mind of the speaker a series of possible events, often quite distant in terms of pragmatic execution. For example, *putting a book on the table* and *putting some jam on the bread*, from the lexical point of view of the verb, are similar events, despite their differences: in the first one, no tools are required[2], and few motoric activations are needed to perform the action; on the contrary, in the second one, a tool is needed (e.g. a knife) even if not explicitly lexicalized, and a sequence of various short actions are performed to complete the task. Moreover, only the

[1]The correlation between entities, thoughts and lexicon is often referred to as the Triangle of Reference (Ogden and Richards, 1923).

[2]According to the generativist point of view, the arm and the hand could be considered as tools (Pastra and Aloimonos, 2012).

ISSN 1135-5948. DOI 10.26342/2019-63-9
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

second event can be predicated correctly also by the verb *spread* which, in its turn, extends its application to a series of events, some of which are other than those activated by *put*, e.g. *people spreading around the room.*

If we extend the focus from one single speaker to a community of speakers, we can see that the mental representation is roughly shared and still holds. But what happens if we try to compare the mental representation of two or more speakers of different languages? We immediately observe a variation in the linking between items of the conceptual space and words of the lexical dimension. In fact, if we ask a Japanese speaker to lexicalize the three aforementioned actions, he would use *oku* (common translation of *put*) to predicate the event of *putting a book on the table*, *tsukeru* for *putting/spreading some jam on the bread*, and *chirabaru* for *people spreading around*, and none of these verbs are overlapping or interchangeable in their primary pragmatic predication.

These examples suggest that even a pure lexical discrimination is able to finely segment the conceptual space, and highlight closeness or distance between concepts, both in a monolingual and a multilingual setting.

Starting from these observations, the work presented in this paper is a first attempt in modelling a language-independent conceptual space representation of actions, using multilingual lexical items as a piece of evidence for action concept discrimination.

2 Connection with other works

The present work is a first attempt in solving a novel NLP task, that we will call Action Type Induction (ATI): it faces the problem of automatically identify similar action instances and group them in types. Similarly to Word Sense Induction (WSI) (Pantel and Lin, 2002), ATI can be addressed as an unsupervised clustering problem, where vectors represent actions, instead of words. This difference is not trivial, and makes ATI an intrinsically multimodal task, in which instances are video elements of performed actions, and the encoded features can be both linguistic and visual. For example, informations about verbs or textual description that refer to actions, as well as features regarding motion and trajectories, become relevant data for this task.

The combination of linguistic and visual features to perform a more accurate classification of actions has been widely used in recent years (Silberer, Ferrari, and Lapata, 2013; Hahn, Silva, and Rehg, 2019; Naha and Wang, 2016), with the development of techniques based on the integration of NLP and computer vision. Within this perspective, our work could be fruitfully exploited to build complex models for action understanding grounded on human knowledge.

Since our instances are actions, and not words, we couldn't use a co-occurrence matrix. Instead our dataset is a co-referentiality matrix, that encodes local equivalence, i.e. the ability of two verbs to refer to the same action concepts (Panunzi, Moneglia, and Gregori, 2018).

A similar approach has been used to represent typological data (Ryzhova, Kyuseva, and Paperno, 2016): a matrix of word references, in which each row corresponds to nouns from a specific semantic field and the dimensions are adjectives from different languages. An intersection of a row and a column is filled with 1 if the adjective can occur in the context and with 0 if it cannot. Given these data, they compute *typological closeness* between nouns. From their work we inherit the notion of typological closeness, that is semantic similarity based on comparison of multilingual data.

3 Action vector representation

3.1 IMAGACT

IMAGACT[3] (Moneglia et al., 2014a) is a multimodal and multilingual ontology of action that provides a video-based translation and disambiguation framework for action verbs. The resource is built on an ontology consisting in a fine-grained categorization of action concepts, each represented by one or more visual prototypes in the form of recorded videos and 3D animations. IMAGACT currently contains 1,010 scenes which encompass the action concepts most commonly referred to in everyday language usage. Action concepts have been gathered through an information bootstrapping from Italian and English spontaneous spoken corpora, and the occurrences of verbs referring to physical actions have been manually annotated (Moneglia et al., 2012). Metaphorical and phraseological usages have been ex-

[3] http://www.imagact.it

cluded from the annotation process, in order to collect exclusively occurrences of physical actions.

The database evolves continuously and at present contains 10 fully-mapped languages and 17 which are underway. The insertion of new languages is obtained through competence-based extension (CBE) (Moneglia et al., 2014b) by mother-tongue speakers, using a method of ostensive definitions inspired by Wittgenstein (Wittgenstein, 1953). The informants are asked to watch each video, to list all the verbs in their language that correctly apply to the depicted action, and to provide a caption describing the event for every listed verb as an example of use.

The visual representations convey action information in a cross-linguistic environment, offering the possibility to model a conceptualization avoiding bias from monolingual-centric approaches.

3.2 Dataset

From the IMAGACT database, we derived our dataset as a binary matrix $C_{1010\times7881}$ with one row per video prototype and one column per verb for the languages considered. Matrix values are the assignments of verbs to videos made by native speakers within the CBE annotation task:

$$C_{i,j} = \begin{cases} 1 & \text{if verb } j \text{ refers to action } i \\ 0 & \text{else} \end{cases}$$

In this way, the matrix C encodes the interlinguistic lexical representation of each video prototype.

Table 1 shows the number of verbs assigned by the CBE annotators for each language. It is important to notice that the task has been performed on the whole set of 1,010 scenes for each language and the differences between the number of verbs depend on linguistic factors: some examples of verb-rich languages are (a) Polish and Serbian, in which perfective and imperfective forms are lemmatized as different dictionary entries, (b) German, that have particle verb compositionality, (c) Spanish and Portuguese, for which verbs belong to both American and European varieties.

Judgments of applicability of a verb to a video scene rely on the semantic competence of annotators. An evaluation of CBE assignments has been made for Arabic and Greek in two thesis (Mutlak, forthcoming; Mouyiaris, forthcoming); results are summarized in Table 2.

Language	**Verbs**
Arabic (Syria)	571
Danish	646
German	990
Greek	638
Hindi	470
Japanese	736
Polish	1,193
Portuguese	805
Serbian	1,096
Spanish	736
TOTAL	**7881**

Tabla 1: Number of verbs per language

Language	**Precision**	**Recall**
Arabic (Syria)	0.933	0.927
Greek	0.990	0.927

Tabla 2: Precision and Recall for CBE annotation task measured on 2 languages

3.3 Creating action vectors

In order to provide an exploitable vector representation of action prototypes, an approximated matrix C' has been created from C, by using *Singular Value Decomposition* (SVD) for dimensionality reduction.

SVD is a widely used technique in distributional semantics to reduce the feature space. The application of SVD to our dataset allowed us to obtain a fixed-size feature space, that is independent of the number of languages, and an approximation matrix, that smooths language-specific semantic differences. These results are highly desirable, considering that the number of languages in IMAGACT is growing continuously, and that the provided representation should be shared as far as possible, abstracting from lexico-semantic properties of single languages. Moreover, SVD approximation leads to some advantages in terms of computational processing, by removing the matrix sparsity and creating a relatively low dimensional space.

Finally, the output C' is a dense matrix 1010×300 that encodes lexical features of action prototypes.

4 *Actions clustering*

In order to obtain a language-independent classification of action concepts, we computed similarity between action prototype vectors, and then we applied an unsupervised clustering algorithm to this data. The resulting classification bypasses differences in lexicalization among languages, in favour of an *average* conceptual representation of actions. In fact, a classification based on data from only one language leads to a representation that is consistent with the semantic space segmentation operated by that language, but it may not be cross-linguistically generalized. Considering more languages together, instead, language-specific differences can be leveraged, highlighting similarities that may remain in shade if comparing monolingual classifications of actions.

Data clustering in this scenario is a complex task because we do not have any information on the number of clusters, that must be found by the clustering algorithm. Moreover, a proper evaluation of the resulting clusters is not trivial, since we need to compare one speaker's conceptual representation with the average representation resulting from summing lexical information from multiple languages.

4.1 Clustering algorithm

An ATI task can be properly considered as a variation of a Word Sense Induction (WSI) task: instead of grouping word occurrences with similar meaning based on word contexts, here we are grouping similar action occurrences based on lexical features. In this experiment we used *Affinity Propagation* (AP) (Frey and Dueck, 2007), a state-of-the-art unsupervised clustering algorithm, that has been successfully applied to accomplish WSI tasks in recent works (Alagić, Šnajder, and Padó, 2018; Arefyev, Ermolaev, and Panchenko, 2018). AP automatically identifies the optimal number of clusters for a given dataset; each cluster consists of one *exemplar* (i.e. one element of the dataset that is representative of the cluster) plus its neighbouring elements.

Results of clustering on C' matrix are summarized in Table 3.

4.2 The map of action concepts

A visual map has been created for data exploration purposes, available at http://lablita.it/app/imclust/map.php. In this map each point is a cluster (i.e. a set of action videos); the spatial position of points is derived as follows:

Number of clusters	178
Min # of scenes per cluster	2
Max # of scenes per cluster	24
Average # of scenes per cluster	5.67

Tabla 3: Results of clustering algorithm

- exemplar vectors are chosen as cluster representatives;
- the feature space has been reduced to 2 dimensions with t-SNE[4];
- this 2D representation of exemplars has been projected on x and y axes.

The action map is interactive: by clicking on each point it is possible to see the set of videos belonging to the cluster. In order to ease data exploration, cluster regions have been manually drawn and named with an English verb that roughly describes the related semantic area.

5 *Evaluation*

5.1 Evaluation task

To our knowledge, no similar previous evaluation work are available and, due to the peculiarity of this kind of work (i.e. language independent clustering obtained by summing of multilingual data), the possibility that a speaker of a language L may find reasonable all the clusters could be far from being positive. Nevertheless, in order to evaluate the obtained Action Type clusters and the applied methodology, we designed a task to perform an evaluation based on human percepts on scenes similarity.

The evaluation consists in a two-alternative forced-choice similarity task in which, given as target a scene s from a cluster c, participants were asked to chose the most similar scene to the target among two other scenes. One of the scene used for comparison belongs to the same cluster c of the target, whereas the other was selected among scenes not belonging to it. We expected annotators to choose the scene belonging

[4] t-SNE is a dimensionality reduction algorithm, specifically designed for visual representation of high dimensional data (Maaten and Hinton, 2008).

to the same cluster as more similar to the target one. If human judgments mirror our unsupervised clusters, we interpret it as a reliability result. Figure 1 shows one item of the evaluation test: annotators judge which video, between 1 and 2, is more similar to the target.

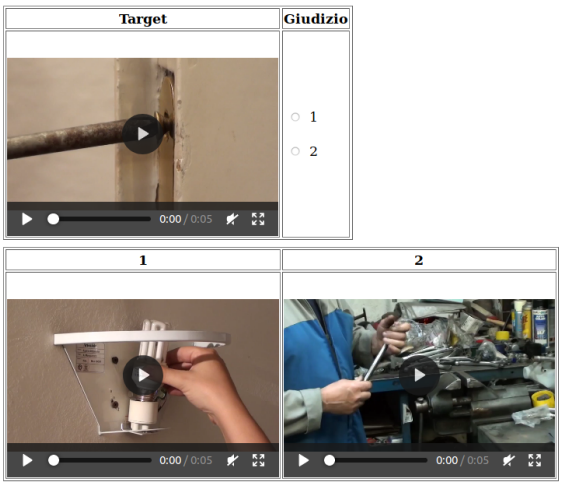


Figura 1: One item of the evaluation test

We conducted a preliminary evaluation on 11 randomly selected clusters. These were selected considering the number of scenes they contain: the clusters considered were in the middle range for number of scenes (one standard deviation around the mean[5]).

We presented to participants every possible couple of scenes belonging to the same cluster, thus resulting in a dataset of 125 datapoints. The third scene of each evaluation item was randomly selected among scenes not belonging to the considered cluster. Moreover, we checked for the similarity distance of this scene from the cluster, and prepared two version of the test: the third scene was alternatively selected among the 15th most similar scenes not belonging to the target cluster or among other farther scenes, and this selection was inverted in the second version of the test. In this way, we can evaluate finely the precision in the categorization of closer concepts. Furthermore, we divided each test in two, in order to avoid a very time consuming test. As a result we obtained 2 different tests with 2 alternative version each (4 tests in total).

[5]We preferred clusters with a number of scenes in the middle range for two reasons: 1- to avoid outlier clusters; 2- to keep the number of items per test small enough, feasible in less then 30 minutes

The first test was performed by 10 annotators, 5 for each version, while the second test was performed by 6 annotators, 3 for each version. All the annotators are Italian native speakers. Note that Italian is not among the languages included in our multilingual matrix. We believe this is an additional strength of our evaluation, because results are not biased from the presence of annotators' mother language into the multilingual matrix.

This preliminary evaluation is meant as a first check on the task suitability. An extensive evaluation with a crowd sourcing platform has been planned, which will consider a larger number of clusters and will collect judgments from speakers of different languages. It will be interesting to analyze the influence of participants mother language in the results, observing language-specific differences.

5.2 Evaluation results

Tables 4 and 5 show results of Test 1 for each annotator; tables 6 and 7 show results of Test 2 for each annotator. Values range from 0 to 1, where 1 indicates that the scenes judged as more similar to the target correspond to the cluster's internal ones. The two versions of the test had a small difference in the percentage of correct pairs (0.88 for Test 1.1, 0.91 for test 1.2 and 0.97 for Test 2.1, 0.92 for Test 2.2), suggesting us that the choice of the third (not belonging to the cluster) scene is relevant. Moreover, results show a small difference between near and far scenes: if the third scene was farther to the cluster, annotators judged the intra-cluster scene as similar to the target scene more easily. Finally Cohen's k has been measured between each annotator and the automatic clustering assignments; the resulting agreement is high: 0.75 for Test 1.1; 0.81 for Test 1.2; 0.83 for Test 2.1; 0.80 for Test 2.2.

	near	**far**	**total**	***k***
annot 1	0.80	0.92	0.86	0.84
annot 2	0.80	0.88	0.84	0.71
annot 3	0.80	0.92	0.86	0.80
annot 4	0.96	0.88	0.92	0.67
annot 5	0.88	0.92	0.90	0.72
Average	**0.85**	**0.90**	**0.88**	**0.75**

Tabla 4: Evaluation results - Test 1.1

Table 8 and 9 report the distribution of shared judgments among participants, i.e. how many items are evaluated according to

	near	far	total	*k*
annot 6	0.88	0.92	0.90	0.84
annot 7	0.92	0.92	0.92	0.84
annot 8	0.88	0.92	0.90	0.80
annot 9	0.88	0.92	0.90	0.79
annot 10	0.92	0.92	0.92	0.79
Average	**0.89**	**0.92**	**0.91**	**0.81**

Tabla 5: Evaluation results - Test 1.2

	near	far	total	*k*
annot 11	0.97	0.97	0.97	0.88
annot 12	0.97	1.00	0.98	0.74
annot 13	0.97	0.97	0.97	0.86
Average	**0.97**	**0.98**	**0.97**	**0.83**

Tabla 6: Evaluation results - Test 2.1

	near	far	total	*k*
annot 14	0.88	1.00	0.93	0.76
annot 15	0.91	0.95	0.93	0.87
annot 16	0.88	0.97	0.92	0.78
Average	**0.89**	**0.97**	**0.92**	**0.80**

Tabla 7: Evaluation results - Test 2.2

the clustering and by how many annotators. For example, the first row of Table 8 reports the number of evaluated items where all of the 5 annotators (5/5) identified the scene belonging to the cluster as more similar to the target.

In general, the evaluation confirmed the validity of the cluster obtained: clusters seem to correspond to what we call Action Type, since they mirror with a high percentage to human judgments. The matrix and the clusters obtained can be, thus, interpreted as interlinguistic and cognitively valid representation of Action Types. With further evaluations we want to investigate to which extent linguistic (lexical) representation mirrors cognitive categorization, thus providing new insights on how we organize concepts.

6 Conclusion and future works

This work describes an experiment of Action Type Induction, that has been addressed through an unsupervised clustering on a matrix of multilingual lexical features. Vectors encode the information about the applicability of verbs (in 10 languages) to action concepts and are extracted from the IMAGACT Ontology of Action. Affinity Propagation clustering algorithm has been applied to these data and the 1,010 video scenes have been grouped in 178 clusters, that represent a conceptualization of action domain. A preliminary evaluation has been made on a set of 11 clusters by 16 annotators, obtaining encouraging results. The very next step will be the implementation of an evaluation campaign to obtain a reliable measure of clustering accuracy without a gold standard.

Pair agr.	Test 1.1	Test 1.2	Total
5/5	39	40	79
4/5	2	4	6
3/5	3	2	5
2/5	2	1	3
1/5	3	3	6
0/5	1	0	1

Tabla 8: Agreement summary on scenes internal to the target cluster in Test 1

Pair agr.	Test 2.1	Test 2.2	Total
3/3	71	67	138
2/3	1	3	4
1/3	2	1	3
0/3	0	3	3

Tabla 9: Agreement summary on scenes internal to the target cluster in Test 2

The action representation provided in this experiment is based on lexical features only, but ATI tasks can exploit several kinds of action features. For this reason, this work can be considered as a first step in the creation of a more complete vector representation of actions that encodes other aspects of action domain. Some examples of features that, in principle, could enrich these vectors regard event structures (e.g. goal, semantic frames), linguistic properties of linked verbs (e.g. thematic roles, aktionsart), visual properties of the scenes (e.g. action trajectory, spatial relation among objects), motor description of performed movements.

References

Alagić, D., J. Šnajder, and S. Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Arefyev, N., P. Ermolaev, and A. Panchenko. 2018. How much does a word weigh? weighting word embeddings for word sense induction. *arXiv preprint arXiv:1805.09209.*

Frey, B. J. and D. Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Hahn, M., A. Silva, and J. M. Rehg. 2019. Action2Vec: A Crossmodal Embedding Approach to Action Learning. *arXiv.org*, page arXiv:1901.00484, January.

Maaten, L. v. d. and G. Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Moneglia, M. 1996. Prototypical vs. non-prototypical predicates: ways of understanding and the semantic partition of lexical meaning. *QDLF - Quaderni del Dipartimento di Linguistica*, 7:163–181.

Moneglia, M., S. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, and A. Panunzi. 2014a. The imagact visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Moneglia, M., S. Brown, A. Kar, A. Kumar, A. K. Ojha, H. Mello, Niharika, G. N. Jha, B. Ray, and A. Sharma. 2014b. Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In G. N. Jha, K. Bali, S. L, and E. Banerjee, editors, *Proceedings of WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation at LREC'14*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Moneglia, M., F. Frontini, G. Gagliardi, I. Russo, A. Panunzi, and M. Monachini. 2012. Imagact: deriving an action ontology from spoken corpora. *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pages 42–47.

Mouyiaris, A. forthcoming. I verbi d'azione del greco nell'ontologia IMAGACT. Master's thesis.

Mutlak, M. forthcoming. *I verbi di azione dell'arabo standard nell'ontologia dell'azione IMAGACT*. Ph.D. thesis.

Naha, S. and Y. Wang. 2016. Beyond verbs: Understanding actions in videos with text. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1833–1838. IEEE.

Ogden, C. and I. A. Richards. 1923. The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism. *8th ed. 1923. Reprint New York: Harcourt Brace Jovanovich.*

Pantel, P. and D. Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.

Panunzi, A., M. Moneglia, and L. Gregori. 2018. Action identification and local equivalence of action verbs: the annotation framework of the imagact ontology. In J. Pustejovsky and I. van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Pastra, K. and Y. Aloimonos. 2012. The minimalist grammar of action. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1585):103–117.

Rosch, E. 1978. Principles of Categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Lawrence Elbaum Associates.

Ryzhova, D., M. Kyuseva, and D. Paperno. 2016. Typology of adjectives benchmark for compositional distributional models. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 1253–1257.

Silberer, C., V. Ferrari, and M. Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582. Association for Computational Linguistics.

Wittgenstein, L. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag.

Técnicas de Generación de Resúmenes y Simplificación Textual

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 95-102 recibido 28-03-2019 revisado 08-05-2019 aceptado 15-05-2019

Lexical simplification approach using easy-to-read resources

Enfoque de simplificación léxica utilizando recursos de lectura fácil

Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, Paloma Martínez
Human Language and Accessibility Technologies group (HULAT)
Universidad Carlos III de Madrid
Leganés,Madrid,Spain
{ralarcon, lmoreno, isegura, pmf}@inf.uc3m.es

Abstract: This work aims to facilitate the understanding and readability of Spanish texts in a generic domain through the design of a lexical simplification system that provides support to the task of Complex Word Identification (CWI) and selection of a simpler substitute. Considering the limited resources available in Spanish, we explore different features that allow us to discern between a complex word and a simpler one. Some of these features are obtained from easy-to-read resources. The evaluation shows good results by obtaining an F1-Score of 0.7497 on the CWI Task with the BEA Workshop 2018 competition's dataset.
Keywords: Lexical Simplification, Accessibility, easy-to-read

Resumen: Este trabajo tiene como objetivo facilitar la comprensión y legibilidad de textos en español en un dominio genérico a través del diseño de un sistema de simplificación léxica que da soporte a la tarea de Complex Word Identification (CWI) y elección de sustituto más sencillo. Considerando la limitación de recursos disponibles en español, exploramos diferentes características que nos permitan discernir entre una palabra compleja y una simple. Algunas de estas características son obtenidas de lectura fácil. La evaluación muestra buenos resultados al obtener 0.7497 en F1-score en la tarea de CWI con el dataset de la competición de BEA Workshop 2018.
Palabras clave: Simplificación Léxica, Accessibilidad, lectura fácil

1 Introduction

Information and communication technologies (ICT), especially the Internet, have transformed the way we live and communicate. People access ICT and all the services offered, however, many of these services are not accessible to all people (Chayle et al., 2017). Although people with disabilities are the user group directly affected, accessibility barriers affect all citizens. There are initiatives, legislation and normative that promote and regulate accessibility in ICT, however, accessibility remains a challenge (Moreno et al., 2018).

The needs of people with sensory and physical disabilities are better known, ignoring the cognitive barriers caused by the difficult understandability to the textual content that mainly affects people with intellectual and learning disabilities. Texts that contain unusual words can cause barriers of cognitive accessibility for people with intellectual disabilities. The need for simplified texts becomes increasingly important as the incidence of disability increases as the population ages. The benefits of providing accessible interfaces and simplified content not only benefit people with intellectual and learning disabilities, but also the deaf, deaf-blind, elderly, illiterate and immigrants with a different native language (Saggion, 2017).

In order to provide universal information and make texts more accessible, there are cognitive accessibility guidelines such as Web Content Accessibility Guidelines (WCAG) (W3C, 2019) and easy-to-read guidelines (Smith, Hallam, and Ghosh, 2012) (Freyhoff et al., 1998) and plain language guidelines addressed to all citizens (www.plainenglish.co.uk/free-guides.html) (www.plainlanguage.gov). These resources provide helpful documentation; however, these guidelines are complicated to comply

ISSN 1135-5948. DOI 10.26342/2019-63-10 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

within a systematic way.

As a solution space, there are methods that support this systematic compliance with these guidelines of cognitive accessibility as Natural Language Processing (NLP), which provides methods to simplify texts and thus promote readability and understandability to people with intellectual disabilities. With this motivation, this work arises, which proposes a system that supports the process of lexical simplification to improve cognitive accessibility.

This article is structured as follows: Section 2 presents related work, Section 3 presents the methodology and resources taken in order to build the lexical simplifier. In Section 4, we discuss the results obtained at the CWI task. Finally, Section 5 offers conclusions and future works.

2 Related Work

Since 1996 the automatic simplification of texts began (Shardlow, 2014) doing a superficial analysis of the text to identify verbs and nouns of complex phrases. Between many ways of approaching this task, there is the syntactic simplification which consists in identifying grammatical complexities and turn it in a much simpler version (Gonzalez-Dios, 2017) and lexical simplification, which can be described as the task of substituting words in a given phrase to make it simple, without applying any modification to its syntactic structure. For the Spanish language, there are different ways to accomplish this task, from supervised, unsupervised or recently proposed hybrid approaches (Štajner, Saggion, and Ponzetto, 2019). Supervised approaches need annotated datasets to achieve their objective (Štajner, Calixto, and Saggion, 2015), this leads to a great disadvantage when dealing with languages with few annotated corpora for text simplification (Saggion et al., 2011) (Mitkov and Štajner, 2014). Unsupervised approaches, while outperforming supervised approaches in coverage, have the disadvantage of only making one-to-one substitutions, not being able to deal with phrases, they also tend to change the meaning of the sentence and have problems dealing with ambiguous words (Glavaš and Štajner, 2015) (Paetzold and Specia, 2016b). While hybrid approaches use methods from the previous two, such as (Ferrés, Saggion, and Guinovart, 2017), which uses a corpus-based approach and a combination of a free lexicon, decision trees and context-based rules. Concerning methodological approaches (Paetzold and Specia, 2017) proposes that the lexical simplification should be conducted in the following four steps: Complex Word Identification (CWI), Generation of Substitutes, Selection of Substitutes and Substitutes Ranking. This paper follows this approach.

CWI aims to select the words in a sentence that must be simplified, that is, to detect which words are complex in a given text. Machine learning approaches have demonstrated to overcome other strategies. Shardlow (Shardlow, 2013) conducted a research in which he compares a binary Support Vector Machine (SVM), a Threshold-based and a "Simplify Everything" approach, where in the latter, it is assumed that all the words in a sentence can be simplified. Results show that the SVM classifier exceeds the other approaches in precision. Machine Learning approaches are widely applied to CWI task, as the reader can see in BEA Workshop (Yimam et al., 2017) on the task of CWI. Most of the participating systems presented Machine Learning approaches. This is the case of the work (Hartmann and dos Santos, 2018) that presented three approaches for CWI, one using traditional classification algorithms of Machine Learning, such as Linear Regression, Logistic Regresion, Decision Trees, Gradient Boosting, Extra Trees, AdaBoost and XGBoost based on lexical features(word length, number of syllables and others), N-gram features(probabilities of n-gram). The second one using Word Embedding to get the vector representations of target words and in the last, the context of the target words is modeled using a deep learning model, Long Short Term Memory (LSTM).

In relation to the second step, substitute generation, refers to the process of producing candidate synonyms for complex words detected. Most of the works use existing dictionaries, among them the most used is Wordnet (Lal and Ruger, 2002) (Burstein et al., 2007). Another work that uses this approach is that of (Bott et al., 2012), where lexical simplification in Spanish uses the OpenThesaurus database, which has 21,381 target words and provides a list of synonyms for each word. On the other hand, CASSA (Baeza-Yates, Rello, and Dembowski, 2015) presents an improve-

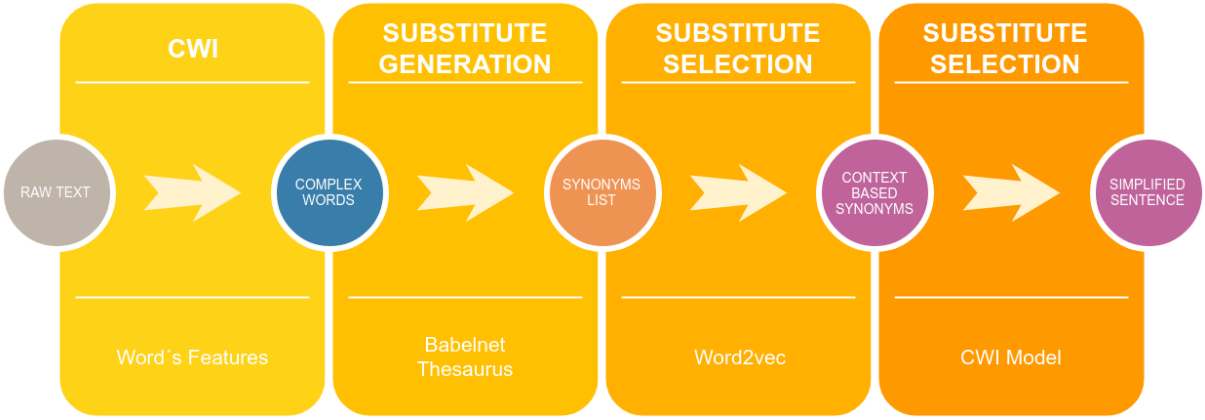


Figure 1: Lexical simplification system pipeline as shown by (Shardlow, 2014) and specified by (Paetzold and Specia, 2017)

ment to the previous approach evaluating the context, the frequency of each substitute and its context. These features are obtained using a Google corpus. The experiments show that the use of these features is a better substitution approach that only using the OpenThesaurus dictionary.

In the third step, the selection of a substitute from the set of the synonyms extracted from in the previous step, the most suitable synonym is selected according to factors such as simplicity and its context. In this stage, the selected synonym should preserve the original meaning of the sentence as well as a correct syntactic structure. In (Paetzold and Specia, 2015), the authors select the final synonym using the cosine distance in a word embedding model. In particular, given a word to be simplified, the word with the closest vector based on cosine similarity is chosen.

3 Lexical Simplifier

As has already been mentioned, in this work the approach of (Paetzold and Specia, 2017) is followed. Figure 1 shows our system pipeline, highlighting the different processes and the resources used in each step. In the following sections, these steps are detailed.

3.1 Complex Word Identification

In order to distinguish which words are complex and which are not, a machine learning approach is followed. In particular, we apply a Support Vector Machine (SVM) because its successful performance for text classification tasks. Moreover, SVM was also one of the most used algorithms for this task on SemEval2016 (Paetzold and Specia, 2016a).

3.1.1 Datasets

We use the datasets provided in shared task of the multilingual Complex Word Identification (CWI) in BEA Workshop 2018 (sites.google.com/view/cwisharedtask2018). The datasets for the Spanish language contains texts from Wikipedia pages in Spanish, which were annotated by 48 native speakers and 6 non-native speakers. The dataset provides a list of words and their corresponding classification (1 for complex words or 0 for simple ones), along with more useful information that can help in the classification. For example, the sentence where the word occurs, the position of the word in the sentence, among others, the number of annotators who classified it as a complex word, among others. The training dataset contains a total of 13,747 instances, 11962 monotoken and 1785 multitoken from which 40% represent complex words and the test dataset contains a total of 2233 instances, 1955 monotoken and 278 multitoken from which 41% represent complex words. Each word is represented by a set of features, which are described in the following section. More available description of this task and dataset can be found in the competition report (Yimam et al., 2018).

3.1.2 Proposed Features

In order to train the algorithm, we need to represent each word (instance) as a set of features that should help to distinguish between a complex word and a simple word. Some of the most discriminative features for the CWI task are (Paetzold and Specia, 2016a): the

length of a word, sentence and the frequency of a word in a large corpus. The chosen features are described as follows:

- **Length Features:** Word Length, Sentence Length, Number of syllables of the word.
- **Frequency Features** using a Ngram Corpus: the frequency of the word(unigram), the frequency of bigram (word and left/right word), the frequency of trigram (word and left/right two next words).
- **Boolean Features:** if the word is lowercase, if the word is Uppercase, if the word is a digit, if the word has uppercase characters, if the word is composed of punctuation symbols, if the word contains punctuation symbols.
- **E2R Feature:** One of our main objectives is that our lexical simplification system complies with accessibility guidelines such as easy-to-read (E2R) guidelines. In this sense, we propose a indicating if an input word is found or not in an E2R dictionary, which has been built for this research work. If a word exists in the E2R dictionary, the word is automatically qualified as simple. The dictionary feeds from different sources that provide texts in E2R elaborated by experts, some of these sources are: The Noticias facil news page (www.noticiasfacil.es/) and the Easy Reading Association (www.lecturafacil.net/es/). Using a crawler, we obtained a large collection of texts from these sources. Then, texts were tokenized, lematized and part-of-speech (PoS) tagged by using the NLP tool FreeLing (http://nlp.lsi.upc.edu/freeling/). The PoS tags allow us to remove the non-lexical words (such as determiners, pronouns, conjunctions, modal and auxiliary verbs, prepositions), preserving only the content words (nouns, verbs, adjectives, adverbs) for our E2R dictionary. Our E2R dictionary contains a total of 13400 simple words. Comparing this dictionary to the datasets described at the Section 3.1.1 we found that there is 37% and 20% of coincidence on the training and test datasets respectively .
- **Word Embedding Features:** For this feature, the system uses the vectors from two different word embedding models. A Word2Vec model trained on The Spanish Billion Words Corpus (Cardellino, 2016) with a total of 1,000,653 words. We also exploit a word embedding model (Grave et al., 2018) trained on Common Crawl and Wikipedia with the FastText tool with character n-grams of length 5. In both models, word vectors have dimension 300.

3.2 Substitute Generation

The substitute generation process consists of selecting candidate substitutions for complex words taking into account the whole context of the word that can be had. Despite the few synonym dictionaries for Spanish language, we follow a strategy of linguistic database querying. In particular, we exploit resources such as Babelnet (Navigli and Ponzetto, 2010) and Thesaurus (http://thesaurus.altervista.org/). We use a REST API to obtain the synonyms from Thesurus, and a Python API from Babelnet.

In this stage, we only process those words that were annotated as complex ones. Then, for each word, we obtain its set of synonyms from BabelNet and Thesaurus.

Using the CWI process described above, we also filter out those candidates identified as complex words from the final list of synonym candidates.

3.3 Substitute Selection

The substitute selection stage takes the list of synonyms extracted from the previous step. Now, the most suitable should be selected according to factors such as simplicity and its context.

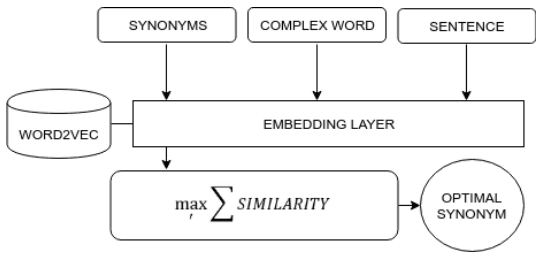


Figure 2: Substitute Selection Step

Our approach considers the context of the word in the sentence and then evaluate the similarity between words and its context

words. Our hypothesis is that considering the context may help to propose an optimal synonym for that specific sentence.

Figure 2 shows this step in the simplification process. For each complex word, the system takes the list of its synonym candidates from the previous step. Then, we calculate the cosine distance between the input word and each of its synonym candidate. In a word embedding model, words are represented as numerical vectors in a low dimensional space. These vectors are able to represent the context of the words, capturing their semantic and syntactic similarities. We assume that similar words will have close vectors. Our hypothesis is that the candidate with the closest vector should be finally selected. Moreover, we also calculate the cosine distance between each candidate and the context words (previous and next words). To obtain these similarities, we exploit the Spanish Billion Words Corpus word embedding model. The three cosine distances are summed, selecting the candidate with a maximum score.

4 *Evaluation*

As it was explained in Section 3.1, we use the datasets provided by the CWI 2018 shared task.

The results are shown as the competition (Yimam et al., 2017) requests in their binary classification task. This is scored using the macro-averaged harmonic mean (F1-score) of precision and recall.

Independent Feature	**F1**
Lenght-Frequency	0.6614
Boolean	0.4485
E2R	0.4011
Word2vec	0.7012
FastText	0.4901

Table 1: F-1 scores for every feature alone

Table 1 shows feature F1-scores independently and Table 2 shows the scores obtained in the CWI task, sorted by feature addition. Using only the length and frequency features, an F1 score of 0.6614 is obtained. These were selected with the aim of having a baseline system, since most of the participating systems the CWI shared tasks in SemEval2016 and BEA 2018 Workshop used them. The features of the length of the word, the number of syllables and the length of the sentence are basic elements in a system due to the fact that this feature are very discriminative. On the other hand, the frequency features used depend to a great extent on the vocabulary of the corpus, since it depends on the occurrence of the terms in the corpus. One way to solve this problem is to enlarge the corpus.

The boolean features were used to represent the morphology of a word, by checking if a word has numbers or special characters, among others. With these features, an improvement of 0.0508 was obtained with respect to the baseline system, due to the fact that the Spanish language contain some words with special characters, being potentially complex words. This is confirmed on Table 1 where the features are evaluated independently.

The E2R dictionary provides a significant improvement of almost 2 points, obtaining an F1 of 0.7314. This dictionary may be a valuable resource for other researches because it gathers a collection of simple words verified by human experts. Although the E2R dictionary is small, it shows a beneficial increase in performance. Additionally, Table 1 shows a F1 of 0.4011 by the feature, this is because there is an important percentage of word coincidence between the dictionary and the datasets of the task.

Likewise, the use of word embedding vectors also provides a significant improvement compared to the baseline results , yielding a F1 score of 0.7341 with Word2vec and a final F1 score of 0.7497 with FastText. Evaluating these features independently shows that Word2vec by itself gets an score of 0.7012 and FastText 0.4901 as shown in Table 1. The main difference between these two libraries is that, FastText in addition to giving similarity information between words in the space, provides morphological information of the words when using n-grams bags. The interesting point is that, when using these two libraries together, a better F1-score of 0.7283 is obtained. This suggests that the two resources complement each other, by extending the dictionary or by being distinct embedding resources.

Additionally, in order to complement the independent scores, Table 2 shows the scores of some combinations between this features, helping us to determine which features are more discriminatory. The best scores are reached with the help of the vectors of the embedding models. Using Word2vec and

Feature	Accuracy	Precision	Recall	F1
L+F	0.6614	0.6819	0.6834	0.6614
B+L+F	0.7519	0.7887	0.7071	0.7122
E+B+L+F	0.7879	0.8015	0.7143	0.7314
E+B+L+F+W	0.7996	0.8544	0.7141	0.7341
E+B+L+F+W+F	0.8137	0.8636	0.7257	**0.7497**
W+F	0.7622	0.7920	0.7214	0.7283
E+W+F	0.7461	0.8250	0.6911	0.6982
E+L+F	0.7891	0.8095	0.7018	0.7205
B+E	0.7281	0.7205	0.7057	0.7097
B+E+W+F	0.7286	0.7299	0.7599	0.7299

Table 2: System CWI scores for feature combinations where L:Lenght, F:Frequency, B:Boolean, E:E2R, W:Word2Vec, F:FastText

FastText models, a F1-score of 0.7283 is obtained and adding the E2R and Boolean features, a F1-score of 0.7299 is obtained. With this final score the system has a better score than the baseline system in BEA Worshoshop for the CWI task.

SPANISH	F-1
TMU	0.7699
NLP-CIC	0.7672
ITEC	0.7637
Our approach	**0.7497**
NLP-CIC	0.7468
CoastalCPH	0.7458
CoastalCPH	0.7458
NLP-CIC	0.7419

Table 3: F-1 scores for the CWI task on BEA Workshop 2018

Finally, Table 3 shows the best seven results in BEA Workshop 2018 for the CWI task for the Spanish language. Comparing our best F1-score of 0.7497 to these participating systems, our system ranks fourth. TMU (Kajiwara and Komachi, 2018) presents a similar approach based on the frequency of the target word in a Wikipedia Corpus and a learner corpus, later on trained on random forest classifiers. NLP-CIC (De Hertog and Tack, 2018) presents a different approach by implementing a deep learning architecture with similar features to this work like word/char embeddings, word length and frequency counts. ITEC (Aroyehun et al., 2018) also implements a deep learning architecture, training a Convolutional Neural Network of three layers, the first two using a linear activation function and the last using a sigmoid activation function.

5 Conclusions and Future Work

The main objective of this work is the design and development of a lexical simplification system for the identification of complex words (CWI) and replacement of complex words with simpler synonyms in the Spanish language in a generic domain. The aim is to improve cognitive accessibility by increasing understanding and readability of texts.

To fulfill this objective, as a first step, we propose a supervised machine learning system using an SVM classifier. The experiments show that the use of the ER2 dictionary as well as the word embeddings provide a significant improvement, with a final F1 of 74.97%.

As future work, we plan to extend the feature set by adding information from Sense2Vec[1] and Char2Vec[2] models. On the part of the classifier, the use of Deep Learning techniques should be gauged. On the other hand, about easy-to-read (E2R) resources, other approaches are going to be considered such as a rule-based approach.

Finally, for the selection of substitutes stage, an evaluation with users is necessary, in order to demonstrate the level of satisfaction obtained with this approach in levels of readability and understanding ease.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and

[1] `github.com/explosion/sense2vec`
[2] `github.com/tannerbohn/char2vec`

Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R)

References

Aroyehun, S. T., J. Angel, D. A. P. Alvarez, and A. Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 322–327.

Baeza-Yates, R., L. Rello, and J. Dembowski. 2015. Cassa: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385.

Bott, S., L. Rello, B. Drndarevic, and H. Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. *Proceedings of COLING 2012*, pages 357–374.

Burstein, J., J. Shore, J. Sabatini, Y.-W. Lee, and M. Ventura. 2007. The automated text adaptation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 3–4. Association for Computational Linguistics.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March. https://crscardellino.github.io/SBWCE/.

Chayle, C., C. M. Herrera, M. A. Barrera, A. Pauletto, and S. Blanco. 2017. Evaluación de la accesibilidad web. In *XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires).*

De Hertog, D. and A. Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334.

Ferrés, D., H. Saggion, and X. G. Guinovart. 2017. An adaptable lexical simplification architecture for major ibero-romance languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47.

Freyhoff, G., G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken. 1998. Make it simple.

Glavaš, G. and S. Štajner. 2015. Simplifying lexical simplification: do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 63–68.

Gonzalez-Dios, I. 2017. Análisis de la complejidad y simplificación automática de textos. el análisis de las estructuras complejas en euskera. *Procesamiento del Lenguaje Natural*, (58):155–158.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Hartmann, N. and L. B. dos Santos. 2018. Nilc at cwi 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340.

Kajiwara, T. and M. Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199.

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL.*

Mitkov, R. and S. Štajner. 2014. The fewer, the better? a contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40.

Moreno, L., P. Martínez, J. Muguerza, and J. Abascal. 2018. Support resource based on standards for accessible e-government transactional services. *Computer Standards & Interfaces*, 58:146–157.

Navigli, R. and S. P. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of*

the 48th annual meeting of the association for computational linguistics, pages 216–225. Association for Computational Linguistics.

Paetzold, G. and L. Specia. 2015. Lexenstein: A framework for lexical simplification. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90.

Paetzold, G. and L. Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Paetzold, G. H. and L. Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Paetzold, G. H. and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Saggion, H. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Saggion, H., E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Smith, K., G. Hallam, and S. Ghosh. 2012. Guidelines for professional library/information educational programs-2012. *IFLA Education and Training Section, IFLA, The Hague, available at: www. ifla. org/publications/guidelines-for-professional-libraryinformationeducational-programs-2012 (accessed 25 August 2014)*.

Štajner, S., I. Calixto, and H. Saggion. 2015. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.

Štajner, S., H. Saggion, and S. P. Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

W3C, W. 2019. Web content accessibility guidelines (wcag) overview. https://www.w3.org/WAI/standards-guidelines/wcag/.

Yimam, S. M., C. Biemann, S. Malmasi, G. H. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Yimam, S. M., S. Stajner, M. Riedl, and C. Biemann. 2017. Multilingual and cross-lingual complex word identification. In *RANLP*, pages 813–822.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 103-110 recibido 12-04-2019 revisado 15-05-2019 aceptado 17-05-2019

Generación de resúmenes extractivos de múltiples documentos usando grafos semánticos

Multi-document extractive summarization using semantic graph

Oleyda del Camino Valle[1], Alfredo Simón-Cuevas[2], Eduardo Valladares-Valdés[2], José A. Olivas[3], Francisco P. Romero[3]
[1] Empresa Nacional de Software (DESOFT), La Habana, Cuba
[2] Universidad Tecnológica de La Habana "José Antonio Echeverría",
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
[3] Universidad de Castilla La Mancha, Paseo de la Universidad, 4, Ciudad Real, España
oleyda.camino@cfr.desoft.cu, {asimon, evalladares}@ceis.cujae.edu.cu
{JoseAngel.Olivas, FranciscoP.Romero}@uclm.es

Resumen: La generación automática de resúmenes consiste en sintetizar en un texto corto la información más relevante contenida en documentos, y permite reducir los problemas generados por la sobrecarga de información. En este trabajo se presenta un método no supervisado de generación de resúmenes extractivos a partir de múltiples documentos. En esta propuesta, la conceptualización y estructura semántica subyacente del contenido textual se representa en un grafo semántico usando WordNet y se aplica un algoritmo de agrupamiento de conceptos para identificar los tópicos tratados en los documentos, con los cuales se evalúa la relevancia de las oraciones para construir el resumen. El método fue evaluado con corpus de textos de MultiLing 2015, y se usaron métricas de ROUGE para medir la calidad de los resúmenes generados. Los resultados obtenidos se compararon con los de otros sistemas participantes en MultiLing 2015, evidenciándose mejoras en la mayoría de los casos.
Palabras clave: Generación de resúmenes multi-documentos, grafos semánticos, desambiguación, agrupamiento de conceptos

Abstract: The automatic texts summarization consists in synthesizing in a short text the most relevant information contained in text documents, and allows to reduce the generated problems by the information overload. In this paper, an unsupervised method for extractive multi-document summarization is presented. In this proposal, the conceptualization and underlying semantics structure of the textual content is represented in a semantic graph using WordNet, and a concept clustering algorithm is applied to identifying the topics of the documents set, with which the relevance of the sentences is evaluated to build the summary. The method was evaluated with texts corpus from MultiLing 2015, and ROUGE metrics were used to measure the quality of the generated summaries. The obtained results were compared with those other participant systems in MultiLing 2015, evidencing improves in most of the cases.
Keywords: multi-document summarization, semantic graph, word sense disambiguation, concept clustering

1 Introducción

En la actualidad, el crecimiento exponencial de la información textual disponible se presenta como una amenaza para su uso efectivo en la toma de decisiones, ya que, si bien su acceso parece estar garantizado, no ocurre lo mismo con el tiempo necesario para su análisis y procesamiento. Por tanto, la obtención de información relevante dentro de ese gran volumen de textos constituye un importante y desafiante objetivo alcanzar en ese escenario. La generación automática de resúmenes de textos es una de las líneas que actualmente se trabaja para enfrentar este desafío y reducir el impacto de la sobrecarga de información.

La generación automática de resúmenes tiene el propósito de condensar en un texto

ISSN 1135-5948. DOI 10.26342/2019-63-11 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

corto la información más relevante y esencial contenida en uno o varios documentos de textos (Das y Martins, 2007), considerando los criterios de concisión por parte del usuario (Padmapriya y Rajasekaran, 2012; Kumar y Salim, 2012), y preservando el contenido principal de la información, así como su significado general (Das y Martins, 2007; Padmapriya y Rajasekaran, 2012). Los resúmenes se pueden obtener mediante métodos extractivos (seleccionan las oraciones más importantes de los textos) o abstractivos (usan palabras o frases diferentes a las incluidas en los documentos originales) (Bhatia y Jaiswal, 2015), y a partir de enfoques supervisados, no supervisados o semi-supervisados (Gambhir y Gupta, 2017). Esta problemática ha sido ampliamente abordada, sin embargo, sigue siendo una tarea desafiante, sobre todo cuando se quiere obtener el resumen de múltiples documentos (Gambhir y Gupta, 2017).

En el trabajo se presenta un método no supervisado que combina enfoques basados en grafos y conceptos (Moratanch y Chitrakala, 2017) para generar resúmenes extractivos de múltiples documentos (o multi-documentos). En esta nueva propuesta, la conceptualización y estructura semántica subyacente del contenido de los documentos se representa mediante grafos semántico generados automáticamente, a partir de la identificación de conceptos y relaciones semánticas entre ellos desde WordNet (Miller y Fellbaum, 1998). Estos grafos se fusionan para obtener una única representación integrada del contenido de los documentos, soportado en la aplicación de un algoritmo de desambiguación para resolver las ambigüedades presentes. A partir del grafo que representa el conjunto de documentos, se aplica un algoritmo de agrupamiento de conceptos para identificar los tópicos más relevantes tratados en ese contenido, con los cuales se mide la relevancia de las oraciones a través de evaluar su similitud con los clústeres que representan los tópicos con un enfoque semántico. El uso de este tipo de grafos, inspirado en la propuesta de generación de resúmenes mono-documentos de Plaza y Díaz (2011), permite lograr mayor granularidad en el procesamiento semántico del contenido de los documentos, respecto a otras soluciones reportadas, así como identificar los temas principales en el contenido y generar resúmenes que se correspondan más a esos temas. También la inclusión de un proceso de desambiguación, como soporte a la integración del contenido de los documentos, constituye otra contribución a este ámbito, dado que ha sido poco tratado en otras propuestas reportadas.

El método propuesto fue evaluado con corpus de textos en español e inglés ofrecidos en MultiLing 2015, y la calidad de los resúmenes obtenidos fue medida usando métricas de ROUGE-N (Lin, 2004). Los resultados del método fueron comparados con los obtenidos por otros sistemas evaluados con esos corpus en el contexto de MultiLing2015, resultando mejores en la mayoría de los casos.

El trabajo se ha organizado de la siguiente forma: en la Sección 2 se sintetiza el análisis de los trabajos relacionados; en la Sección 3 se describe el método propuesto; en la Sección 4 se muestran y analizan los resultados experimentales obtenidos; y en la Sección 5 se exponen las conclusiones arribadas y líneas de trabajo futuro.

2 *Trabajos relacionados*

La generación de resúmenes multi-documentos consiste en crear de forma automática un texto corto con la información más relevante y esencial de un conjunto de documentos, en principio, relacionados a un tema específico. En estos procesos surgen problemáticas adicionales que repercuten en mayores complejidades y desafíos, respecto a la generación de resúmenes mono-documento, por ejemplo: mezcla de información inconexa y no relacionada (los documentos pueden no compartir los mismos temas), redundancia (los documentos pueden contener información en común), falta de coherencia (las unidades de información provienen de diferentes fuentes), entre otras (Gambhir y Gupta, 2017).

Los métodos extractivos de generación de resúmenes se han diseñado usando técnicas supervisadas y no supervisadas, y estas últimas han sido aplicadas con diferentes enfoques (Moratanch y Chitrakala, 2017). En la generación de resúmenes multi-documentos específicamente se han aplicado: técnicas basadas en grafos (los documentos se representan en modelos de grafos) (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013; Ferreira et al. 2014; Zore y Deshpande, 2014; Mirchev y Last, 2014; Yan y Wan, 2014; Al-Saleh y Menai, 2018); basadas en conceptos (se identifican conceptos de fuentes externas con los que se realiza el

análisis de relevancia) (Baralis et al., 2013; Sankarasubramaniam, Ramanathan, y Ghosh, 2014); lógica difusa (algunas características del contenido textual se procesan usando técnicas difusa, tales como: sistemas basados en conjuntos difusos y reglas de inferencia) (Bhoir y Gulati, 2015); LSA (*Latent Semantic Analysis*) (para procesar la estructura semántica subyacente de oraciones y palabras) (Steinberger, 2013); evaluación de características del contenido textual (Bhoir y Gulati, 2015; Naserasa, Khosravi, y Sadegh, 2018); métodos de agrupamiento (Zhong et al., 2017; Puspaningrum et al. 2018); entre otras. Según esta caracterización, el uso de grafos constituye uno de los enfoques más aplicados, siendo común en estas soluciones la representación de las oraciones como vértices en los grafos (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013; Ferreira et al. 2014; Zore y Deshpande, 2014; Al-Saleh y Menai, 2018), y el establecimiento de relaciones entre ellas a partir de medidas de similitud, en su mayoría basadas en aspectos sintácticos del contenido (Erkan y Radev, 2004; Hariharan, Ramkumar y Srinivasan, 2013), y no considerando aspectos semánticos (Ferreira et al. 2014). Sin embargo, este tipo de esquema de representación limita el poder realizar un procesamiento computacional del contenido con un mayor nivel de granularidad, por ejemplo, a nivel de conceptos. Esto resulta ser poco conveniente en soluciones que incluyen el modelado de tópicos como parte de la evaluación de la relevancia de las oraciones, donde los resúmenes se construyen con las oraciones que mayor representen los tópicos o temas principales tratados en los documentos, así como para darle solución los problemas de ambigüedad presentes en el contenido textual.

La ambigüedad de las palabras es otro gran problema que aún no se ha resuelto en el área de la construcción automática de resúmenes (Soriyan y Omodunbi, 2014), aunque este problema puede afectar en menor medida en la construcción de resúmenes mono-documentos. En el contexto de resúmenes multi-documentos, tal situación puede ocurrir con mayor probabilidad y, por tanto, tener mayor impacto en la calidad de los resultados (Ferreira et al. 2013). Inicialmente se puede identificar que una oración en un documento A puede tener un alto grado de similitud a otra del documento B (sin tener en cuenta las posibles ambigüedades en el contenido), y esto incluso puede sugerir que esa oración sea una buena candidata para el resumen, pero solo una de ellas debe ser incluida. Sin embargo, el tratamiento de la ambigüedad, puede resultar que esa semejanza aparente no sea tal, ya que se pueden estar usando términos con diferentes significados, logrando con ello una mayor precisión en el procesamiento y reduciendo la ocurrencia de redundancias. A pesar de lo argumentando, en la mayoría de las propuestas estudiadas no se incluye la desambiguación como parte del proceso de generación de resúmenes, esto solo se reporta en (Baralis et al., 2013). En (Baralis et al., 2013) los conceptos presentes en el texto son identificados y desambiguados usando la ontología Yago. El proceso de desambiguación se realiza analizando el grado de pertinencia de cada entidad de la ontología al contexto en el que se encuentran los conceptos ambiguos en el texto. El impacto de ambigüedad en la calidad de los resúmenes generados también se puede reducir usando medidas de similitud semántica, por ejemplo, entre oraciones en las soluciones basadas en grafos, pero esto tampoco ha sido común en este tipo de propuestas.

3 Método propuesto

El método fue concebido teniendo en cuenta los procesos que suelen considerarse en el diseño de este tipo de soluciones (Allahyari et al., 2017): (1) representación de los documentos en grafos semánticos; (2) fusión de los grafos semánticos; (3) agrupamiento de conceptos; (4) evaluación de relevancia de las oraciones; y (5) construcción del resumen.

3.1 Representación de los documentos en grafos semánticos

En esta fase se construye automáticamente un grafo semántico de cada uno de los documentos, en los cuales se representan conceptos presentes en el texto y las relaciones semánticas entre ellos, usando WordNet (Miller y Fellbaum, 1998) como referencia para la captura de esa información. Previo a la generación de los grafos, se aplican un conjunto de tareas básicas de procesamiento de lenguaje natural sobre los documentos, tales como: extracción del texto plano, segmentación en oraciones, análisis morfo-sintáctico y la eliminación de palabras desprovistas de significado (o *stop words*). En este pre-procesamiento se utiliza el analizador sintáctico

Freeling, dado que brinda soporte para procesar textos en español e inglés.

La construcción de los grafos parte de extraer de WordNet los sentidos (*synset*) de los conceptos incluidos en cada una de las oraciones, así como las relaciones semánticas de hiperonimia, hiponimia, meronimia y holonimia existentes entre ellos. A partir de cada uno de los *synset* identificados también se extraen aquellos *synset* incluidos en una vecindad de radio 2 con los cuales están relacionados (según los tipos de relación mencionados). Estos elementos capturados son usados para obtener el grafo que representa a cada oración, donde los *synset* constituyen los nodos (los conceptos ambiguos estarían representados en más de un nodo por sus *synset*) y las relaciones se establecen según los vínculos identificados en WordNet. Luego, los grafos generados son integrados para obtener un único grafo del documento, cuyo proceso se realiza unificando los *synset* comunes en los diferentes grafos de las oraciones.

3.2 Fusión de los grafos semánticos

El propósito de este proceso es obtener una representación del contenido del conjunto de documentos en un único grafo, a partir del cual se identifican los tópicos más relevantes con el algoritmo de agrupamiento que se aplica en la siguiente fase. Este proceso se lleva a cabo mediante la integración de los grafos semánticos que representan a cada uno de los documentos. La fusión de los grafos está basada en la identificación e integración de los conceptos representados que tienen el mismo significado. Por lo tanto, como paso previo es necesario resolver las ambigüedades presentes en los conceptos representados en esos grafos. En este sentido, se aplica el algoritmo de desambiguación reportado en (Hojas et al. 2018) sobre cada uno de los grafos de los documentos. Este algoritmo combina heurísticas de dominio, contexto y glosa para determinar el sentido más apropiado de un concepto representado en un grafo usando WordNet. A partir del análisis de cada heurística se le otorga un valor de peso a cada sentido posible, cuyos valores son agregados en una suma ponderada para obtener un valor de peso global, siendo el de mayor peso el seleccionado como sentido del concepto. Esta combinación de heurísticas reduce las limitaciones derivadas de aplicar una sola de ellas en este proceso (Navigli, 2009), como ocurre en el caso del algoritmo basado en glosa usado en (Plaza y Díaz, 2011).

El proceso de desambiguación incluido en esta propuesta constituye un factor clave para mejorar los resultados en la obtención de los resúmenes multi-documentos, aun cuando la mayoría de las soluciones similares estudiadas no lo consideran. Esto no solo permite evitar posibles redundancias (Ferreira et al. 2013), sino también contribuye a realizar un uso más eficaz de Wordnet, y con ello reducir los efectos de otros problemas que pueden padecer este tipo de soluciones como: mezcla de información inconexa o no relacionada, y la falta de coherencia en el resumen resultante.

Al concluir la desambiguación, son refinados los grafos que representan los documentos, eliminándose aquellos *synset* de los términos ambiguos que recibieron una menor votación como resultado del algoritmo aplicado. Finalmente, todos los grafos de los documentos son fusionados mediante la integración de los vértices que representen los mismos *synset*, obteniéndose un único grafo semántico que representa el contenido del conjunto documentos.

3.3 Agrupamiento de conceptos

En esta fase, los conceptos representados en el grafo resultante del proceso anterior son agrupados en clústeres, los cuales representan los tópicos principales abordados en el conjunto de documentos. Los conceptos centroides en estos clústeres son los que aportan la información más relevante sobre cada tópico. Este proceso se lleva a cabo usando un algoritmo de agrupamiento basado en la conectividad, similar al aplicado en (Plaza y Díaz, 2011). Los conceptos son agrupados de acuerdo al grado de conectividad entre ellos, considerando que, en el tipo de grafo que se genera en esta propuesta, unos pocos vértices estarán altamente conectados, mientras que el grado de conectividad del resto de los vértices será relativamente más bajo.

El algoritmo de agrupamiento inicia identificando el grado de conectividad de cada vértice del grafo, otorgando un valor de relevancia para cada uno mediante la suma de sus vértices adyacentes; con lo que se genera un ranking de vértices. A partir de este ranking son seleccionados lo *n* vértices de mayor relevancia (mayor grado de conectividad), siendo el valor

de *n* determinado por un usuario. Los vértices de mayor relevancia son denominados vértices *HUB* y son agrupados en conjuntos denominados *HVS* (*HUB Vertex Sets*), considerando que *HVS* es un conjunto de vértices fuertemente conectados entre sí y constituyen los centroides de los clústeres. El algoritmo de agrupamiento construye los *HVS* buscando, iterativamente y para cada vértice *HUB*, aquel vértice *HUB* más conectado a este, y los agrupa en un único *HVS*. Posteriormente, se comprueba para cada par de *HVS* si sus conectividades internas son menores que la conectividad entre ellos, agrupándose ambos *HVS* si esta condición se cumple. Finalmente, en un proceso iterativo se agrupan los vértices no clasificados como *HUB* al *HVS* correspondiente (con el que existe una mayor conectividad) para obtener los clústeres finales.

3.4 Evaluación de relevancia de las oraciones

En esta fase es evaluada la relevancia de cada oración presente en los documentos, obteniéndose un ranking a partir del cual se seleccionan las oraciones que conformarán el resumen. Este proceso se lleva a cabo mediante una evaluación de la similitud existente entre el contenido de cada una de las oraciones y los clústeres generados en la fase anterior, con un enfoque basado en el análisis semántico.

La similitud entre la oración y el clúster es evaluada combinando el uso de la medida de similitud semántica de Lin (1998) (incluida en el paquete de *WordNet::Similarity* (Pedersen, Patwardhan y Michelizzi, 2004)) para evaluar la similitud entre los conceptos de la oración y los incluidos en el clúster, con un mecanismo de ponderación similar al usado en (Plaza y Díaz, 2011), y calculada según la ecuación (1). En este caso, para cada uno de los conceptos (c_i) de la oración, se calcula su similitud semántica con cada uno de los conceptos (v_j) del clúster (C_j) (*sim_semántica*(c_i,v_j)), ponderado con un peso ($w_{i,j}$) diferente, según el tipo de vértice del clúster con el que se realiza el análisis. Considerando que los conceptos pertenecientes al *HVS* tienen mayor importancia que el resto, se considera el peso 1.0 ($w_{i,j} = 1$) si v_j pertenece al *HVS* del clúster y la mitad del peso ($w_{i,j} = 0.5$) en caso contrario. Los valores de similitud obtenidos son promediados para obtener el valor de relevancia global de las oraciones.

$$similitud(O_i, C_j) = \sum_{c_i / c_i \in O_i} \sum_{v_j / v_j \in C_j} w_{i,j} * sim_semántica(c_i, v_j) \quad (1)$$

3.5 Construcción del resumen

Luego de haber calculado la relevancia de las oraciones, en esta fase se procede a la construcción del resumen seleccionando las N oraciones con mayor relevancia, siendo N un parámetro que representa la tasa de compresión deseada. En este proceso de selección son considerados otros dos aspectos: el orden de los documentos en el que aparecen las oraciones más relevantes, y luego el orden en que aparecen dentro del documento. Estos criterios contribuirían a evitar inconsistencias en el resumen resultante.

4 Resultados experimentales

El método propuesto fue evaluado con los corpus de textos en español e inglés ofrecidos en MultiLing 2015 para evaluar soluciones en la tarea MMS (*Multilingual Multi-document Summarization*), siendo la última edición en la que ha sido incluida la generación de resúmenes multi-documentos como tarea de evaluación en este escenario. Los corpus seleccionados están formados por artículos de noticias provenientes de *WikiNews* asociados a 15 tópicos, y los mismos se han agrupado en dos colecciones: ENCol y SPCol, respectivamente. Ambas colecciones se caracterizan en la Tabla 1.

	Colecciones	
Características	**ENCol**	**SPCol**
Idioma	Inglés	Español
Cantidad de Corpus	15	15
Documentos en cada Corpus	10	10
Cantidad de Oraciones Totales	4469	5000
Prom. de Oraciones en Corpus	297,9	333,3
Prom. de Oraciones en Doc.	28,92	33,34

Tabla 1: Caracterización de las colecciones

Los resúmenes obtenidos son evaluados usando las métricas de precisión (P), *recall* (R) y medida-F (F), en el contexto de ROUGE. Específicamente, ROUGE-1 y ROUGE-2, teniendo en cuenta que ambas medidas funcionan relativamente bien en este tipo de tarea, según se reporta en (Lin, 2004). La Tabla 2 muestra los resultados obtenidos y su comparación con las soluciones participantes en la tarea MMS de MultiLing 2015 y evaluados

con los corpus seleccionados. También se incluye los resultados del sistema *Lead*, el cual fue propuesto como *baseline* para esta tarea.

Según se aprecia en la Tabla 2, los resultados obtenidos por el método propuesto superan los valores del *baseline* en cada una de las métricas y colecciones de prueba. Los valores obtenidos para ROUGE-1 también mejoran los del resto de los sistemas en ambas colecciones, lográndose un incremento superior al 20 %, con respecto a los valores promedio obtenidos por esos sistemas en cada una de las métricas. Constituye un resultado muy positivo también los valores superiores al 50 % obtenidos de *recall* y medida-F en las pruebas con SPCol. En el caso de las mediciones con ROUGE-2, los valores de *recall* y medida-F obtenidos en ambas colecciones fueron superiores a los reportados por el resto de los sistemas. En este caso, la precisión obtenida por el método no fue la mayor, pero el valor resultante es muy cercano al obtenido por los sistemas que mejores resultados reportaron, específicamente, BGU-MUSE (Litvak et al., 2016) y UWB (Steinberger, 2013). BGU-MUSE (Litvak et al., 2016) es un método supervisado, basado en la aplicación de algoritmos genéticos y entrenado con colecciones ofrecidas en MultiLing 2015. Esta cualidad constituye una desventaja frente al acercamiento no supervisado propuesto, si se desea generalizar su aplicación en otros dominios donde no se disponga de colecciones de entrenamiento. Por otra parte, UWB (Steinberger, 2013) se basa en la técnica *LSA* para modelar los tópicos y el tamaño del vector de las oraciones es usado como medida para evaluar su relevancia dentro de los tópicos. Aunque los resultados que obtiene son buenos, esta estrategia puede requerir que más de una oración exprese toda la información asociada a los tópicos, lo que puede constituir una limitación en algunas colecciones (Nenkova y McKeown, 2012). El método propuesto no se ve afectado por este problema dado que los temas son modelados de una forma más granular, representando e integrando estructuras semánticas de conceptos y aplicando luego un algoritmo de agrupamiento de conceptos.

Sistemas	**ENCol**						**SPCol**					
	ROUGE-1			**ROUGE-2**			**ROUGE-1**			**ROUGE-2**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SCE-Poly	.207	.196	.200	.127	.117	.121						
BUPT-CIST	.125	.110	.116	.019	.015	.017	.130	.122	.124	.028	.027	.028
BGU-MUSE	.294	.277	.284	**.188**	.173	.180						
NCSR/SCIFY	.154	.135	.142	.053	.043	.046	.226	.202	.211	.068	.059	.062
UJF-Grenoble	.142	.121	.129	.043	.034	.038						
UWB	**.344**	**.327**	**.333**	.186	**.178**	**.181**	**.412**	**.390**	**.399**	**.261**	**.247**	**.253**
ExB	.229	.227	.227	.084	.086	.085	.253	.242	.245	.098	.097	.097
ESIAllSummr	.159	.150	.153	.039	.034	.036	.208	.202	.204	.054	.055	.054
IDAOCCAMS	.230	.230	.229	.066	.068	.067	.252	.244	.246	.085	.081	.082
GiauUngVan	.147	.123	.134	.037	.027	.031						
Lead (baseline)	*.391*	*.409*	*.401*	*.111*	*.126*	*.116*	*.451*	*.458*	*.453*	*.158*	*.170*	*.165*
Solución propuesta	**.427**	**.460**	**.442**	.186	**.185**	**.185**	**.485**	**.558**	**.518**	.259	**.285**	**.271**

Tabla 2: Resultados experimentales

5 *Conclusiones y trabajo futuro*

En el trabajo se ha presentado un nuevo método no supervisado de generación de resúmenes extractivos multi-documentos. En este método se usan grafos semánticos capturados de WordNet para representar el contenido del conjunto de documentos, incluyendo la aplicación de un algoritmo de desambiguación que combina varias heurísticas. La aplicación de este algoritmo, no solo posibilita reducir las ambigüedades presentes en el contenido, sino también lograr mayor precisión en la integración del contenido de los documentos representados en los grafos y reducir la ocurrencia de redundancias en el resumen resultante. El uso de los grafos semánticos

basado en conceptos y relaciones entre ellos permitió lograr mayor granularidad en el procesamiento del contenido textual, así como a reducir la mezcla de información inconexa o no relacionada, y la falta de coherencia en los resúmenes generados. Por otra parte, también facilitó la identificación de los temas principales tratados en los documentos, a través del algoritmo de agrupamiento de conceptos aplicado, y con ello poder generar resúmenes sobre la base de la selección de oraciones relevantes para esos temas. Los experimentos realizados sobre colecciones de textos en español e inglés, arrojaron como resultado mejoras en la calidad de los resúmenes generados por esta propuesta, con respecto a otros sistemas. En todas las pruebas se superaron los valores de ROUGE del *baseline*, y en el caso de ROUGE-1 se superaron los resultados de todos los sistemas (en ambos idiomas). En el caso de ROUGE-2, los resultados obtenidos en ambas colecciones fueron, en general, superiores al resto de los sistemas, destacándose mejoras en la cobertura y en la medida-F, y alcanzándose una precisión muy cercana a la obtenida por los sistemas que mejores resultados reportaron.

En trabajos futuros se pretende evaluar el uso de otros recursos externos, como BabelNet, para generar los grafos, con el objetivo de lograr una mayor cobertura en la representación del contenido de los documentos. Además, se evaluarán otras alternativas para identificar los temas principales del contenido a partir de los grafos generados, por ejemplo, aplicando otros algoritmos dirigidos a determinar nodos relevantes en grafos. En este sentido, se considerarían también variantes de ponderación de las relaciones representadas en el grafo en función de su significado semántico, de tal forma que se logre un mayor aprovechamiento de esa información en este proceso.

Agradecimientos

Este trabajo ha sido parcialmente soportado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio Español de Economía y Competitividad, bajo la subvención del proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET) Ref. TIN2016-76843-C4-2-R (AEI/FEDER, UE)

Bibliografía

Allahyari, M., S., Pouriyeh, S., Safaei, E. D., Trippe, J. B., Gutierrez, y K., Kochut. 2017. Text Summarization Techniques: A Brief Survey, *Int. J. of Advanced Computer Science and Applications*, 8(10): 397-405.

Al-Saleh, A. y M. El B. Menai. 2018. Solving Multi-Document Summarization as an Orienteering Problem. *Algorithms*, 11(7):1-27.

Baralis, E., L. Cagliero, S. Jabeen, A. Fiori, y S. Shah. 2013. Multi-document summarization based on the Yago ontology, *Expert Systems with Applications*, 40:6976–6984.

Bhatia, N., y A., Jaiswal. 2015. Trends in Extractive and Abstractive Techniques in Text Summarization, *International Journal of Computer Applications*, 117(6):21-24.

Bhoir, A. S., y A. Gulati. 2015. A Multi-document Hindi Text Summarization Technique using Fuzzy Logic. *Int. J. of Advance Research in Science and Engineering*, 4(1):468-476.

Das, D. y A. F. Martins. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4: 192-195.

Erkan, G., y D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research*, 22:457-479.

Ferreira, R., L. S. Cabral, R. D. Lins, G. Pereira e Silva, F. Freitas, G. D.C. Cavalcanti, R. Lima, S. J. Simske, y L. Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, 40:5755-5764.

Ferreira, R., L. S. Cabral, F. Freitas, R. D. Lins, G. F. Silva, S. J. Simske, y L. Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41:5780–5787.

Gambhir, M., y V. Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1-66.

Hariharan, S., T. Ramkumar, y R. Srinivasan. 2013. Enhanced graph based approach for

multi document summarization. *Int. Arab J. Inf. Technol.*, 10:334-341.

Hojas, W., A. Simón, M. de la Iglesia, F. P. Romero, J. A. Olivas. 2018. A Concept-Based Text Analysis Approach Using Knowledge Graph. *Communications in Computer and Information Science*, 854:696-708.

Kumar, Y. J., y N. Salim. 2012. Automatic Multi Document Summarization Approaches. *Journal of Computer Science*, 8(1):133-140.

Lin, C.-Y. 2004. ROUGE: a package for automatic evaluation of summaries. En *Proceedings of the ACL-04 workshop*, páginas 74-81.

Lin, D. 1998. An information-theoretic definition of similarity. En *Proceedings of the International Conference on Machine Learning*.

Litvak, M., N. Vanetik, M. Last, y E. Churkin. 2016. MUSEEC: A Multi-lingual Text Summarization Tool. En *Proceedings of the 54th Annual Meeting of the ACL - System Demonstrations*, páginas 73 - 78.

Miller, G. y C. Fellbaum (Eds.). 1998. WordNet: An Electronic Lexical Database, The MIT Press: Cambridge, MA. 1998.

Mirchev, U., y M. Last. 2014. Multi-document summarization by extended graph text representation and importance refinement. En A. Fiori (Ed): Summarization Techniques: Revolutionizing Knowledge Understanding, IGI Global, páginas 28-53.

Moratanch, N. y S. Chitrakala. 2017. A Survey on Extractive Text Summarization. En *Proceedings of the IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)*.

Navigli, R. 2009. Word sense disambiguation: A survey, *ACM Computing Surveys*, 41(2):1-69.

Naserasa, A., H. Khosravi, y F. Sadegh. 2018. Extractive multi-document summarization based on textual entailment and sentence compression via knapsack problem. *Natural Language Engineering*, 25 (1):121–146.

Nenkova, A. y K. McKeown. 2012. A Survey of Text Summarization Techniques. En C.C. Aggarwal and C.X. Zhai (eds.), Mining Text Data, Springer, páginas 44-76.

Padmapriya, K. D. G., y V. G. Rajasekaran. 2012. A View On Natural Language Processing and Text Summarization. *Int. Journal of Communications and Engineering*.

Pedersen, T., S. Patwardhan y J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. En *Proceedings of the AAAI-04,* páginas 1024-1025.

Plaza, L., y A. Diaz. 2011. Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. *Procesamiento del Lenguaje Natural*, 47:97-105.

Puspaningrum, A., A. Nurilham, E. F. Bisono, K. Umam, y A. Z. Arifin. 2018. Inter and Intra Cluster on Self-Adaptive Differential Evolution for Multi-Document Summarization. *Journal of a Science and Information*, 11(2):86-94.

Soriyan, A., y T. Omodunbi. 2014. Trends in Multi-document Summarization System Methods. *International Journal of Computer Applications*, 97(16):46-52.

Sankarasubramaniam, Y., K. Ramanathan, y S. Ghosh. 2014. Text summarization using Wikipedia. *Information Processing & Management*, 50(3):443-461.

Steinberger, J. 2013. The UWB Summariser at Multiling 2013. En *Proc. of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, páginas 50-54.

Yan, S., y X. Wan. 2014. SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization. *IEEE/ACM Transactions on Audio, Speech, and Learning Processing,* 22(12):2048-2058.

Zhong, Y., Z. Tang, X. Ding, L. Zhu, Y. Le, K. Li, y K. Li. 2017. An Improved LDA Multi-Document Summarization Model Based on TensorFlow. En *Proceedings of the 2017 International Conference on Tools with Artificial Intelligence*, páginas 255-259.

Zore, A. S, y A. Deshpande. 2014. Extractive Multi Document Summarizer Algorithm. *IJCSIT*, 5(4):5245-5248.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 111-118 recibido 29-03-2019 revisado 08-05-2019 aceptado 15-05-2019

Applying Siamese Hierarchical Attention Neural Networks for multi-document summarization

Aplicando Redes Neuronales Siamesas con Atención Jerárquica para resúmenes multi-documento

José Angel González, Julien Delonca, Emilio Sanchis, Fernando García-Granada, Encarna Segarra

VRAIN Valencian Research Institute for Artificial Intelligence
Universitat Politècnica València
Camino de Vera s/n
46022 Valencia
{jogonba2, esanchis, fgarcia, esegarra}@dsic.upv.es

Abstract: In this paper, we present an approach to multi-document summarization based on Siamese Hierarchical Attention Neural Networks. The attention mechanism of Hierarchical Attention Networks, provides a score to each sentence in function of its relevance in the classification process. For the summarization process, only the scores of sentences are used to rank them and select the most salient sentences. In this work we explore the adaptability of this model to the problem of multi-document summarization (typically very long documents where the straightforward application of neural networks tends to fail). The experiments were carried out using the CNN/DailyMail as training corpus, and the DUC-2007 as test corpus. Despite the difference between training set (CNN/DailyMail) and test set (DUC-2007) characteristics, the results show the adequacy of this approach to multi-document summarization.
Keywords: Siamese Hierarchical Attention Neural Networks, multi-document summarization

Resumen: En este artículo presentamos una aproximación al problema de resumen automático multi-documento, basada en Redes Siamesas Jerárquico-Atencionales. El mecanismo de atención de las redes Jerárquico-Atencionales permite asignar un peso a cada frase en función de su relevancia en el proceso de clasificación. Durante la generación del resumen sólo se tienen en cuenta los pesos asociados a las frases para seleccionar aquellas más relevantes. En este trabajo exploramos la posibilidad de adaptar estos modelos al problema de resumen multi-documento (típicamente documentos muy largos donde la aplicación directa de redes neuronales no se comporta correctamente). Se ha experimentado utlizando el corpus CNN/DailyMail para entrenamiento, y el corpus DUC-2007 para evaluación. A pesar de la heterogeneidad de las características entre el corpus de entrenamiento (CNN/DailyMail) y el corpus de test (DUC-2007), los resultados muestran la adecuación de esta propuesta al resumen multi-documento.
Palabras clave: Redes Neuronales Siamesas Jerárquico-Atencionales, resúmenes multi-documento

1 *Introduction*

Nowadays, the development of automatic summarization systems is an important challenge, due to the necessity of tackling with the large amount of information that is accessible in the web or in other repositories. There are many applications that could be enriched with summarization systems, such as news and tourist information websites, seminars or conference abstracts, etc.

Although there are some attempts to address the problem of audio and video summarization, the main efforts until now have been done for developing systems that consider text documents as input. Different strate-

ISSN 1135-5948. DOI 10.26342/2019-63-12 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

gies to the summarization problem have been proposed (Lloret and Palomar, 2012) (Tur and De Mori, 2011) (Saggion and Poibeau, 2013). It must be distinguished among abstractive summarization, where the summary is composed by sentences that does not appear in the document but contain almost all the meaning; extractive summarization, where the summary consist of a selection of the more salient sentences of the document; and mixed summarization, where summaries are generated by combining abstractive and extractive methods (See, Liu, and Manning, 2017). Due to the difficulty of developing good abstractive and mixed strategies, most of the approaches are extractive. These approaches are a good solution in some tasks, such as summarization of news, because the journalistic writing style tends to contain the main information in some few sentences, that usually appear at the beginning of the article.

Related to methodologies, due to the difficulty of obtaining training corpus of document-summary pairs to train supervised systems, most of the initial works were based on unsupervised techniques. This is the case of the statistical word features extraction (Carbonell and Goldstein, 1998), the obtention of latent concepts by means of Latent Semantic Analysis (Deerwester et al., 1990), the graph based approaches such as LexRank (Erkan and Radev, 2004), among others (Tur and De Mori, 2011)(Lloret and Palomar, 2012). On the other hand, some systems based on supervised techniques were proposed when manually training corpus were built. This is the case of summarization based on Support Vector Machines (Begum, Fattah, and Ren, 2009) or Conditional random Fields (Shen et al., 2007).

In order to promote the comparison of different summarization techniques, some conferences were organized. Two of the most important were DUC[1] and TAC[2] conferences, where the corpus used for evaluation consists of news obtained from different press agencies. The summaries are provided by humans in both cases.

Given the success of deep learning methods for Neural Networks (NN) in many applications of language technologies, some attempts to apply these techniques to document summarization have been done (Cheng and Lapata, 2016) (Nallapati et al., 2016) (Nallapati, Zhai, and Zhou, 2017) (See, Liu, and Manning, 2017) (Paulus, Xiong, and Socher, 2017) (Narayan, Cohen, and Lapata, 2018). One of the problems for estimating accurate NN-based models is the availability of large and high-quality corpora. An important resource in this field is the CNN/DailyMail summarization corpus (Cheng and Lapata, 2016)(Nallapati, Zhai, and Zhou, 2017). It consists of news from CNN and DailyMail, and is composed of 312,084 document-summary pairs. Other corpora, as NewsRoom have been recently created (Grusky, Naaman, and Artzi, 2018). NewsRoom's summaries were written by authors and editors in the newsroom of news, sports, entertainment, financial, and other publications. To create the dataset, the NewsRoom's authors performed a Web-scale crawling of over 100 million pages from a set of online publishers.

In this paper, we present an approach to multi-document summarization based on Siamese Hierarchical Attention Networks (SHA-NN). One advantage of this kind of models is that they can learn from positive and negative samples, that in our case are document-summary pairs. A positive sample is a document and its corresponding summary, and a negative sample is a document and a summary of other document randomly chosen. This way, the model is trained as a binary classifier and it doesn't need any kind of apriori assignation of scores to sentences as it is the case of other NN-based summarization systems (Nallapati, Zhai, and Zhou, 2017).

For training purposes, the input of the Siamese network consists of document-summary pairs along with the information about if it is a positive or negative sample. The document is processed by a subnetwork and the summary is processed by the other subnetwork of the SHA-NN system. Furthermore, it has an attention mechanism that can be used to provide an score to each sentence of the input document. For test purposes, only a document is provided, that is, only a subnetwork is used, and the output is a weight associated to each sentence of the input document. This way, the summary is generated by a selection mechanism applied on the weighted sentences. Additionally, the training process converge in few hours, dif-

[1] https://duc.nist.gov/
[2] https://tac.nist.gov/tac

ferently from other NN-based systems that converge after several days. In a preliminary research we applied our system to the CNN/DailyMail corpus for single document summarization (González et al., 2019). The obtained results are in line with the state-of-the-art.

A limitation of NN-based approaches is that input documents must not be very long. This is due to the fact that current models have not enough capability to capture long term dependencies. Moreover, generally, there are some space and time constraints. Therefore, NN-based approaches can work reasonably well with short documents, as news of journals, but it is necessary to adapt them when longer documents must be summarized. This is the case of DUC-2007 summarization task, where each multi-document is composed by the addition of different short documents related to a topic. In order to address this problem with the SHA-NN system, in this work we have proposed an iterative process that successively provides the most salient sentences from shorter pieces of the multi-document until the 250 words length summary (required by DUC competition) is obtained.

2 Related work

The use of Deep Neural Networks have made substantial progress in many language technologies, such as extractive document summarization. Some initial works, such as (Cheng and Lapata, 2016), addressed the summarization process as a sequential binary classification problem where sentences are classified as candidates to be extracted or not. This is done by an encoder-decoder system enriched with an attention mechanism that is used to score the sentences. The sentences are encoded by Convolutional Neural Networks and Recurrent Neural Networks are used to score them, taking into account the encoded representations, and the previous labeled sentences. In (Nallapati, Zhai, and Zhou, 2017) the sentence selection is addressed as a sequence classification problem by using Hierarchical Attention Networks (Yang et al., 2016) but modeling more features than in the work of (Cheng and Lapata, 2016). Both works first assign a score to each candidate sentence, and then extract the most salient sentences.

Other recent work presents some variants of NN-based approaches to enrich the systems. This is the case of Wu and Hu, (2018) and Narayan, Cohen, and Lapata, (2018) were a reinforcement learning algorithm is applied, considering ROUGE Lin, (2004) measures as reward. In Zhou et al., (2018) a system where the sentence scoring and the selection mechanism are jointly learnt is presented. At each step during extraction, the sentence extractor reads the representation of the last extracted sentence, and uses it to score the relevance of the remaining sentences. Finally, in Al-Sabahi, Zuping, and Nadher, (2018) an attempt to take into ac-count the structure of the document as in-formation to be considered in the selection of sentences is presented. The model com-putes the score of each sentence by modeling several features as: content richness, salience with respect the document, redundancy re-spect the summary and the position in the document.

3 Corpus description

Since there are no large enough corpora to train complex supervised systems for multi-document summarization, we used the CNN/DailyMail corpus for training the SHA-NN system in order to evaluate it in a multi-document summarization task. The corpus was built from the journals news and the associated summary, consisting in some highlights manually done by journalists. It consist of 312,084 document-summary pairs and three sets were defined from it: a training set of 287,226 pairs, a development set of 13,368 pairs, and a test set of 11,490 pairs. The mean compression ratio is 14:1, i.e. the reference summaries have, in average, approximately 14 times less words than the articles.

The DUC-2007 corpus consists of a collection of newswire documents. Documents were organized in 45 topics, and each topic is composed by 25 documents. The summarization problem consist of obtaining a summary of 250 words for each topic. The average number of words for document is 11,927 and the mean compresion ratio is 50:1. The gold standard summaries were done by human experts and there are 4 summaries for each one of the 45 topics.

Some statistics of both corpora are shown in Table 1. It can be seen that the lengths of the articles and summaries are extremely high for DUC-2007 in compari-

	Sents	Words
CNN/DM Articles	28.2	765.4
CNN/DM Summaries	3.8	53.4
DUC-2007 Articles	1028.0	12065.0
DUC-2007 Summaries	13.1	244.0

Table 1: Average number of sentences and words of CNN/DailyMail and DUC-2007 corpora for articles and summaries

son to CNN/DailyMail. Concretely, the articles of DUC-2007 have 36 times more sentences than the articles in CNN/DailyMail. Something similar happens with the number of words, where DUC-2007 articles have 15 times more words than their counterparts in CNN/DailyMail. It seems that the sentences in the CNN/DailyMail are twice as long as in the DUC-2007. However, regarding to the summaries, although in DUC-2007 they have more sentences and words than the summaries in CNN/DailyMail, the proportionality between the lengths is lower than for the articles.

4 Siamese Hierarchical Attention Networks

The SHA-NN architecture is shown in Figure 1. The left subnetwork represents the model for the document, and right subnetwork is the model for the summary. Both subnetworks are Hierarchical Attention Networks (Yang et al., 2016) composed by Bidirectional Long Short Term Memory (BLSTM) (Hochreiter and Schmidhuber, 1997). For training purposes, the input of the Siamese network consists of both, the sequence of words $\boldsymbol{x} = \{\boldsymbol{w_{11}}, ..., \boldsymbol{w_{1W}}, ..., \boldsymbol{w_{T1}}, ..., \boldsymbol{w_{TW}}\}$ of the document, and the sequence of words $\boldsymbol{x'} = \{\boldsymbol{v_{11}}, ..., \boldsymbol{v_{1V}}, ..., \boldsymbol{v_{Q1}}, ..., \boldsymbol{v_{QV}}\}$ of the summary, as well as the information about if it is a positive or negative sample, that is coded as 0 or 1 in the output $\boldsymbol{y}$. The input words are coded as d-dimensional embeddings, that are estimated with a skip-gram model from the CNN/DailyMail corpus. The output of the word level are the vector representation of sentences (in Figure 1, $\boldsymbol{s}_i$ for the document, and $\boldsymbol{q}_i$ for the summary), and the output of the sentence level are the vector representations of the document $\boldsymbol{r}$ and the summary $\boldsymbol{p}$. The boxes labeled as α_i and β_i represent the attention mechanism that assigns a score to each word or a sentence in the document side and the summary side.

Finally, the vector representation of the document $\boldsymbol{r}$ and the summary $\boldsymbol{p}$ as well as the difference between them ($|\boldsymbol{r}-\boldsymbol{p}|$) are concatenated in an output layer with a softmax activation function that works as classifier, as shown in Eq 1.

$$\widehat{y} = softmax(W_{\widehat{y}}[\boldsymbol{p}, \boldsymbol{r}, |\boldsymbol{r} - \boldsymbol{p}|] + b_{\widehat{y}}) \quad (1)$$

When the system is working on summarization mode, only the left side of the full model is considered (the subnetwork that processes the input document). A forward pass is performed on it to obtain the attention output, α, associated to each sentence of the input document, which allows us to generate a ranking of the most salient sentences to build summaries. Although many approaches can be used to select the most relevant sentences, in this work we have chosen the sentences considering only these attention outputs.

5 Multi-Document summarization process

The straightforward application of NN-based models to multi-document summarization on extremely long documents has several drawbacks. First, this kind of models have not got enough capacity to capture long term dependencies on extremely long sequences. Moreover, the longer these documents, the more complex these dependencies are and then, models must be more complex to explain these dependencies. That is, more parameters have to be estimated and therefore, more training data are required.

In addition, generally, there are space and time constraints. The most known and useful strategy in order to train NN-based models efficiently by using mini-batch training mode, consists of truncating the input document to a limited length of words and sentences. However, by doing this, some fragments of the input are discarded without a relevance criterion and some of these fragments could be relevant for computing the output. This is specially important in multi-document summarization where documents are compositions of many short single documents.

In order to address this problem, we have developed an iterative mechanism that first

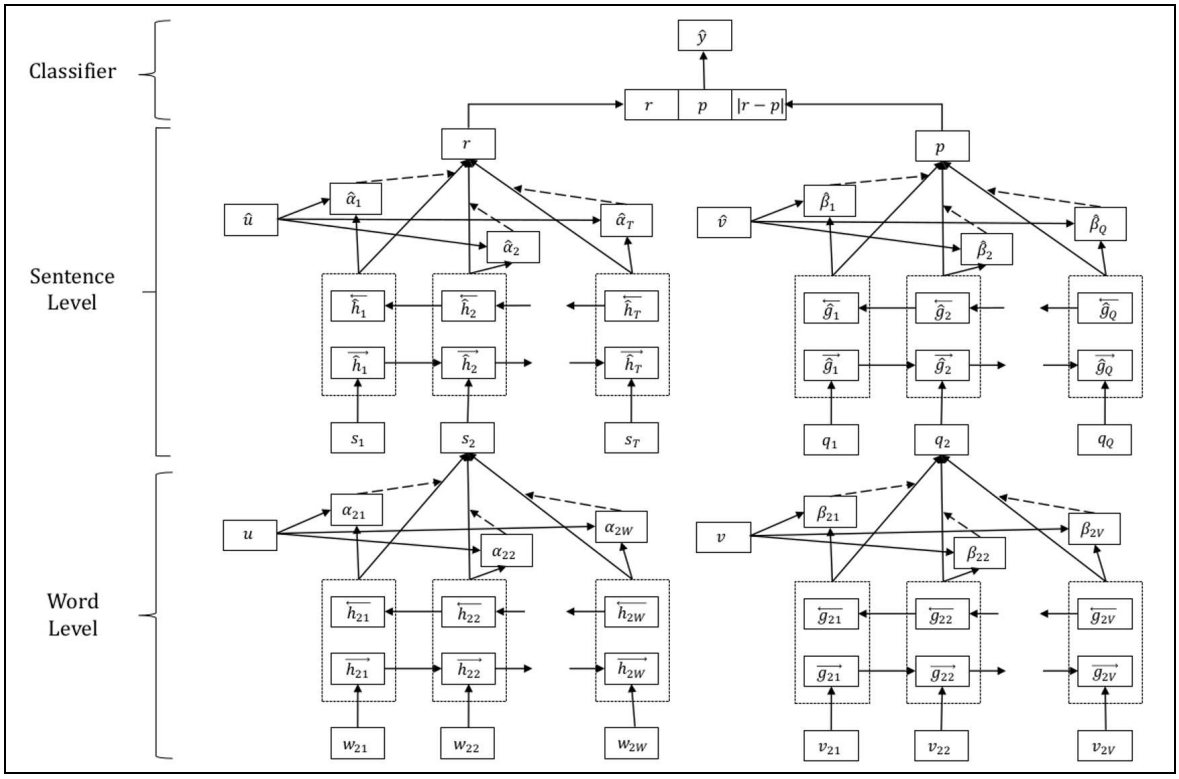


Figure 1: SHA-NN architecture proposed in (González et al., 2019)

obtains summaries from short fragments of the whole multi-document, and successively new summaries are generated from these previous summaries. Concretely, in each iteration of this iterative process, each fragment is separately summarized and then a new document is created for the new iteration by concatenating these summaries. This process is repeated until a summary of the desired length is obtained. The process is shown in Figure 2. Furthermore, this approach can be specially appropriate for the DUC corpus, because each long document is composed by the addition of shorter documents.

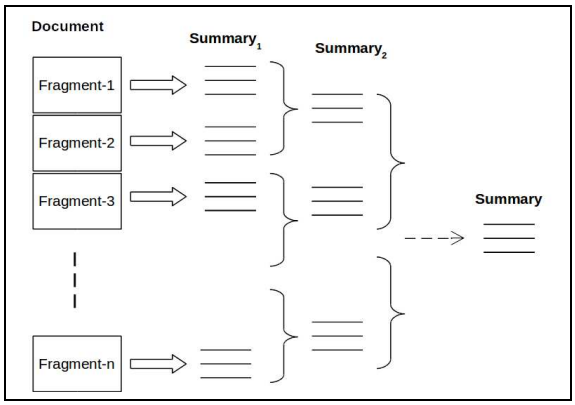


Figure 2: Iterative mechanism to obtain summaries from short fragments of the whole multi-document

6 *Experiments*

Some experiments were performed with the DUC-2007 corpus. Results were evaluated in terms of some ROUGE measures (Lin, 2004). In order to compare with other systems, we used the evaluation software given by the organizers of the competition, as well as the gold standard summaries also provided by them.

As the DUC-2007 corpus was designed only for evaluation, there is not possible to learn models with it. For this reason we have trained SHA-NN with the CNN/DailyMail corpus, and we have studied if a system trained with a type of corpus can generalize features that can be successfully applied for other types of corpora. One of the main differences between training and test corpus is the length of the input documents, that is much longer in the test. Another characteristic of the training corpus is that the named entities are anonimized, diferently than in the DUC corpus.

The summarization experiments were done as follows. For each topic, a first summary of each document belonging to this

topic is performed. To do this, each document is splitted in fragments of 20 sentences, and for each block, the 3 most salient sentences are extracted. The ratio between the number of input sentences of each fragment and the number of selected sentences for the summary, 20:3, is similar to the ratio considered in the training process with the CNN/DailyMail corpus. Once the multi-document associated to a topic is summarized, the obtained summaries are concatenated, and a new process of summarization is performed on that set of sentences in the same way. That is, the new set of sentences is splitted in blocks of 20 sentences and for each block the three most salient sentences are extracted. This process is done iteratively until arriving to 250 words, the length established by the competition. In the case of the experiments with the DUC-2007 corpus only two iterations have been performed, because after these two iterations the 250 word summaries were obtained. All this process is done for each one of the 45 topics.

Table 2 shows the results obtained in terms of Precision (P), Recall (R), and F-measure (F) for each one of the ROUGE measures. The results obtained are in line with those published in the DUC-2007 competition.

We did another experiment that uses a more simple way for selecting the summary sentences. In this experiment only one iteration of the summarization process is performed. This way, a set of sentences containing the three sentences of each fragment, as well as the associated score assigned by the SHA-NN system to each sentence, are obtained. These sentences are ranked considering their score, and they are sequentially selected until arriving to 250 words. Table 3 shows that results are slightly lower than those of the previous experiment. It seems that the iterative process takes advantage of the context generated by the salient sentences, instead of only consider the original context where relevant and not-relevant sentences participate.

	P	R	F
ROUGE-1	0.37098	0.37557	0.37204
ROUGE-2	0.07122	0.07240	0.07158
ROUGE-3	0.02209	0.02253	0.02225
ROUGE-4	0.01057	0.01074	0.01063
ROUGE-L	0.34084	0.34520	0.34190
ROUGE-W-1.2	0.18082	0.09862	0.12719
ROUGE-SU4	0.12767	0.12956	0.12819

Table 2: Results of the full iterative summarization process

	P	R	F
ROUGE-1	0.36946	0.37359	0.37113
ROUGE-2	0.06959	0.07028	0.06986
ROUGE-3	0.01958	0.01976	0.01965
ROUGE-4	0.00790	0.00800	0.00794
ROUGE-L	0.34123	0.34497	0.34274
ROUGE-W-1.2	0.18070	0.09844	0.12728
ROUGE-SU4	0.12366	0.12496	0.12418

Table 3: Results with only one iteration of the summarization process

7 Conclusions

We have presented in this work an approach to adapt a document summarization system based on Siamese Hierarchical Attention Neural Networks to a multi-document summarization task, the DUC-2007 competition. In the absence of an adequate enough large training corpus for this domain, the SHA-NN system has been trained on the CNN/DailyMail corpus, that presents some structural differences compared to the DUC-2007 corpus. It has been necessary to define a specific mechanism to allow the SHA-NN system to be applied to that multi-document summarization task. The results obtained with the DUC-2007 corpus are in line with those published in the DUC-2007 competition. As future works we will study if different reordering on the sentences obtained at each iteration can improve the results. It can be also interesting to use more information than just the score assigned to each sentence for selecting the most salient sentences. For example, in order to avoid including similar sentences in the summary, some distance among the candidate sentences can be taken into account.

Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R). Work of José-Ángel González is also financed by Universitat Politècnica de València under grant PAID-01-17.

References

Al-Sabahi, K., Z. Zuping, and M. Nadher. 2018. A hierarchical structured self-attentive model for extractive document summarization (hssas). *IEEE Access*, 6:24205–24212.

Begum, N., M. Fattah, and F. Ren. 2009. Automatic text summarization using support vector machine. 5:1987–1996, 07.

Carbonell, J. and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336.

Cheng, J. and M. Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Erkan, G. and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

González, J.-Á., E. Segarra, F. García-Granada, E. Sanchis, and L.-F. Hurtado. 2019. Siamese hierarchical attention networks for extractive summarization. *Journal of Intelligent & Fuzzy Systems (JIFS)*, To be published.

Grusky, M., M. Naaman, and Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lloret, E. and M. Palomar. 2012. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41.

Nallapati, R., F. Zhai, and B. Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3075–3081.

Nallapati, R., B. Zhou, C. N. dos Santos, Ç. Gülçehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*, pages 280–290. ACL.

Narayan, S., S. B. Cohen, and M. Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of NAACL-HLT*, pages 1747–1759. Association for Computational Linguistics.

Paulus, R., C. Xiong, and R. Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Saggion, H. and T. Poibeau, 2013. *Automatic Text Summarization: Past, Present and Future*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

See, A., P. J. Liu, and C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Shen, D., J.-T. Sun, H. Li, Q. Yang, and Z. Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2862–2867.

Tur, G. and R. De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech.* John Wiley & Sons.

Wu, Y. and B. Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

Zhou, Q., N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *CoRR*, abs/1807.02305.

Procesamiento del Lenguaje y Técnicas Basadas en Grafos

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 121-128 recibido 29-03-2019 revisado 14-05-2019 aceptado 14-05-2019

Benchmarking Entity Linking for Question Answering over Knowledge Graphs

Evaluación del Enlazado de Entidades para Sistemas Pregunta-Respuesta sobre Grafos de Conocimiento

Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas
Natural Language Universidad Nacional de Educación a Distancia (UNED)
{gblanco, alvarory, anselmo}@lsi.uned.es

Abstract: Entity Linking (EL) is the process of anchoring a part of a question to a node (entity) already known in a Knowledge Base (KB). Although EL has been widely studied with large documents such as webpages, there have not been studies about its impact on Question Answering (QA). In this paper, we study benchmarks for QA and how they are composed, providing insights about its suitability for a real evaluation about the state of the art in QA, specillay if we want to take into account the subtask of EL. We propose a semi-automatic method to generate an EL dataset linked to the QA task taking advantage of pre-existing QA datasets. We apply this method to benchmarking QA collections, analyze the results and release the created dataset to the research community, including a subset focused on complex EL in QA. We believe that EL effectiveness in the context of QA can be better assessed through the use of the proposed dataset.
Keywords: Question Answering, Knowledge Bases, Entity Linking, DBPedia

Resumen: El Enlazado de Entidades (EE) consiste en asociar partes de un texto con nodos de una Base de Conocimiento (BC). A pesar de que se ha prestado bastante atención a la tarea de EE en documentos, apenas hay estudios relativos a su impacto en el campo de la Búsqueda de Respuestas (BR). En este trabajo estudiamos la composición de varias colecciones de BR y realizamos varias observaciones relativas a su adecuación para evaluar sistemas BR, especialmente en lo relativo a realizar EE. También proponemos un método semiautomático para crear colecciones de EE en el contexto de BR reaprovechando colecciones existentes de BR. Posteriormente, aplicamos nuestro método a varias colecciones actuales de BR, analizamos los resultados obtenidos y ponemos a disposición de la comunidad científica la colección de EE generada, incluyendo un subconjunto que contiene los ejemplos donde es más difícil realizar EE. Consideramos que la disponibilidad de esta nueva colección permitirá una mejor evaluación de la tarea de EE en el contexto de la BR.
Palabras clave: Búsqueda de Respuestas, Bases de Conocimiento, Enlazado de Entidades, DBPedia

1 Introduction

Digital information grows everyday. As a consequence, users must face an enormous amount of data when searching for specific information. A great amount of this information is stored as structured data inside Knowledge Bases (KB), databases, etc. A desirable feature of such structured data is the ability to query it in a user-friendly manner, or even better with natural language questions.

Question Answering (QA) systems over Knowledge Graphs (KG) receive Natural Language questions and return answers from structured databases such as DBpedia (Lehmann et al., 2014) or Freebase (Bollacker et al., 2008). For this purpose, these systems must translate the natural language question into a structured query understandable using, for instance SPARQL (Shekarpour et al., 2016).

One challenge of these QA systems is to link parts of the Natural Language question to their corresponding nodes in the Graph.

ISSN 1135-5948. DOI 10.26342/2019-63-13 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

When the element to be linked is an entity, the task is called Entity Linking (EL) and is usually at the upfront of any Question Answering System (Unger et al., 2014). More formally, EL is the process of mapping a given string containing an entity, onto a resource of the target KB.

In general, mapping an entity based on a string comprises two steps: (1) recognizing the entity mention in the text and (2) disambiguating it. Let us consider the following example[1]: given the question *Who all have been a manager of english under twenty one football team?*, we would like to map the string *english under twenty one football team* to its corresponding entity in a KG (we use DBPedia here), which in this case is *dbr:England_national_under-21_football_team*. To do so we need to recognize the entity mention (*english under twenty one*) and map it to the correct KB resource (disambiguation). In many cases, this process implies lemmatizating words (e.g. *English* → *England*), parsing numbers (*twenty one* → *21*), dealing with missing words (*national* is not present at all in the source question) or even considering alias with higher lexical gap.

The process of EL is just one part inside the QA pipeline, although very important: linking to a certain node will restrict the search process inside the KG, yielding poor results when the entity is linked incorrectly or simply not linked at all. Therefore, it is important to properly asses the impact of EL over QA Systems. For this purpose, we need a procedure for EL benchmarking in the context of QA. However, there is no collection, to the best of our knowledge, focused on evaluating EL for QA over KGs.

In this paper, we describe a novel method for building a collection oriented to evaluate EL in QA. Our method reuses collections of QA. So, we show the results of applying our method to common collections.

The main contributions of this work are:

- We characterize the main QA datasets
- We propose a semi-automatic method to generate an EL dataset linked to the QA task
- We create and release a large benchmark dataset for EL in QA
- We create and release a subset focused on complex EL in QA

We have found that the standard collections used to evaluate QA do not pay so much attention to the EL task. We show how the EL task over these collections is too easy given the nature of the collections. Thus, we consider that this work points out potential problems in the QA datasets. Moreover, we release a dataset focused on EL in QA, where the more complicated examples are already detected. This dataset can help other researchers to improve their systems.

The paper is structured as follows: Section 2 analyses the work done up to date with respect to Question Answering over Knowledge Graphs datasets. We thoroughly describe the task we propose the benchmark for. In section 3 the technical aspects of the dataset are explained, as well as its construction and peculiarities. Section 4 describes the proposed method to generate EL datasets and the baseline against which we measured it. Section 5 shows the results we obtained with our method and opens a debate about Entity Linking datasets. We further extend the discussion in section 6. In section 7 some details about the released resources are given. We finish the article with some conclusions and further research in section 8. A brief description of the scripts employed for every part of the process as well some pointers to the source code can be found on the appendix (A).

2 Related Work

We have not found collections of EL related to QA. This is why, in this Section, we focus on collections created for evaluating EL.

The Web People Search (WePS) Evaluations proposed a task similar to that of Word Sense Disambiguation (WSD), consisting in disambiguating person names based on internet results or wikipedia pages (Artiles et al., 2010). The organizers provided more than 4.5K documents with human annotations.

The Text Analisys Conference (TAC) organize an EL task since 2009. Systems participating in this task have to link entities in texts with entities in a KB. Organizers provide in each edition a new set of documents with human annotations (Mcnamee and T Dang, 2009).

These works try to link or disambiguate

[1] Taken from LC-QuAD (Trivedi et al., 2017), qid:1

entities given a big context (webpages or documents). However, in QA the available contextual information consists solely of a natural language question. This is why we now focus on benchkmarks related to EL with smaller contexts. This problem is known as short-text EL (Chen et al., 2018).

Webscope[2] is a good example of a collection for short-text EL. Webscope contains more than 2k search-engine queries where humans have identified links of entities to Wikipedia articles.

The dataset of the #Microposts2016 Named Entity rEcognition and Linking (NEEL) Challenge contains more than 6k tweets with entities annotated by humans and linked to DBpedia (Rizzo et al., 2016). In the scope of Twitter, there exists also the MSR-TEL dataset (Guo, Chang, and Kıcıman,). This dataset contains less than 1k tweets annotated with entities in the text linked to Wikipedia articles.

All these datasets are usefull for evaluating EL, but they cannot be applied to measure the impact of EL in the context of QA. Even though we could evaluate the EL task in a similar scenario to QA over KG, we still suffer from the lack of assessing the effect of EL performance in the final QA system. We think this fact hinders evaluating the impact of the EL task over the whole performance of a QA System and motivates our work in this paper.

3 Description of the QA datasets

There are many large datasets oriented to evaluate QA over KGs. In these datasetes, a sample or datapoint consists of a natural language question accompanied with the SPARQL Query that gathers the correct answers. The queries are attached to a particular KG such as DBPedia or Freebase. For instance, QALD (Unger et al., 2014) and LC-QuAD (Trivedi et al., 2017)) are two important datasets for QA over DBPedia. An example from QALD is depicted in listing 1.

Dataset creators employed different methods. In the case of QALD, the creators manually compiled the questions from real-world question and query logs (Unger et al., 2014). LC-QuAD, on the other hand, went for an automatic system that creates questions based on a few given templates and a target KG

[2]http://webscope.sandbox.yahoo.com/

```
{
  "id": "37",
  "query": {
    "sparql": "SELECT ?uri ?string WHERE{
         ?uri dbo:series dbr:The_Sopranos
         . ?uri dbo:seasonNumber 1 .
        OPTIONAL {?uri rdf-schema#label ?
        string . FILTER (lang(?string) =
        'en') } }"
  },
  "answers": {
    "answer": [{
      "uri": "http://dbpedia.org/resource
          /46_Long",
      "string": "46 Long"
    },{
      "uri": "http://dbpedia.org/resource
          /A_Hit_Is_a_Hit",
      "string": "A Hit Is a Hit"
    },
    ...
    {
      "uri": "http://dbpedia.org/resource
          /College_%28The_Sopranos%29",
      "string": "College (The Sopranos)"
    }]
  },
  "question": [
    {
      "string": "List all episodes of the
           first season of the HBO
          television series The Sopranos
          !",
      "language": "en"
    }
  ]
}
```

Listing 1: Simplified sample from the original QALD 1 dataset.

(Trivedi et al., 2017)

Table 1 shows some basic details, like the number of total questions, number of questions containing entities, average number of entities per question and total number of unique entities covered by each dataset. The difference in size between QALD 1 through 3 is notable, but the biggest one appears between LC-QuAD and the rest, being one order of magnitude. This difference accounts for the fact that LC-QuAD was semi automatically constructed. It is to be noted that the first edition of QALD features a rather low number of questions containing entities.

In Table 2 we depict the identified EL casuistry by dataset. The the first row of the Table shows the number of questions for which an exact match between the mention in the question and the resource is present. We find this number surprisingly high across all analysed datasets. Also, we acknowledge that the number of different EL cases

```
{
  "id": 2,
  "question_id": 37,
  "question": "List all episodes of the
      first season of the HBO television
      series The Sopranos!",
  "dbr": "http://dbpedia.org/resource/
      The_Sopranos",
  "mention": "The Sopranos"
}
```

Listing 2: Sample from QALD 1 dataset after processing for Entity Linking. The mention must exclude punctuation as it is not present in the resource.

in QALD grows with the editions of the task, being QALD-4 the most richer in terms of number of different cases. Given the size of LC-QuAD dataset, we have excluded it from the manual cases classification.

Based on the observation that lots of mentions can be directly inferred from the entity uri (DBPedia resource, or *dbr*), we propose a custom method to automatically extract the part of the question (mention) that gives the best hints to the database resource. We present this method in the following sections.

Dataset	Q	Q-E	Avg E	U. E
QALD-1	50	12	1.25	13
QALD-2	100	72	1.17	70
QALD-3	100	72	1.24	72
QALD-4	200	160	1.27	162
LC-QuAD	5000	5000	1.32	3962

Table 1: Number (Q)uestions, (Q-E) Questions with entities, (Avg E) Average entities per question and number of (U. E) Unique Entities by dataset.

4 *Method for EL dataset generation*

We propose a method for the generation of a new dataset for EL where a sample comprises: 1) the original natural language question; 2) one of the entities pertaining to the question and 3) the part of the questions that relates to the entity (mention). A sample is depicted in listing 2.

Since a question may contain various entities in the original QA datasets, our method can create several samples for the same question. To build each sample we need to:

1. Identify the mention of the entity in the question.
2. Disambiguate this mention according to the KG.

However, this last step is trivial because such information is already provided by the QA dataset itself. Therefore, we only need to focus on the task of mention detection. In this case, we can take the expected uri as a clue.

4.1 Trigram-based mention detection

We propose a method to construct and semi-automatically annotate an EL dataset based on any QA dataset that provides a set of natural language questions with their corresponding SPARQL queries that retrieves answers from a KG. The fundamental challenge we address here is *identifying the mention* in the Natural Language Question (NLQ) that tallies with a resource in the Knowledge Base. These mentions can be used later to aid in the disambiguation process of any EL system.

We are going to apply a strategy consisting in grouping contiguous highly overlapping trigrams character sets from the question, similar to labeling the mentions with a BIO scheme[3]. This way, we obtain a set of candidate mentions from which we choose the most promising one and apply a simple alignment to clean it.

Also, as a preprocessing step for every text, we apply a series of regular expressions that trim characters like commas and periods, substitute underscores and hyphens with spaces and trim multiple consecutive spaces (*clean* procedure used in alg. 1).

The proposed method is depicted by algorithm 1 that works as follows:

1. Apply the previously explained simple clean procedure both to the question and the entity.
2. Divide every token in trigrams and get the possible mentions. If no mention is found, return the word with the maximum number of matching trigrams (*find_mentions* procedure, alg. 2). The probability of a match between trigrams

[3]The BIO scheme suggests to learn classifiers that identify the **B**eginning, the **I**nside and the **O**utside of the text segments

EL Casuistry	QALD-1	QALD-2	QALD-3	QALD-4
Mentions identical to DBP uri	11	72	75	160
Typographic		1	1	1
Missing tokens in the mention		4	5	20
Additional tokens in the mention	3	1	1	1
Synonyms			1	1
Acronyms				2
Lexical variation	1	5	5	17
Lexical variation + reordering		1	1	
Total	15	84	89	203

Table 2: Identified Entity Linking Typographic quantities by dataset.

Algorithm 1: Dataset construction

```
Data: QuAD.json
Result: QuAD-EL.json
mentions ← []
for point in Questions do
    question ← clean(point.question)
    entities ← clean(point.entities)
    for entity in entities do
        best_score ← ∞
        best_candidate ← null
        candidates ←
         find_mentions(question,entity)
        for candidate in candidates do
            score ← distance(entity,
             candidate)
            if score < best_score or
             (score == best_score and
             len(candidate) <
             len(best_candidate))
             then
                best_score ← score
                best_candidate ←
                 candidate
            end
        end
        mentions ← mentions ∪
         best_candidate
    end
end
write(mentions)
```

sets is measured as the number of trigrams that match divided by the total number of trigrams.

3. Get the mention that minimizes Levenshtein distance (Levenshtein, 1966) to the entity with the minimum length (*distance* procedure).

Algorithm 2: Find Mentions

```
Data: question, entity
Result: mentions
mentions ← []
mention ← []
e_grams ← ngrams(entity)
best_prob ← 0
best_index ← null
best_score ← ∞
for word, index in question do
    w_grams ← ngrams(word)
    /* get matching grams        */
    max_common ← max(1,
     len(w_grams))
    common_tris ← 0
    for gram in w_grams do
        if gram in e_grams then
            common_tris ←
             common_tris +1
        end
    end
    prob ← common_tris/max_common
     if prob > 0.7 then
        mention ← mention ∪ word
    else
        if prob > best_prob then
            best_prob ← prob
            best_index ← index
        end
        mentions ← mentions ∪
         mention mention ← []
    end
    if mention is not Empty then
        mentions ← mentions ∪
         mention
    end
    if mentions is Empty then
        mentions ← question
         [best_index ]
    end
end
return mentions
```

4.2 Baseline mention detection

In order to asses the effectiveness of our method, we have also developed a simple and fast baseline method.

The method searches in the question for a contiguous span with size in tokens equal to that of the database resource such that the Levenshtein distance between the span and the resource is minimized.

This is a simple baseline that, even though it works only at the lexical level, is robust enough.

5 Results

We manually annotated and classified the cases by type for all the questions. Then, we evaluated the proposed method and the baseline. We acknowledge that the number of exact matches found is high on all datasets. This fact arises some questions about the quality of the datasets that will be discussed in the discussion section (sect. 6).

The obtained results are depicted on table 3. Because both methods yielded the same exact mention matches, that row is omitted from this table (and showed only in table 2). Table 3 shows the number of correctly extracted mentions by each method on each dataset, classified by EL case type. At a first sight, the proposed method out stands in all scenarios. From the second error type (*missing tokens in the mention*), we observe that the baseline does not adjust correctly when the question lacks part of the resource. It is clear that this happens because the method is only allowed to choose segments of the same size as the resource, yielding more words when it should not.

Among the failed questions we have found some questions from which a mention linking the entity and the natural language query cannot be established directly. In the process of constructing the LC-QuAD dataset some questions got transformed in a way that the SPARQL Query is unrelated to the question itself, although this accounts for a small part of the dataset (about 6 questions).

Regarding the other errors, the first and most obvious is related to spelling. In English there are no accents, in the automatic construction step most of them got stripped off, resulting in misspelled words. Also, there are lots of abbreviations and acronyms which hampers joining on matching trigrams without an expansion mechanism. When speaking or writing in many languages we usually inflect words, but on the other hand, DBPedia stores entities in its lemmatized form, rising the necessity to analyse the words further.

The *lexical variation* error type is a good example of this effect. Let us consider *Which monarchs of the United Kingdom were married to a German?*[4] as the source question, one necessary resource to answer the question is the one that relates to *German*, but the database resource is *dbr:Germany*. This implies that any given system must be able to select it as the mention.

For further details, refer to each annotated dataset, which contains the list of questions classified by error type.

6 Discussion about Entity Linking in QA Datasets

In our study, we have found that the number of questions containing the exact mention to the resource is significantly high (more than a 70% in each QALD dataset and an 80% in LC-QuAD). Thus, it is unclear if these datasets might be helpful for training QA systems working in real environments, where users may include questions using only a surname, part of the entity, etc.

We think that this issue arises as a consequence of how collections are created. All the analysed datasets automate parts of the process to some extent. While QALD datasets use logs and already available queries as the basement to formulate questions, LC-QuAD uses a slot-filling + templates technique to generate queries and then translate to NLQ.

We have also reviewed WebQuestions (Berant et al., 2013), which was created from non-experts using questions that begin with a wh-word and contain exactly one entity. The answers were searched using Freebase. In this dataset, 73% of questions contain the exact mention of the entity in Freebase. Therefore, these datasets do not pay too much attention to the problem of Entity Linking.

Given the big amount of cases where we can extract exact mention matches, we envision that a resource containing all the questions containing different cases could be valuable for the research community. It can be useful to assess the effectiveness of EL systems against a complex dataset. This will be

[4] Taken from QALD 3, qid: 38

Cases	QALD-1		QALD-2		QALD-3		QALD-4	
	B	T	B	T	B	T	B	T
Typographic			1/1	1/1	0/1	0/1	0/1	0/1
Missing tokens in the mention			0/4	4/4	1/5	4/5	1/20	14/20
Additional tokens in the mention	0/3	0/3	1/1	1/1	1/1	1/1	1/1	0/1
Synonyms					1/1	0/1	1/1	0/1
Acronyms							0/2	0/2
Lexical variation	0/1	0/1	5/5	5/5	5/5	5/5	16/17	15/17
Lexical variation + reordering			1/1	1/1	0/1	1/1		
Total	0/4	0/4	8/12	12/12	8/14	11/14	19/42	29/42

Table 3: Correct results obtained on (B)aseline and (T)rigrams based methods over each dataset (identical mentions ommited)

addressed in the next section.

Besides, all these collections usually contain KB resources that can be easily disambiguated, but in a real world scenario we can expect questions where there is not enough context to disambiguate the entity. A good example of this is when the name of the person that the question refers to cannot be disambiguated, i.e.: *What was the wealth of Aristoteles?*, *When did Armstrong die?*. This leads to a scenario where a QA system might ask for additional information about the mention before returning a possible resource. However, current datasets are not useful for such scenarios.

7 Release of datasets for EL in QA

One of the contributions we provide is the Entity Linking prepared version of each dataset, namely *LC-QuAD-EL* and *QALD-X-EL*. In the case of QALD based datasets from 1 to 4 we manually classified all the questions by error type, as well as annotate the correct mention. For LC-QuAD, given its size, we only annotated the mentions.

To construct each *EL* dataset, we have applied our method to automatically annotate all exact matches and manually correct the rest.

The contributed datasets follows two objectives: (1) Evaluate the proposed method itself and (2) Produce high quality and reliable datasets for evaluating this task. We have produced and openly released the following resources:

1. QALD datasets 1 through 4: We obtained some few hundred samples in total. The resources are named: QALD-X-EL, with X from 1 to 4.
2. LC-QuAD, that features 5K natural language questions accompanied with its corresponding SPARQL queries, was expanded to 6.6K samples, comprising 3.9K different resources. The final dataset is LC-QuAD-EL
3. Complex-EL4QA: A curated compilation of all the questions across all the analysed datasets that were not exact matches.

Table 4 shows the number of unique questions, unique entities covered as well as the total number of samples obtained for each Entity Linking dataset.

Dataset	U. Q.	U. E.	Total
QALD-1-EL	3	3	4
QALD-2-EL	11	11	12
QALD-3-EL	13	13	14
QALD-4-EL	38	40	45
LC-QuAD-EL	1204	997	1292
C-EL4QA	1307	1083	1469

Table 4: Number (U)nique questions, (U)nique entities covered and number of (T)otal.

8 Conclusions and Future Work

Entity Linking is still an open challenge in the NLP community and an important task in Question Answering. However, there is no specific collection for evaluating Entity Linking in QA.

In this paper, we introduce a benchmark to measure linkage accuracy between KB entities and mentions in a natural language question. In doing so, we classify common cases that add complexity to the task. Also, a

simple yet effective baseline approach is proposed and it can be used as the ground foundation of new, more complex systems. With this new benchmark we expect to help narrowing down the problem of EL intermediate step when answering questions.

Moreover, we deliver to the research community a new dataset that comprises all non-exact matched questions across all datasets that can be used to assess the effectiveness of the previous methods proposed over LC-QuAD, QALD and WebQuestions associated tasks. We also suggest to pay more attention to Entity Linking in QA collections.

We think the real world scenario where we can expect ambiguous questions deserves more attention. Also, we introduce the research question about how to evaluate the interaction with the user to capture the disambiguated version of the information needed.

Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación) LIHLITH project (PCIN-2017-085/AEI) in the framework of EU ERA-Net CHIST-ERA and RTI2018-096846-B-C21 (MCIU/AEI/FEDER,UE).

References

Artiles, J., A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. 2010. Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *CLEF 2010*.

Berant, J., A. Chou, R. Frostig, and P. Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP 2013*, pages 1533–1544.

Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Chen, L., J. Liang, C. Xie, and Y. Xiao. 2018. Short text entity linking with fine-grained topics. In *CIKM '18*, pages 457–466, New York, NY, USA. ACM.

Guo, S., M.-W. Chang, and E. Kıcıman. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of NAACL-HLT*, pages 1020–1030.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, and C. Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6.

Levenshtein, V. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10.

Mcnamee, P. and H. T Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Proceedings of the Second Text Analysis Conference*.

Rizzo, G., M. van Erp, J. Plu, and R. Troncy. 2016. Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. In *6th Workshop on Making Sense of Microposts*, pages 50–59.

Shekarpour, S., K. M. Endris, A. J. Kumar, D. Lukovnikov, K. Singh, H. Thakkar, and C. Lange. 2016. Question answering on linked data: Challenges and future directions. In *WWW 2016*, pages 693–698.

Trivedi, P., G. Maheshwari, M. Dubey, and J. Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *International Semantic Web Conference*, pages 210–218. Springer.

Unger, C., C. Forascu, V. Lopez, A.-C. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. 2014. Question Answering over Linked Data (QALD-4). sep.

A Appendix 1

Along the paper, we deliver all the necessary scripts to replicate our experiments. There is a README explaining each source code file. All the source code is released under the GPL-v3 license and can be found here.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 129-136 recibido 01-04-2019 revisado 08-05-2019 aceptado 15-05-2019

Landscaping Language Technologies using Topic Modeling and Graph Analysis: Overview of the Spanish Contribution

Caracterización del sector de Tecnologías del Lenguaje mediante modelado de tópicos y análisis de grafos: Visión general de la participación española

Doaa Samy[1,3], David Pérez-Fernández[1], Jerónimo Arenas-García[1,2]
[1]Secretaría de Estado de Avance Digital, Spain
[2]Universidad Carlos III de Madrid, Spain
[3]Instituto Ing. del Conocimiento (IIC), Madrid, Spain and Cairo University, Egypt
doaa.samy@iic.uam.es, dperezf@mineco.es, jeronimo.arenas@uc3m.es

Abstract: This paper aims at landscaping the Human Language Technologies (HLT) sector by applying topic modeling and graph analysis to study the scientific literature in ACL Anthology with special emphasis on the Spanish participation. The analysis takes into account the structured and unstructured data to offer an overview of the HLT landscape in Spain identifying main underlying themes and its evolution in the last years compared to the international HLT community. Results obtained are represented through an interactive visualization to allow the exploration of the HLT landscape in the time frame 1983-2018.
Keywords: Human Language Technologies, Topic Modeling, Latent Dirichlet Allocation (LDA), Louvain modularity algorithm, Natural Language Processing

Resumen: El presente trabajo aplica herramientas de modelado de tópicos y análisis de grafos para caracterizar el sector de Tecnologías del Lenguaje (TL) en España. Para ello, se estudian el repositorio de ACL Anthology. Este análisis tiene en cuenta los datos estructurados y no-estructurados en dichas fuentes con el fin de retratar el panorama actual en términos de temáticas subyacentes y su evolución en los últimos años en comparación con la comunidad internacional. Los resultados se presentan mediante una visualización interactiva que permite navegar en el espacio de TL en el intervalo temporal 1983-2018.
Palabras clave: Tecnologías del Lenguaje, Modelado de Tópicos, Latent Dirichlet Allocation (LDA), Algoritmo de modularidad de Louvain, Procesamiento del Lenguaje Natural

1 *Introduction*

The amount of data available from scientific production is increasing rapidly. Dynamic innovative tools are needed to analyze this data in its structured and unstructured formats to offer better insights on the situation and the progress in different fields. Bibliometrics and Scientometrics approaches have opted for statistical approaches and indicators based on structured data (Xu et al., 2017) (Serenko et al., 2010) (Langhe, 2016) (Clauset, Newman, and Moore, 2004). However, Natural Language Processing (NLP) and Artificial Intelligence can offer a complementary approach by discovering the underlying semantics in the unstructured text.

In this paper, we opt for a comprehensive hybrid approach which has been previously adopted in Corpus Viewer, a tool developed by the authors that analyzes the R&D&i space in general, including scientific publications, funded projects, patents, etc.

Building on the above previous experience, we address the HLT field as a rapidly growing and interdisciplinary domain which would rather benefit from this approach. Moreover, HLT has recently been in the centre of AI advances as an enabling technology

ISSN 1135-5948. DOI 10.26342/2019-63-14 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

in different fields: Human Computer Interactions, Internet of Things, Smart Solutions, etc. This is reflected in an increasing demand by the different sectors to include NLP components in their solutions. Also, it is reflected in several areas of applications and research lines.

A key starting point for a strategic plan to promote HLT in Spain and eventually in Hispanic countries, is: 1) to understand and characterize the sector and its dynamics in general and 2) to characterize the Spanish sector. In this line, the present study stems out from the following basic questions:

- What are the main thematic lines addressed by HLT and how these lines are reflected in the scientific literature?
- How these thematic lines have evolved along the years?
- What is the participation of Spanish key-players and their contribution within the general HLT landscape?

Answers to these questions would provide insights for the current strategic plan in Spain and would add value to the sector for a better self-positioning within the international landscape.

Although our approach is completely general and could be used with other datasets, to landscape the HLT sector we rely primarily on the ACL Anthology[1](Bird et al., 2008), a repository of computational linguistics and NLP papers, including open access to the full text in pdf format. In addition to this, we used Semantic Scholar[2] to enrich the information from ACL adding abstracts, entities and citations. The ACL Anthology is a valuable resource that has been used in a number of studies for example (Hall, Jurafsky, and Manning, 2008) applied topic modeling to study the history of ideas in the ACL Anthology. The ACL Searchbench was developed in 2011[3] (Schäfer et al., 2011) allowing the search by keywords, full text, affiliation, etc. However, our approach is completely different since it pretends to landscape the information space of the ACL Anthology semantically and dynamically in time. Others enriched the Anthology with semantic (Gábor et al., 2016) or co-reference annotation (Schäfer, Spurk, and Steffen, 2012). More studies included scientific term mining (Jin et al., 2013), studying gender aspects (Vogel and Jurafsky, 2012), etc.

The rest of the paper is divided into 3 sections. The methodology is described in Section 2 underlining the data sources and the experiments. Results and conclusions driven are discussed in Sections 3 and 4.

2 *Methodology*

Our methodology consists of the following stages, that will be described in the next subsections:

1. Identifying relevant datasets representing HLT scientific literature and extracting the metadata fields relevant to the study.
2. Selecting the data subset representing the Spanish participation.
3. Processing the unstructured text using basic tokenization and segmentation through an NLP Pipeline. The final output of this stage is the text lemmatized and a set of n-grams.
4. Topic modeling.
5. Construction of semantic graphs based on the interdistance between the topic vectors identified in the previous step.
6. Semantic community identification and characterization.
7. Visualization of the results using a graph visualization tool, such as Gephi (Bastian, Heymann, and Jacomy, 2009).

2.1 Enriching ACL Anthology metadata and identifying Spanish contributions

Two main datasets are used in this study: 1) ACL Anthology as a representative dataset of HLT scientific literature and 2) Semantic Scholar as an additional resource used for enriching the dataset with extra information which is not available in ACL Anthology. In Semantic Scholar we have access to the following metadata that is not available in ACL: abstract, citations and "entities". It is important to point out that the term "entities" applied by Semantic Scholar is not necessarily *named entities*, but rather candidates of

[1] https://www.aclweb.org/anthology/
[2] https://www.semanticscholar.org/
[3] http://aclasb.dfki.de/#

domain specific terms or n-grams in the abstract.

As of March, 2019, ACL Anthology hosts 48,286 entries. The whole website was downloaded using a web crawling tool, including paper, author, and event metadata. A total of 47,198 full papers in pdf format where also crawled from the web. In order to enrich the information with metadata available in Semantic Scholar, we carried out a matching procedure, so that for every paper in the ACL collection we searched for an appropriate match in Semantic Scholar. To avoid wrong assignations, the process was carried out ensuring that matched papers were published in the same year, and that at least one of the following conditions was satisfied:

- DOI match: matched papers should have identical DOIs.
- Title match: matched papers should have identical lowercase titles.
- URL match: the URL from which the ACL paper was retrieved is listed in the URL field of the matched paper from Semantic Scholar.

As a result of this process, we obtained a dataset of 42,396 papers. A simple inspection of the unmatched ACL papers revealed that most of them correspond to lists of authors, cover pages of conference proceedings, book reviews, etc.

In addition to the general ACL corpus, we created a subcorpus of ACL Anthology reflecting the Spanish HLT contribution in the field (to which we will refer in the following as ACL-SPA). Inclusion of papers in the ACL-SPA subset was based on the criteria that at least one of the authors is affiliated to a Spanish institution.

Author affiliation is not available in either ACL or Semantic Scholar. Though affiliations are available in ACL Searchbench, they can not be readily applied since institutions names are not normalized and it does not contain all papers currently available in ACL. Therefore, we used a simple approach for detecting author affiliation: using regular expressions for email detection. Based on detected emails, we filtered the following domains: '.es', '.cat', '.eus', and '.gal'. We selected also '.edu' domains of Spanish Universities. Finally, we also searched for appearances of the word 'Spain' in the first page of the paper followed by human check to make sure only papers with Spanish affiliations were selected. The ACL-SPA subset contains 1,408 papers, representing barely 3.32% of the total number of papers in ACL.

2.2 Unstructured Text Preprocessing

Available text for all the papers in the ACL Anthology was processed to obtain a valid document representation for topic modeling. For the study, we used the following unstructured text information:

- The title of the paper
- The list of entitities provided by Semantic Scholar
- The abstract, also available from Semantic Scholar
- The text of the full paper extracted from the pdf files using the *tika-python* library[4], which is based on the popular Apache Tika toolkit[5].

In order to build the vocabulary and lemmatize the text, we used the librAIry NLP toolkit (Badenes-Olmedo, Redondo-Garcia, and Corcho, 2017) which provides the following NLP tools: 1) Part-of-Speech Tagger (we keep nouns, adjectives, verbs, and adverbs), 2) Lemmatizer, 3) N-Grams Identifier and 4) Wikipedia Resource Finder. In addition to this, we removed stopwords using a generic list of stopwords, as well as a list of typos and common terms in the field identified by the authors that provide little semantic content in this context (e.g., 'document' , 'word'). Finally, we also removed from the vocabulary any word that appeared in less than 7 documents or in more than 40% of them.

After text preprocessing, we obtained a vocabulary of 16,075 entries, which increased to 97,053 when using the full papers. In the latter, many typos are still present due to common errors in the pdf extraction.

2.3 Topic Modeling using LDA

To analyze the ACL document collection and extract its underlying topics in an automatic way, we recur to Latent Dirichlet Allocation (LDA), a well-known tool for probabilistic

[4] `https://github.com/chrismattmann/tika-python`

[5] `https://tika.apache.org/`

topic analysis first proposed in (Blei, N, and Jordan, 2003). Although later refinements of LDA have appeared in the literature, the original algorithm suffices our needs in the paper, and has the advantage of availability of very efficient implementations. In this paper, we use the Mallet implementation (McCallum, 2002), which is based on Collapsed Gibbs Sampling and supports multithreading for very fast and scalable execution.

In LDA, a topic is characterized by a probability distribution over the complete vocabulary. Its generative model assumes also that each document is generated as a mixture of the different topics, where the sum of the topic proportions for each document is one. Therefore, to train the LDA model we provide as input the bag of word representations of all documents in the corpus and the number of topics (K), and the output we obtain consists of:

1. The length-K vectors representing the topic proportions for each document, $\boldsymbol{\theta}_d$, $d = 1, \ldots, \#\text{docs}$.
2. The vectors representing the vocabulary distribution for each topic, $\boldsymbol{\beta}_w$, $w = 1, \ldots, \#\text{vocabsize}$.
3. The relative importance of each topic.

Selection of the number of topics is always subject to a tradeoff between granularity of the topics (with a larger number of topics we observe very specific topics that do not appear for very few topics) and appearance of garbage topics (more numerous when using a large number of topics).

2.4 Semantic Graph Construction and Community Detection

Topic models are a very useful tool for finding out relevant topics inside a collection of text documents. However, there are some questions that remain unanswered with an LDA model. For instance, if we would like to know the number of papers about Machine Translation, the information on vectors $\boldsymbol{\theta}_d$ can only give an approximation, since many different papers will have a non-zero contribution to the topic. In other words, in principle all papers can belong to all topics but with a variable degree.

In this paper, we follow an alternative approach consisting of the following two steps: 1) building a semantic graph of papers, in which two papers are connected if they are semantically similar, and 2) finding groups of connected papers in the graph.

Unlike co-citation graphs, in our analysis we aim to calculate semantic graphs using the topic representation for each paper. The underlying hypothesis for prefering semantic graphs is that co-citation graphs do not contain links for all pairs of semantically related papers. Besides, some of the links in co-citation graphs may also be between papers from different topics. In short, we expect the semantic graph to be more complete and less noisy than the corresponding co-citation graph. An additional advantage is that link weights are real in contrast to co-citation graphs links.

In order to measure the semantic similarity between two papers, we compute the Jensen-Shannon divergence between their topic vectors,

$$\mathrm{JS}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = \frac{1}{2} \sum_i \frac{\theta_d^{(i)} \log 2\theta_{d'}^{(i)} + \theta_{d'}^{(i)} \log 2\theta_d^{(i)}}{\theta_d^{(i)} + \theta_{d'}^{(i)}}, \tag{1}$$

where we are using base 2 logs and $\theta_d^{(i)}$ represents the ith component of vector $\boldsymbol{\theta}_d$. Using base-2 logs, the divergence between two documents is between 0 (for $\boldsymbol{\theta}_d = \boldsymbol{\theta}_{d'}$) and 1 (for $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_{d'}$ with disjoint non-zero components).

Regarding the computation of the semantic graph, evaluating (1) for all pairs of documents can easily become unfeasible for very large datasets. A first possibility to facilitate the calculation is to use parallel computing. For further scalability we use a bound on the Jensen-Shannon based on Hellinger distance, namely

$$\mathrm{JS}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) \geq 1 - \sum_i \sqrt{\theta_d^{(i)} \theta_{d'}^{(i)}}. \tag{2}$$

Since (2) is more economic to evaluate than (1), we can use this bound to skip the calculation of Jensen-Shannon divergences above a predetermined threshold.

Once the semantic graph has been constructed, we still need to find papers that share a high density of connections among them. For that, we will use a well-known algorithm for community detection in graphs, the Louvain algorithm (Blondel et al., 2008), that pursues to maximize an objective criterion known as modularity, which measures

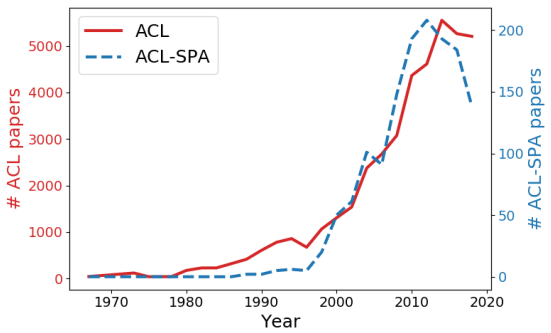


Figure 1: Papers per 2-year interval (production is larger in odd years given frequency of certain events). Spanish contribution is also represented in dashed blue (note the different scaling of the y-axis)

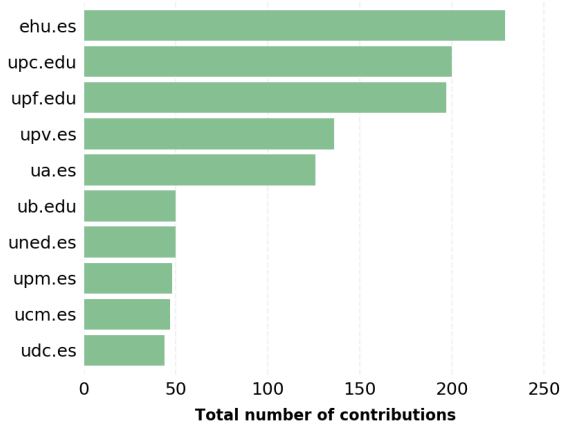


Figure 2: Number of contributions from most active Spanish institutions

precisely the density of links inside identified communities with respect to the density of links between nodes across different communities. In order to visualize the graphs and apply the Louvain algorithm, we have used Gephi.

3 *Experiments*

In this section we present the results of the analysis, and discuss the thematic composition and evolution of the HLT field according to the ACL Anthology papers. We depict also the role of the contribution from Spanish researchers.

The section starts with a description based on metadata features. We continue our discussion presenting the results from the topic model, and analyzing the relative importance of ACL-SPA papers in the identified topics. The last subsection presents the results from the graph analysis, describing the detected communities, and their time evolution using graph representations.

3.1 Overall perspective of the Spanish contribution

We start analyzing the volume of the scientific contributions in the field over time. Fig. 1 shows that scientific production has increased very significantly from year 2000, showing approximate exponential increase. Spanish contribution shows approximately the same trend, with presence of Spanish authors in approximately 3.5% of the papers used for the study.

Next, we analyze the scientific production of Spanish institutions. For this, we consider a paper is assigned to an institution if any of the authors is declaring such affiliation. Affiliation detection is based on the email addresses provided by the authors in the paper. We also applied some rules for equivalent domains (e.g., `ub.edu` is equivalent to `ub.cat`). Fig. 2 shows the number of contributions of the 10 most active Spanish institutions in the ACL Anthology. These institutions alone are responsible for approximately 75% of available contributions in the ACL-SPA subset. Most relevant private companies are Barcelona Media, Vicomtech and Telefonica I+D with 20, 17, and 16 identified papers, respectively.

3.2 Topic distribution of contributions

We have created an LDA topic model using $K = 25$, as an appropriate tradeoff between topic granularity and number of meaningless topics. All topics have been checked and annotated considering both the most significant words for each topic and the output of an automatic annotation algorithm. One topic was removed from the model because of a very poor representation in terms of most significant words (mostly typos). Table 1 describes the 24 restating topics. We can see that, in most cases, LDA found very relevant topics for the field, providing a good insight into the different subfields in HLT.

Figure 3 shows also topic size illustrating the relative contribution of Spanish institutions. Spanish sector has a relative larger size for topics 3, 4, 9, 12, and 16. Observing topic description in Table 1, many of these topics have to do with the creation of

tpc	Size	Description	tpc	Size	Description
0	0.075	supervised_learning, deep_learning,...	12	0.039	machine_translation,...
1	0.058	classification, supervised_learning,...	13	0.039	embedding, deep_learning,...
2	0.057	computational_linguistics,...	14	0.037	part-of-speech, multiword...
3	0.053	resource, annotation, corpus,...	15	0.037	natural_language_generation,...
4	0.050	resource, format, platform	16	0.035	lexical_resources, lexico-...
5	0.048	question_answering, summarization,...	17	0.035	sentiment_analysis, opinion_mining,...
6	0.045	computational_semantics, reasoning,...	18	0.035	speech_processing, dialog_system,...
7	0.044	grammar, parsing, treebank	19	0.032	information_extraction, BioNLP,...
8	0.042	parsing, dependency_grammar, treebank	20	0.030	speech_recognition, transcription,...
9	0.041	semantics, word_sense_disambiguation	21	0.029	parallel_corpora, alignment,...
10	0.040	tagger, part-of-speech,...	22	0.026	morphological_analysis,...
11	0.039	information_retrieval,...	23	0.025	computational_pragmatics, discourse,...

Table 1: Description of topic model with 24 topics

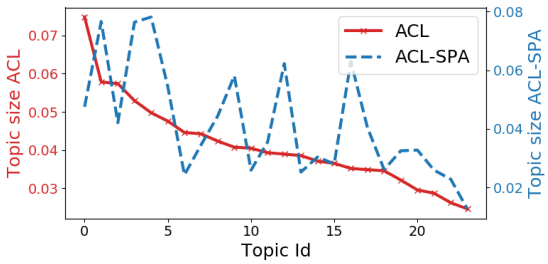


Figure 3: LDA topic size for the ACL corpus and estimated topic size for ACL-SPA papers

HLT resources (most likely for co-official languages in Spain). Topic 12 regarding Machine translation shows also a representation clearly above the average size of the Spanish sector as a whole.

3.3 Evolution of graph communities

Different experiments were conducted to select the best results for community detection and graph analysis. Community detection and graph analysis revealed to have better results when using the 40 topic-model based on the abstracts with a modularity of 1 (distinct values could be used for changing community size or radius) compared to those based on topic-models using the full paper.

Analyzing the data of the detected communities and the generated graph, 21 relevant communities were clearly identified reflecting the main areas of HLT. Each included an average of 1600 paper-entries. These communities were validated manually to assign a Community Description for each (see Table 2). It is interesting to point out that Evaluation, Validation and Quality Measures were clearly detected in a community. Parallel bilingual and multilingual resources were also detected within one community together with alignment, translation memories, etc. It is also interesting to observe a detected community of NLP toolkits and platforms or a community with clear presence of the Biomedical domain.

Figure 4 shows the evolution of some of the main areas of HLT such as Grammars, Statistical Machine Translation, Language Resources, Sentiment Analysis.

Semantics, Word Sense Disambiguation, Named-Entities are clearly represented in a highly populated community with over 3800 paper-entries. Evolution of this subgraph shows how Word Sense Disambiguation and Word2Vec started to populate more areas in this community in the last years (See Figure 5). The Spanish participation is represented in blue nodes, where it is clear that Spanish groups are more active in Semantics, WordNet and Ontologies. Moreover, the graph analysis allows zooming up into certain communities. Figure 6 illustrates a zoom of the communities related to Statistical Machine Translation. Evaluation of MT and Proceedings of the Workshop on MT (WMT) are detected as sub communities.

4 *Conclusions*

In this work, we have landscaped the HLT Sector through scientific literature included in the ACL Anthology, considered as one of the most important resources for HLT.

The methodology for the analysis heavily relies on Topic Modeling and Graph Analysis tools. Our approach allows to automatically detect the most significant thematics, as well as learning the semantic relation among pa-

#Papers	Community Description
3,858	Semantics, WSD, Anaphor, Semantic_Role_Labeling, Metaphor, Coreference, Syntax
3,109	Information_Extraction, Named_Entities, Temporal_Expressions, Factoids, Sentiment_Analysis
3,027	Lexicology, Semantics, Terminology, Dictionaries, WSD, Word_Vectors, Word_Embeddings
2,813	Language_Understanding, Language_Generation, Machine_Translation, Semantics, WSD, MWEs
2,074	Grammars, HSPG, Finite_State_Transducer, Automata, Formalism, Syntax
2,068	Parsing
2,055	Machine_Translation
1,833	Language_Modelling, Statistical_Language_Processing, HMM, Finite_State, Decision_Trees, SVM, CRF, Unsupervised_Learning, Deep_Learning, Neural_Networks, Chinese, Japanese
1,759	Language_Resources, Corpora, Multilingual_Resources, Minority_Languages
1,739	Speech_Recognition, Phonology, Acoustics, Spoken_Language, Dialog
1,541	Question_Answering, Information_Extraction, Information_Retrieval, Clustering, Cross_Language
1,526	Dialog, Speech_Processing, Interaction, NL_Understanding, Assitants, Agents
1,486	Domain_Specific, News, Email, Author_Profiling, Financial_Text, Political_Text, Narrative, Sports, Social_Media, Chat, Stance, Emotions, Opinions, Recommendation_System
1,419	Parallel_Corpora, Bilingual_Resources, Alignment, Translation_Memories, Multilingual_Resources
1,310	Discourse, Speech, Dialog, Pragmatics, Natural_Language_Generation, Anaphor
1,271	Evaluation, Validation, Evaluation_Metrics, Quality_Measures, Error_Detection
1,264	Ontology, Terminology, Information_Extraction, Taxonomy, Medical Domain, BioNLP
1,149	Summarisation, Language_Generation, Information_Extraction, Indexing, Document_Clustering
1,037	Annotation_Tools, Toolkits, Architecture, Interfaces, Text_to_Speech, Interaction
737	Morphology, Morphological_Analysis, Morphotactics, Inflection, Agglutinative, Stemming
728	Sentiment_Analysis, Subjectivity, Opinion_Mining, Polarity, Stance, Social_Media, Twitter

Table 2: Detected communities.

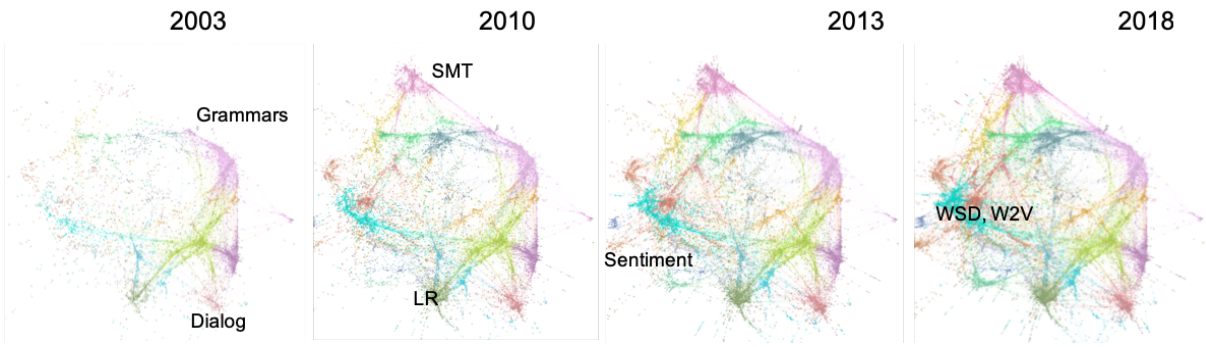


Figure 4: Evolution of main detected HLT communities (1983-2018)

pers. Semantic graphs are calculated, and communities of semantically similar papers have been detected and characterized.

Spanish contribution in the field represents 3.32% of the international scientific production represented by the ACL Anthology. According to our analysis, this participation is more significant in the areas of Lexicology & Semantics, Machine Translation, Parsing, and Speech Technologies.

These observations give insights for strategic plans to promote HLT taking into consideration the rapidly growing interest in the last years for its role in the recent advances in cognitive technologies and Artificial Intelligence. Moreover, Spanish HLT key players could contribute with a significant role in the Hispanic Language Industry.

In this study, we limited the experiments to the scientific publications in the ACL Anthology. Further studies will include more resources, especially the SEPLN journal. Also, more experiments will be conducted to landscape not only the scientific literature, but also funded R&D projects.

Acknowledgments

This work has been carried out in the framework of the Spanish State Plan for Natural Language Technologies. The work of J. Arenas-García has also been partly funded by MINECO projects TEC2014-52289-R and TEC2017-83838-R.

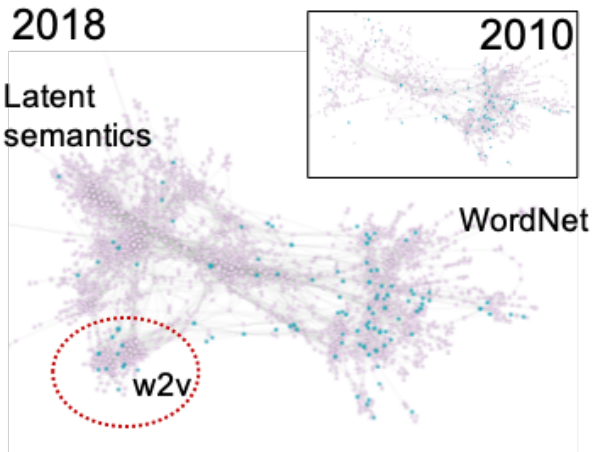


Figure 5: Latent semantic, wordEmbeddings and WordNet related papers. Contributed papers with Spanish affiliations are represented in blue

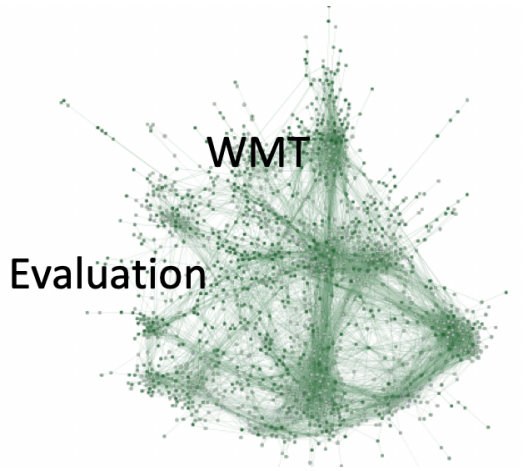


Figure 6: Graph Representation of SMT knowledge area

References

Badenes-Olmedo, C., J. L. Redondo-Garcia, and O. Corcho. 2017. Distributing text mining tasks with librAIry. In *Proc. 2017 ACM Symposium on Document Engineering*, DocEng '17, pages 63–66. ACM.

Bastian, M., S. Heymann, and M. Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks.

Bird, S., R. Dale, B. J. Dorr, B. R. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. LREC*.

Blei, D. M., A. Y. N, and M. I. Jordan. 2003. Latent dirichlet allocation. In *Proc. NIPS*.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10).

Clauset, A., M. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*.

Gábor, K., H. Zargayouna, D. Buscaldi, I. Tellier, and T. Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *Proc. LREC*.

Hall, D. L. W., D. Jurafsky, and C. D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP*.

Jin, Y., M.-Y. Kan, J.-P. Ng, and X. He. 2013. Mining scientific terms and their definitions: A study of the ACL anthology. In *Proc. 2013 Conf. Empirical Methods in Natural Language Processing*.

Langhe, R. D. 2016. Towards the discovery of scientific revolutions in scientometric data. *Scientometrics*, 110:505–519.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit.

Schäfer, U., B. Kiefer, C. Spurk, J. Steffen, and R. Wang. 2011. The ACL Anthology Searchbench. In *Proc. ACL*.

Schäfer, U., C. Spurk, and J. Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In *Proc. COLING*.

Serenko, A., N. Bontis, L. D. Booker, K. W. Sadeddin, and T. Hardie. 2010. A scientometric analysis of knowledge management and intellectual capital academic literature (1994-2008). *J. Knowledge Management*, 14:3–23.

Vogel, A. and D. Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proc. ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*.

Xu, H., Z.-H. Yue, C. C. Wang, K. Dong, H. Pang, and Z. Han. 2017. Multi-source data fusion study in scientometrics. *Scientometrics*, 111:773–792.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 139-142 recibido 01-04-2019 revisado 04-04-2019 aceptado 17-05-2019

MENHIR: Mental health monitoring through interactive conversations

MENHIR: Monitorización de la salud mental mediante conversaciones interactivas

Zoraida Callejas[1], **Kawtar Benghazi**[1], **Manuel Noguera**[1], **María Inés Torres**[2], **Raquel Justo**[2]
[1]Universidad de Granada (UGR)
[2]Universidad del País Vasco (UPV/EHU)
{zoraida, benghazi, mnoguera}@ugr.es, {manes.torres, raquel.justo}@ehu.eus

Abstract: Mental health is an essential component of health; however depression and anxiety are common disorders across the European Union. Many people may be described as "living with" a mental illness, and managing their own symptoms. However, they are often unsure of the thresholds for treatment, how to control their mental health, what are the best coping strategies or which resources are available to them. Conversational systems may facilitate watchful waiting and symptom monitoring, by initiating contact and symptom checking at various times of the day and night. The MENHIR project aims to research and develop conversational technologies to promote mental health and assist people with mental ill health (depression and anxiety) to manage their conditions. The participants will exchange and share their expertise toward this common objective, building a network for exchanging ideas and knowledge between Spain, Germany, Italy and United Kingdom, including academic, non-profit organisations and companies.
Keywords: Mental Health, dialogue system, chatbot, anxiety, depression, wellbeing

Resumen: La salud mental es un componente esencial de la salud, sin embargo la depresión y la ansiedad son desórdenes muy comunes en la Unión Europea. Muchas personas conviven con un problema de salud mental y gestionan sus propios síntomas. Sin embargo, suelen estar inseguras acerca de cuál es el umbral para iniciar la búsqueda de tratamiento, cómo controlar su salud mental, cuáles son las mejores estrategias de afrontamiento o qué recursos están a su disposición. Los sistemas conversacionales pueden facilitar la monitorización de síntomas o la "espera vigilante", iniciando el contacto y comprobando los síntomas en diversos momentos del día. El proyecto MENHIR propone investigar y desarrollar tecnologías conversacionales para promover la salud mental y asistir a personas con problemas de depresión o ansiedad para la gestión de sus síntomas. Los participantes intercambiarán su experiencia para alcanzar este objetivo común, construyendo una red para el intercambio de ideas y conocimiento entre España, Alemania, Italia y Reino Unido, incluyendo universidades, organizaciones sin ánimo de lucro y empresas.
Palabras clave: Salud mental, sistema de diálogo, chatbot, ansiedad, depresión, bienestar

1 *Project funding and duration*

The MENHIR project (`http://menhir-project.eu`) has been funded by the European Commission H2020-MSCA-RISE-2018 grant number 823907, under the funding scheme Marie Skłodowska-Curie Research and Innovation Staff Exchange.

MENHIR started on 1st February 2019 and will have a duration of 4 years, finishing on 31st January 2023.

2 *Consortium*

The consortium brings together 8 partners including universities, a non-profit organisation and companies:

ISSN 1135-5948. DOI 10.26342/2019-63-15
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

- Universidad de Granada, Spain (coordinator).
- Universidad del País Vasco, Spain.
- Action Mental Health, United Kingdom.
- University of Ulster, United Kingdom.
- Intelligent Voice Limited, United Kingdom.
- Ulm University, Germany.
- Università degli Studi della Campania Luigi Vanvitelli, Italy.
- GLOBIT - Globale Informationstechnik GmbH, Germany.

3 Description

The World Health Organisation (World Health Organization, 2018) defines mental health as an essential component of health that should not be considered only as the absence of mental disorders or disabilities, but as an state of mental well-being. This implies that mental health care should not be understood only from the point of view of the treatment of symptoms, but it should have an important focus on its understanding, promotion and protection.

MENHIR is aligned with this idea and can be contextualised in the EU policies designed for the promotion of mental health and well-being. The Joint Action (on) Mental health and Well-being launched in 2013 resulted in the European Framework for Action on Mental Health and Well-being (European Commission, 2018), which supports EU countries to review their policies and share experiences in improving policy efficiency and effectiveness. MENHIR aligns with this framework and its recommendations, as the conversational system developed is designed considering the relevance of sustained mental health care and prevention by means of continuous user-system interaction, reduces mental health inequalities as it provides easy and universal access to mental health monitoring, and facilitates gaining very valuable feedback on the benefits of mood and anxiety monitoring that can help policymakers promote mental health.

Treatments such as cognitive therapy and medication are effective in reducing the symptoms of anxiety and depression and many people do experience a complete recovery following an episode of mental illness. However, for others, depression and anxiety disorders follow a different trajectory and there may be recurrences of the symptoms or episodes with varying levels of severity. Many people may be described as "living with" a mental illness, and managing their own symptoms using coping strategies and using their own support networks together with the formal support services and therapies such as counselling and medication.

MENHIR will pursue objective symptom monitoring and relapse prevention for people with mental illness. Symptom monitoring, whereby the patient assesses and records their moods, behaviour, and symptoms over time, can provide clinicians with additional, real time, more ecologically valid information to inform treatment strategies and empower the client to manage their own mental health symptoms.

The capture of real time information facilitates rapid intervention and appropriate feedback such as prompts, supportive messages and symptom summaries. The system will therefore promote user empowerment and collaborative care. The limitations of previous technology-based monitoring approaches are the high dropout rates and poor adherence in particular user groups. This is due in part to a lack of interactive features and appropriate prompts, difficulties for people with low levels of technological literacy and failure to provide human support (Walsh, Golden, and Priebe, 2016). The MENHIR chatbot aims to address these difficulties by using more realistic, natural and smart interactions. The project will also generate data on how people with anxiety and depression interact with and respond to symptom monitoring chatbots.

4 Aims and specific objectives

The aim of the project is to develop and test the efficacy and user experience of the MENHIR chatbot for the self-management of mental illness.

The objectives are as follows:

1. Co-create the MENHIR integrated chatbot/programme.
2. Monitor the use of the chatbot to develop an understanding of how users engage with it and how that impacts on

their behaviour and well-being.

3. Test the effect of the chatbot on the user's mental health over time.

The first goal of the project is to develop a smart mental health monitoring and communication chatbot for people who are living with a mental illness. The system will go some way to imitate a "virtual therapist" or "concerned friend" and provide a natural conversational experience that also tracks the user's moods, behaviours, social interactions and mental health symptoms. It can provide personalised responses based on the user's profile, recent mood trends and provide the appropriate coping strategies whilst also signposting to online digital resources, social communities and local geolocated services that are relevant at the point of care and need. The chatbot knowledge base will be curated from therapies and coping strategies identified in the scientific literature and techniques recommended by the psychology experts in this project.

The chatbot empowers the user to manage their own mental health by facilitating mood tracking over time and by linking symptom level data. This is akin to the quantified-self movement except the application is mental health as opposed to physical health. It is important to visualise mood trends to instil user reflection allowing for knowledge-driven motivational efforts. Trends will be depicted visually to promote reflection and understanding of the user's unique patterns of mood behaviour and symptom patterns, diurnal trends and links with daily activities. The chatbot will have a therapist personality with humor, empathy and rapport and conversational variability to promote engagement. Variability in the conversations will be a key design feature to imitate human-to-human conversations and to avoid tedious predictable dialogues.

Following, the data collected will be subjected to analysis to generate an understanding of the extent to which user engagement impacts on subsequent moods, behaviours and activity. Information on mental health symptoms will be collected regularly to establish the effect of engagement with the programme over time on the person's mental health generally. These processes will be performed with the users consent and controlled by the Ethical Board.

5 *Scientific relevance*

As shown in Figure 1, the chatbot incorporates a wide range of scientific and technical challenges including natural language and speech processing, knowledge management, emotion/sentiment analysis, dialogue management, user modeling, user experience design, conversation modeling and management, communicative style definition, system personality, and establishing trust to maximise user adoption. To a lesser extent, the project will explore the potential for the augmentation of the chatbot with data from wearable devices.

Voice Assistive Computer Interfaces can provide automated on-demand health assistance reducing costs. However, speech is intrinsically complex, and emotional speech is even more requiring the need for an holistic approach that accounts for several factors including personality traits, social and contextual information. Such interfaces can be exploited as automatic diagnostic tools for the diagnosis of different degrees of depressive states and in general for detecting from interactional exchanges reliable behavioural and contextual information. Our goal is to considerably extend the results of earlier research expanding the set of identified depressive speech acoustic features and automate their detection so that depressed and anxiety speech can be accurately distinguished from healthy speech.

A recent well conducted meta analysis reported that users valued the convenience of mood tracking, and the ecological validity of the data was of value in clinical sessions. However there was frustration that current methods did not always capture what patients wanted to report or missed the essential qualities of conversations. Automated feedback messages were sometimes well received or drew mixed reactions, with some users finding them patronising or inappropriate, repetitive and impersonal (Walsh, Golden, and Priebe, 2016).

MENHIR will provide novel approaches focused on making the interaction personalised to the users in several ways. The acoustic analysis of the user input and the corresponding system responses must be adapted to the specific user groups. A big challenge to be addressed in the system is that prospective users are very heterogeneous in terms of age, gender, mental health condition (in

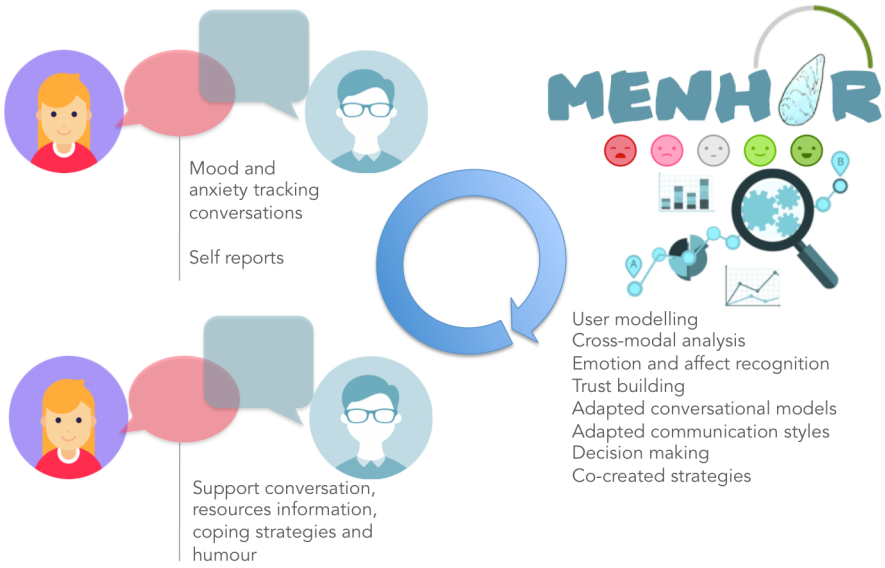


Figure 1: Summary of MENHIR's conversational loop

our case depression and/or anxiety), progression (some users may have more problems than others in reaching a healthy status) and responses to the system's strategies (different users may react differently to the same system advice). The user model generated within MENHIR will advance current state-of-the-art models which focus on specific user groups and represent the user by means of a series of pieces of information and preferences that do not vary over time.

From the perspective of conversation management, the system interactions must be tailored to knowledge that it has about the user at that specific moment of the interaction, during the previous turns of that session, and for the whole history of user-system interactions. During each interaction, the system must have a calculation on how far the user is from the objective, what his/her mood is, emotional state and anxiety level and how they have progressed over time. It will have to consider also the history of previous interactions to consider information about which strategies have been more effective in a similar context for that particular user or users in a similar condition.

In addition, as shown in Figure 1, MENHIR involves several types of conversations: tracking and self report and support. On the one hand, the system will try to track the mood and anxiety level of the user for which it would have to foster trust and rapport. These conversations can be more system-directed, e.g. answering a questionnaire that measures the level of anxiety, and sometimes will be more user-directed, for example, when the user explains how they felt during the previous days. On the other hand, the system must also give support to the user, in this case the conversation must be more persuasive as it will help the user cope with their current mental health situation. Developing conversation models and management strategies suited to these disparate conversation structures and being able to figure out when the system must lead the conversation or let the user talk to favour self-disclosure will be a very important scientific result of the project.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907.

References

European Commission. 2018. EU-Compass for Action on Mental Health and Well-being.

Walsh, S., E. Golden, and S. Priebe. 2016. Systematic review of patients' participation in and experiences of technology-based monitoring of mental health symptoms in the community. *BJM Open*, 6(6).

World Health Organization. 2018. Mental health: Strengthening our response.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 143-146 recibido 03-04-2019 revisado 08-05-2019 aceptado 14-05-2019

Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad

Extraction of Geospatial References from Free Text Based on Natural Language Processing and Accessibility Theory

Alejandro Molina-Villegas[1], **Oscar S. Siordia**[1],
Edwin Aldana-Bobadilla[2], **César Aguilar**[3], **Olga Acosta**[3]
[1]CONACyT – Centro de Investigación en Ciencias de Información Geoespacial, México
[2]CONACyT – Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional, México
[3]Pontificia Universidad Católica de Chile, Chile
amolina@centrogeo.edu.mx

Resumen: La extracción automática de información geoespacial en tiempo real, a partir de discurso libre, tendrá un enorme impacto en tecnologías disruptivas, tales como los asistentes inteligentes y los motores de búsqueda. Generar modelos capaces de decodificar el discurso para luego transformarlo en datos estructurados aborda la solución de problemas como: la identificación automática de frases que se refieran a alguna entidad geoespacial; el manejo de sinónimos referidos a una misma locación (ambiguedad); la caracterización taxonómica de expresiones locativas; así como la automatización de procesos complejos de interpretación del lenguaje para la determinación de coordenadas geográficas en mapas y bases de datos. El proyecto presentado tiene por objetivo la automatización de procesos de georeferencia de documentos digitales para extraer conocimiento. La propuesta metodológica incluye compilación de un corpus, caracterízación lingüística y desarrollo de algoritmos.
Palabras clave: Geoparsing, Geocodificación, Reconocimiento de Entidades Nombradas

Abstract: The automatic extraction of geospatial information in real time, from free speech, will an important impact on disruptive technologies, such as AI assistants and search engines. Models capable of decoding discourse and then transforming it into structured data addresses the solution of challenging problems such as: the automatic identification of phrases refering geospatial entities; the handling of synonyms referring to the same location (ambiguity); the taxonomic characterization of locative expressions; as well as the automation of complex processes of natural language interpretation to determine of geographical coordinates, maps and databases. The presented project aims to automate georeferencing processes of digital documents to extract georeferenced knowledge. The methodological proposal includes compilation of a corpus, linguistic characterization and algorithms.
Keywords: Geoparsing, Geocoding, Named Entity Recognition

1 Introducción

En los últimos años, el estudio de la expresión de nociones espaciales por medio del lenguaje natural, ha cobrado una gran relevancia, particularmente para la inteligencia artificial. Un buen ejemplo de esto son asistentes como Siri, Cortana o Alexa, por mencionar los más conocidos en el mercado. Uno de los motivos que subyace en tal relevancia —aunque no es el único— es la masificación de datos georreferenciados, implícita o explícitamente.

Este estudio propone la selección, recolección y procesamiento de expresiones en la variante de español mexicano (ya sean palabras o frases), que codifiquen algún concepto georreferencial asociable a entidades, escenarios o eventos locativos. En concreto, nos interesa

ISSN 1135-5948. DOI 10.26342/2019-63-16 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

resolver casos como:

1. *San Andrés Cholula es un municipio próximo a la Zona Metropolitana de Puebla.*
2. *El choque ocurrió esta mañana en el kilómetro 5 de la carretera México-Puebla*
3. *Los asaltantes irrumpieron en la sucursal que se encuentra en Puebla esquina con Monterrey, en la Colonia Roma*

En los tres casos, la entidad nominal *Puebla* tiene un claro atributo georreferencial, pero de diferente índole: mientras que (1) se refiere a un núcleo urbano, en (2) se alude a una vía que recorre dos provincias diferentes, mientras que (3) se refiere a dos calles situadas en un barrio de la Ciudad de México. Si bien es cierto que la ambigüedad que se genera para ubicar tal entidad en un mapa puede ser resuelta, también es verdad que resulta necesario delimitar el alcance de los atributos locativos asociados a dicha entidad, de tal suerte que puedan resolverse preguntas como: *¿cuál es la salida del metro más cercana a Puebla?*, en donde un asistente inteligente debería inferir que se trata de una calle, y no de una zona urbana, de la capital de una provincia, o de la provincia misma. En ese sentido, este proyecto aborda un problema *AI-Hard* que entrelaza la Inteligencia Artificial, la Lingüística y la Geomática, ya que en él se vislumbra la creación de métodos capaces de transformar voz o texto en identificadores geográficos inequívocos tales como latitud y longitud. Así, nuestra propuesta es multidisciplinaria, pues no solo se trata de caracterizar el fenómeno lingüístico de la georreferenciación discursiva en español, sino que también integrarlo en algoritmos y estructuras de datos de cara a la automatización completa de procesos sofisticados de georreferenciación por medio de la voz o el análisis de documentos de texto (*Geoparsing*). Con estas tecnologías lograríamos explotar, eficientemente, considerables cantidades de documentos existentes para extraer conocimiento georreferenciado.

2 Teoría de la accesibilidad

Un enfoque lingüístico pertinente para identificar entidades nominales con información georreferencial es el que plantea la *Teoría de la accesibilidad*, desarrollada principalmente por (2014), así como Gernsbacher y Givón (1995). Esta teoría explica cómo las unidades nominales y pronominales ofrecen un vínculo directo o indirecto a sus referentes, dependiendo de la cantidad de información que contengan. Tal cantidad de información, se sitúa en los niveles semántico y pragmático y permite que tales referentes sean reconocidos como elementos nuevos en el discurso (p. e., frases nominales largas como: *los vecinos de la nueva casa de enfrente*), o como elementos ya conocidos (p. e., el pronombre personal *ellos*, estableciendo una relación anafórica con la frase anterior). El tomar en cuenta la cantidad de información que contengan unidades nominales de índole locativa será útil para identificar aquellas que tengan un peso referencial relevante (p. e.: *la carretera federal México-Puebla*, con miras a contrastarlas con otras que puedan ser vistas o bien como segmentos nominales con valor anafórico situadas en un mismo contexto discursivo (*la México-Puebla*), o como unidades pronominales que requieren mayor información referencial para ser desambiguadas (p. e.: *por ahí se llega rápido a Puebla*).

3 Aportaciones del proyecto

Uno de los aportes relevantes de este proyecto será la caracterización lingüística de entidades georreferenciables. Para lo cual, nuestro enfoque metodológico considera la compilación del Corpus de Entidades Georreferenciadas de México (CEGEOMEX), el primero de su clase, el cual tendrá un anotado lingüístico manual, lo que facilitará el desarrollo de algoritmos de aprendizaje de máquina, así como la generación de meta-información discursiva. Cabe destacar que la anotación manual ya ha sido ampliamente utilizada en iniciativas internacionales tales como la CoNLL (*Conference on Natural Language Learning*) y que para este proyecto, resulta de especial interés la edición CoNLL 2002 (Sang y De Meulder, 2003) en donde por primera vez se consideró al español como una de las lenguas a procesar. Sin embargo, dado que en CoNLL 2002, los datos fueron recopilados por la Universidad Politécnica de Catalunya y la Universidad Autónoma de Barcelona, la anotación se focalizó en documentos de España, dejando de lado cualquier otra variante dialectal, entre ellas la mexicana. El rezago de recursos similares en México pone de manifiesto la necesidad de contar con un corpus de ca-

lidad con estas características. CEGEOMEX será un corpus de lengua general que se compondrá de documentos periodísticos, así como de segmentos de entrevistas y diálogos extraídos de la radio, con el propósito de conformar una colección balanceada de muestras escritas y orales siguiendo los criterios desarrollados por McEnery (2001) y Gries (2006) para observar variaciones en la expresión de entidades georreferenciadas en español mexicano. CEGEOMEX permitirá abordar la solución de problemas como: a) la identificación automática de frases u oraciones que se refieran a alguna entidad o evento de tipo espacial o georreferenciado implícitamente; b) el manejo de sinónimos referidos a una misma locación (Ciudad de México/Capital de la República/Distrito Federal/CDMX/DF); c) la propuesta de una taxonomía que ayude a clasificar expresiones locativas y entidades nombradas y d) la automatización de procesos complejos de interpretación del lenguaje para la determinación de coordenadas geográficas en mapas y bases de datos.

4 Objetivos

1. Establecer una caracterización lingüística de las entidades georreferenciadas mediante su definición formal, considerando tanto sus atributos lingüísticos, su estructuración en patrones, así como los parámetros contextuales que den indicios de su manifestación en un texto, sea oral o escrito.
2. Crear el primer Corpus de Entidades Georreferenciadas de México.
3. Innovar en la generación de algoritmos que combinando atributos lingüísticos y modelos computacionales detecten entidades georreferenciables para ser visualizadas en cartografía digital o en imágenes satelitales.
4. Consolidar un grupo de investigación multidisciplinario e internacional para desarrollar proyectos relacionados con nuevas áreas de investigación que unifiquen computación, lingüística y geomática.
5. Formar recursos humanos especializados en investigación de frontera.

5 Metas

1. Crear un repositorio de archivos para almacenar documentos periodísticos de México que incluya, al menos cinco de los periódicos principales de cobertura nacional: *El Universal*, *La Jornada*, *El Financiero*, *El Sol de México*, *La Razón*, *Uno Más Uno* y *Reforma*; así como seis estaciones de radio: *MVS noticias*, *Radio Fórmula*, *Radio IPN*, *Radio UNAM*, *Radio Ibero*, *Red FM* y Stereo Cien.
2. Implementar el primer Corpus de Entidades Georreferenciadas de México (CEGEOMEX) que servirá de base para futuros proyectos.
3. Diseñar un sistema para detectar entidades Georreferenciadas en discurso libre (oral y escrito).

6 Metodología

La metodología propuesta cubre 3 etapas, las cuales son descritas a continuación.

6.1 Etapa I

En aras de compilar un volumen masivo de documentos para caracterizar el fenómeno de estudio, en sus variantes escrita y oral, se desarrollarán programas informáticos para automatizar procesos de descarga, almacenamiento, indexado, transcripción y organización de noticias en los medios de comunicación antes mencionados, empleando técnicas de *Web Crawling*. Dado que se procesarán documentos de texto y audio, será necesario adquirir y configurar equipo especializado y software para ejecutar tal procesamiento, de tal suerte que se evaluarán e incluirán modelos de reconocimiento de voz para la transcripción de noticias de radio en español mexicano. El equipo encargado del análisis lingüístico estará integrado por investigadores de la Pontificia Universidad Católica de Chile, quienes han abordado el análisis de relaciones espaciales reconocidas entre términos médicos en español, tomando en cuenta un enfoque cognitivo (ver Acosta y Aguilar (2015)). Los resultados de su análisis refuerzan la hipótesis de que la concepción espacial se extiende a muchos dominios abstractos, tales como tiempo, estado, posesión, corporeidad u organización social, entre los más relevantes. La tarea de estos investigadores será aportar al proyecto una propuesta de taxonomía y especificación de rasgos espaciales para el diseño del etiquetado de entidades georreferenciadas. Estos trabajos ayudarán a

establecer el metalenguaje (etiquetas y atributos) y las consideraciones para el anotado manual del corpus.

6.2 Etapa II

En esta etapa se creará el corpus CEGEOMEX, el cual se etiquetará manualmente. Para apoyar esta tarea, se desarrollará una plataforma de anotación en el Centro de Investigación en Ciencias de Información Geoespacial (México). Dicha herramienta brindará la especificación de las etiquetas y atributos obtenidos de la etapa anterior. La herramienta dará acceso remoto y multiusuario de manera que en seis meses se llegue a la meta de, al menos, 12000 colocaciones anotadas, divididas en 6000 instancias orales y 6000 instancias escritas. Cabe mencionar que ningún corpus anotado cuenta actualmentelas características específicas de esta investigación.

6.3 Etapa III

Finalmente, en esta etapa se desarrollarán los algoritmos y un software que concentrará los resultados de la investigación (Figura 1). Aprovechando los resultados obtenidos de las exploraciones que se hagan al CEGEOMEX, será posible experimentar con algoritmos híbridos que utilizarán tanto atributos simbólicos (tales como posición de entidades y categorías gramaticales), así como variables abstractas generadas mediante técnicas de aprendizaje de máquina y reconocimiento de patrones. El hecho de generar variables abstractas será útil para vincular y representar en forma de vectores los resultados obtenidos (p. e., candidatos a entidades georreferenciales codificadas en nombres o en frases nominales). Dichos vectores, además de ofrecer una descripción numérica sobre el comportamiento de los resultados obtenidos, serán valiosos para categorizar similitudes semánticas identificables entre tales unidades lingüísticas. Cabe señalar aquí que estudios previos se ha demostrado la eficiencia de combinar patrones lingüísticos con algoritmos de aprendizaje de máquina (Sierra et al., 2009).

Agradecimientos

Proyecto FORDECyT 296737 (Consorcio en Inteligencia Artificial) y a la Red Temática en Tecnologías del Lenguaje por el financiamiento parcial de esta investigación. A la Mtra. en Literatura Mariana Tello-Signoret por la revisión de datos.

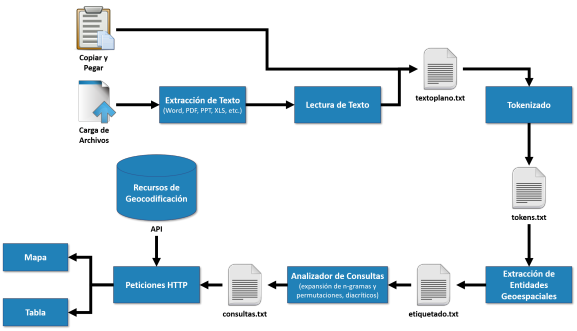


Figura 1: Proceso de extracción de referencias geoespaciales en discurso libre. Los documentos son preprocesados y tokenizados para que un módulo de Reconocimiento de Entidades Nombradas detecte nombres de lugares que serán georreferenciados mediante recursos externos para obtener sus coordenadas

Bibliografía

Acosta, O. y C. A. Aguilar. 2015. Extracting concrete entities through spatial relations. En *Proceedings of the 3rd International Workshop on Artificial Intelligence and Cognition, Turin, Italy, September 28-29, 2015.*, páginas 133–145.

Ariel, M. 2014. *Accessing noun-phrase antecedents*. Routledge.

Gernsbacher, M. A. y T. Givón. 1995. *Coherence in spontaneous text*, volumen 31. John Benjamins Publishing.

Gries, S. T. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2):109–151.

McEnery, T. 2001. Corpus linguistics/tony mcenery, andrew wilson.

Sang, E. F. y F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sierra, G.-E., R. Alarcon, A. Molina-Villegas, y E. Aldana. 2009. Web explotation for definition extraction. En *IEEE Latin American Web Congress, doi 10.1109/LA-WEB.2009.36*, páginas 217–223.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 147-150 recibido 31-03-2019 revisado 16-04-2019 aceptado 14-05-2019

LIHLITH: Learning to Interact with Humans by Lifelong Interaction with Humans

LIHLITH: Aprendiendo a interactuar con las personas mediante la interacción continuada con personas

Eneko Agirre[1], **Arantxa Otegi**[1], **Camille Pradel**[2], **Sophie Rosset**[3], **Anselmo Peñas**[4] and **Mark Cieliebak**[5]
[1]University of the Basque Country UPV/EHU, [2]Synapse Développement, [3]LIMSI, [4]Universidad Nacional de Educación a Distancia, [5]Zurich University of Applied Sciences
e.agirre@ehu.eus

Abstract: The LIHLITH project will research, innovate and validate a new lifelong learning framework for the interaction of humans and machines on specific domains with the aim of improving the quality of existing dialogue systems and lowering the cost of deployment in new domains. LILITH will develop dialogue systems that learn autonomously from their interactions with the users, and retain this new knowledge in order to answer more accurately in future interactions. The key insight is that the dialogue systems will be designed to get feedback from the user. Based on this feedback, the system will keep improving after deployment all modules down in the pipeline. LIHLITH project will also develop and deliver evaluation protocols and benchmarks to allow public comparison and reproducibility.
Keywords: Dialogue systems, lifelong learning, question answering, knowledge induction

Resumen: El proyecto LIHLITH investigará, innovará y validará un nuevo marco de aprendizaje permanente aprendizaje continuo para la interacción entre humanos y máquinas en dominios específicos con el objetivo de mejorar la calidad de los sistemas de diálogo existentes y reducir el coste de implementación en nuevos dominios. LILITH desarrollará sistemas de diálogo que aprenden de forma autónoma de sus interacciones con los usuarios, y retendrá este nuevo conocimiento para responder de manera más precisa en futuras interacciones. La idea clave es que los sistemas de diálogo se diseñarán para obtener *feedback* del usuario. Basándose en esta información, el sistema seguirá mejorando después del despliegue todos los módulos del sistema. El proyecto LIHLITH también desarrollará protocolos de evaluación y *benchmarcks* para permitir la comparación y reproducibilidad en el futuro.
Palabras clave: Sistemas de diálogo, aprendizaje continuo, búsqueda de respuestas, inducción del conocimiento

1 Introduction

A dialogue system makes it possible the human-machine interaction more natural, as the goal of such systems is to converse with humans using ordinary natural language. Traditional dialogue systems are built with hand-crafted rules or they use machine learning techniques trained on a large number of manually annotated sample dialogues. These makes the transferring of the dialogue system to a new target domain exceptionally costly in terms of money and time. Moreover, as the background domain knowledge is fixed and does not expand after deployment, the scope of their applications is limited and is not expanded over time.

Humans use dialogue to learn a great deal of knowledge about the world around us. That is, people have interactive conversations with other people in order to confirm or retract our understanding. Mimicking this behaviour, dialogue systems should also need to learn new knowledge continuously, and this requirement can be fulfilled by means of lifelong learning.

The idea of continuous learning or lifelong learning in machine learning systems has recently gained popularity and is based on the

ISSN 1135-5948. DOI 10.26342/2019-63-17 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

fact that the implemented system, once it has been deployed, continues to improve through interaction with its environment (Chen and Liu, 2016). The application of lifelong learning in the dialogue systems will make the system adapt its strategy to generate the utterances based on the previous interactions that it has with the user, and this will make its results more satisfactory.

Based on the lifelong learning framework, LILITH will develop dialogue systems that learn autonomously from their interactions with the users, and retain this new knowledge in order to answer more accurately in future interactions. This project focuses the research in the scope of question-answering dialogue, that is, dialogues that have the goal to fulfill the information need of the user. The key insight is that the dialogue systems will be designed to get feedback from the user, if necessary, asking explicitly for it to the user ("Are you asking for a specific recipe?" or "I don't know. Can you give me an example of accurate answer?"). Based on this continuous feedback, the system will keep improving after deployment all modules down in the pipeline (dialogue management, knowledge induction, domain knowledge...).

As the evaluation of a dialogue system is challenging and not very well established, LIHLITH project will also develop and deliver evaluation protocols and benchmarks to allow public comparison and reproducibility.

LIHLITH project[1] is a CHIST-ERA project[2] selected for the call 2016 "Lifelong Learning for Intelligent Systems (LLIS)". It started in January 2018 and it will run until December 2020. The project is coordinated by Ixa Group from the University of the Basque Country (UPV/EHU) and these are the other groups of the consortium: Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI) in France, Universidad Nacional de Educación a Distancia (UNED) in Spain, Zurich University of Applied Sciences (ZHAW), and Synapse Développement in France.

2 Objectives

LIHLITH has two general objectives: (i) How to improve the state-the-art in lifelong learning using feedback from users, and (ii) how to apply the previous improvements to dialogue systems, leveraging the interaction with humans to improve the overall quality and the ability to cope with new domains. These main objectives will be achieved through the following more specific objectives:

Objective 1: Define a reproducible evaluation protocol for lifelong learning of dialogue systems. Generate resources for lifelong learning evaluation through benchmarking, in the context of dialogue systems and community question answering (CQA).

Objective 2: Improve dialogue systems using lifelong learning through interaction. The improvements will be sought on better domain models, inference models, and the dialogue management module. Special attention will be put on robustness and resilience in the lifelong learning process. The dialogue system will be improved through the reward provided by the interactions, and will include pro-active behaviour, as well as self-assessment capabilities.

Objective 3: Improve knowledge induction through interaction. The necessary domain knowledge is induced using domain texts, existing knowledge bases, and past interactions. The domain knowledge will be used by the inference and dialogue management modules. The induction process will be improved through the reward provided by the interactions.

Objective 4: Build and evaluate a dialogue system for CQA. The dialogue will involve questions on specific question-answering domains.

Objective 5: Build a real industrial use case for chatbot technology and develop the corresponding chatbot system.

3 First year significant results

The following are the results achieved so far during the first year of the project, organized according to the final objectives in section 2:

- Objective 1: We have produced a survey paper on evaluation of dialogue systems (submitted to a journal). In addition, we have implemented an interface to crowdsource train and evaluation dialogues based on community-based question answering forums, and we have collected a set of dialogues using it.

[1] http://ixa2.si.ehu.eus/lihlith
[2] http://www.chistera.eu/

- Objectives 2 and 3: We have produced the baseline knowledge induction, question answering and dialogue systems, which will be improved during the second year using lifelong learning.

- Objective 3: We have produced the baseline technologies, which will be improved using lifelong learning.

3.1 Survey on evaluation methods for dialogue systems

The survey analyses the methods and concepts developed for the evaluation of dialogue systems. Evaluation is a crucial part during the development process. Dialogue systems are often evaluated by means of human assessment and questionnaires, which tend to be costly and time intensive. As an alternative, techniques to reduce the involvement of human labour have been proposed. The survey differentiates various classes of dialogue systems, such as task-oriented dialogue systems, conversational dialogue systems, and question-answering dialogue systems. This survey is key to set a common understanding on evaluation of dialogue systems, which will be the base for the design of reproducible lifelong learning dialogue benchmarks, to be produced in the second year.

3.2 Collection of dialogues

We have implemented an Amazon Mechanical Turk (AMT) interface to collect dialogues. This interface allows to collect dialogues by an interactive task between two workers on AMT. Each of the workers have a different role: (1) one worker asks questions about a cooking topic (the topic is the title of a thread in the forum), and (2) the other worker has to answer these questions selecting a span of text presented to them (the text is the most voted answer in the thread). The collected dialogues are based on the StackExchange CQA forum, more specifically the forum about cooking[3]. We have already collected around 1700 dialogues, consisting on 8195 questions and 7744 answers. Figure 1 shows one of the collected dialogues. The resulting dataset will be used to train and test the neural dialogue system for CQA that we are going to implement in the second year.

[3]https://cooking.stackexchange.com/

Q: How long can you keep chocolate, and what is the best way to store it?
A: Regardless of type, all chocolate should be stored in a cool and low humidity (dry) place away from direct sunlight.
Q: How long does chocolate last before losing flavor?
A: Dark chocolate will last for years. Milk and white chocolate will last for a much shorter time (a few months), because of their milk content.
Q: Once it gets that white stuff on the outside, is it done?
A: No, This kind of chocolate is still suitable for any application where the chocolate will be fully melted (most baking)

Figure 1: Example of a dialogue collected via AMT

3.3 Demo for generating triples of knowledge from a text

We have made a demo available online to generate triples of knowledge from a text[4]. This system, first, processes the input text with a natural language processing pipeline (coreNLP): sentence splitting, tokenization, part-of-speech tagging, Named Entities recognition, coreference resolution and dependency analysis. Then, departing from this information, it generates sets of triples expressing concepts and relations contained in each of the sentences. Figure 2b shows an example of the output produced by this tool given the input text in 2a.

3.4 Demo of a question answering system

We have a demo for a baseline question answering system, which is a subset of an industrial chatbot system. This system generates all the questions and answers that can be asked about the product or service from client documentation. Questions are generated on the basis of manually defined rules. For example, after analyzing the sentence *"You can reset your password by sending an email to admin@example.com"*, the question *"How can I reset my password?"* is generated. Similarly, if an enumerating structure is found in a document, a question of the type *"What are the steps for [subject of the primer]"* is generated. The answers to these questions are always excerpts of the text.

[4]http://nlp.uned.es/lihlith-project/ki/

John W. Carr (...) was a North Dakota Republican Party politician who served as the 15th Lieutenant Governor of North Dakota under Governor George F. Shafer.

(a) Input text

John W. Carr a **politician**
John W. Carr a **person**
north Dakota Republican Party a **organization**
John W. Carr a **lieutenant governor**
North Dakota a **location**
George F. Shafer a **person**
north Dakota Republican Party organization:has:lieutenant governor **John W. Carr**
15th ordinal:has:lieutenant governor **John W. Carr**
North Dakota location:has:lieutenant governor **John W. Carr**
John W. Carr lieutenant governor:serve under:person **George F. Shafer**

(b) Set of triples for each sentence.

Figure 2: Example of the generated triples of knowledge from a text

3.5 Demo of a task-oriented dialogue system

We have made available online a demo of a baseline task-oriented dialogue system working on the cooking domain[5]. It handles two different types of scenarios: (1) the user wants to find a recipe meeting his criteria, and (2) the user asks a question relative to the cooking domain. For the first scenario, the system carries out an interactive search in a database which contains 1,064 recipes extracted from the Wikipedia textbook "Cookbook". The database contains information about the name, the details, the ingredients, the variations, the procedure and the categories for each recipe. For the second scenario, the system accesses, based on semantic textual similarity methods, the unstructured data consisting in sentences from 784 non-recipe documents that can be found on "Cookbook". An example of a dialogue can be seen in Figure 3.

3.6 "Dialogue systems and lifelong learning" special session at IWSDS 2019

LIHLITH project organized the "Dialogue systems and lifelong learning" special session[6] at the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2019), which was held in Sicily on April 24-26, 2019 (Agirre, Jonsson, and Larcher, 2019). The main objective of this special session was to gather researchers interested in dialogue systems that interact with the users in order to learn about new domains or acquire new knowledge. Thus, continuous learning for question answering and dialogue systems and learning inferences and understanding strategies from the interactive structure of a corpus where some of the topics discussed in the session. The workshop gathered feedback from the community on the ideas and techniques being explored in LIHLITH.

SYSTEM: Hello, Are you hungry?
USER: I want to cook a chocolate mousse
SYSTEM: 3 recipes correspond to your search, including : atkins-friendly chocolate mousse, chocolate mousse cake (vegan) and chocolate mousse . Which one do you want?

Figure 3: Example of a dialogue with the baseline task-oriented dialogue system

Acknowledgments

This work has been supported by ERA-Net CHIST-ERA LIHLITH Project funded by the Agencia Estatal de Investigación (AEI, Spain) projects PCIN-2017-118/AEI and PCIN-2017-085/AEI, the Agence Nationale pour la Recherche (ANR, France) projects ANR-17-CHR2-0001-03 and ANR-17-CHR2-0001-04, and the Swiss National Science Foundation (SNF, Switzerland) project 20CH21 174237.

References

Agirre, E., A. Jonsson, and A. Larcher. 2019. Framing lifelong learning as autonomous deployment: Tune once live forever. In *Proceedings of the Dialogue systems and lifelong learning special session of the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS).*

Chen, Z. and B. Liu. 2016. *Lifelong Machine Learning.* Morgan & Claypool Publishers.

[5]https://lihlith.limsi.fr/dialog.php (user/pass: lihlith/recipe?)

[6]https://iwsds2019.unikore.it/dialogue-systems-and-lifelong-learning/

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 151-154 recibido 31-03-2019 revisado 25-04-2019 aceptado 17-05-2019

Unsupervised Neural Machine Translation, a new paradigm solely based on monolingual text

Traducción Automática Neuronal no Supervisada, un nuevo paradigma basado solo en textos monolingües

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group, HiTZ Basque Center for Language Technologies
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract: This article presents UnsupNMT, a 3-year project of which the first year has already been completed. UnsupNMT proposes a radically different approach to machine translation: unsupervised translation, that is, translation based on monolingual data alone with no need for bilingual resources. This method is based on deep learning of temporal sequences and uses cutting-edge interlingual word representations in the form of cross-lingual word embeddings. This project is not only a highly innovative proposal but it also opens a new paradigm in machine translation which branches out to other disciplines, such us transfer learning. Despite the current limitations of unsupervised machine translation, the techniques developed are expected to have great repercussions in areas where machine translation achieves worse results, such as translation between languages which have little contact, e.g. German and Russian.
Keywords: Machine Translation, Deep Learning, Word Embedding

Resumen: Este artículo presenta UnsupNMT, un proyecto de 3 años del que ha trascurrido la primera anualidad. UnsupNMT plantea un método radicalmente diferente de hacer traducción automática: la traducción no supervisada, es decir, basada exclusivamente en textos monolingües sin ningún recurso bilingüe. El método propuesto se basa en aprendizaje profundo de secuencias temporales combinado con los últimos avances en representación interlingual de palabras ("cross-lingual word embeddings"). Además de ser una propuesta propiamente innovadora, abre un nuevo paradigma de traducción automática con ramificaciones en otras disciplinas como el aprendizaje por transferencia ("transfer learning"). A pesar de las limitaciones actuales de la traducción automática no-supervisada, se espera que las técnicas desarrolladas tengan gran repercusión en áreas donde la traducción automática consigue peores resultados, como la traducción entre pares de idiomas con poco contacto, tales como alemán o ruso.
Palabras clave: Traducción Automática, Aprendizaje Profundo, Word Embeddings

1 Introduction

Machine translation has been one of the most prominent applications of artificial intelligence since the very beginnings of the field. In addition to its intrinsic interest given the difficulty and completeness of the problem, machine translation has a huge practical interest in our increasingly global word, as it promises to break the language barrier while keeping the cultural heritage and diversity of all the languages spoken in the world.

In very recent times, previous approaches have been superseded by neural machine translation (NMT), which has now become the dominant paradigm to machine translation (Bahdanau, Cho, and Bengio, 2014). As opposed to the traditional statistical machine translation (SMT), NMT systems are trained end-to-end, take advantage of continuous representations that greatly alleviate the sparsity problem, and make use of much larger contexts, thus mitigating the locality problem. Thanks to this, NMT has been reported to significantly improve over SMT both in automatic metrics and human evaluation.

Current NMT methods require expensive annotated data, as they fail terribly when the training data is not big enough (Koehn

ISSN 1135-5948. DOI 10.26342/2019-63-18 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

and Knowles, 2017). Unfortunately, the lack of large parallel corpora is a practical problem for the vast majority of language pairs, including low-resource languages (e.g. Basque) as well as many combinations of major languages (e.g. German-Russian). Several authors have recently tried to address this problem using pivoting or triangulation techniques (Chen et al., 2017) as well as semi-supervised approaches (He et al., 2016), but these methods still require a strong cross-lingual signal.

In this project, we introduce unsupervised neural machine translation, a new paradigm where the system learns to translate between two languages without the need of any bilingual dictionary or translation memories. That is, given large bodies of monolingual text (monolingual corpora), the system is able to extract the patterns which allow to translate from one language to another.

Our approach would use standard deep learning models for sequence-to-sequence learning. More concretely, we would follow the encoder/decoder architecture, combining a single language independent encoder that would compose the cross-lingual word embeddings and several language specific decoders that would decompose this representation back into the appropriate language. The system would be trained in an unsupervised manner following the same principle of denoising autoencoders, and we would explore additional techniques like adversarial training and backtranslation to enhance proper learning (Sennrich, Haddow, and Birch, 2016).

The new techniques will open new research avenues on machine translation. We are at a very good position to check whether this new paradigm allows to improve the state-of-the-art in MT, especially for less-resourced language pairs and domains. Moreover, the developed methods to train cross-lingual sentence representations will also be useful for cross-lingual transfer learning, as already shown by cross-lingual word representations. Finally, the viability to induce translation models in a completely unsupervised environment would empirically prove the existence of an inherent connection among all languages, which has a great interest from the point of view of Linguistics.

2 Goals

The overall goal of this project is to develop unsupervised learning methods for neural machine translation. In order to maximize impact of this nascent technology, the project will be structured in three goals:

Goal 1: Develop methods to train NMT models in a completely unsupervised manner, relying solely in monolingual corpora. This is the core goal of this project. The rest of the goals explore practical ramifications and impact of unsupervised NMT.

Goal 2: Improve the state-of-the-art in machine translation in different real-world scenarios where we have access to varying degrees of cross-lingual signals. This goal explores whether the new paradigm has practical applications.

Goal 3: Transfer learning. This goal explores whether the new paradigm has practical implications in transferring natural language processing systems from a resource-rich language to a less-resourced language.

3 Technical approach

We propose a radically new approach to unsupervised machine translation based on deep learning, a direction that has shown to be highly successful in other related areas, including standard supervised machine translation itself through NMT. The core of our approach is to learn to compose cross-lingual word representations in an unsupervised manner. Then, use those embeddings to generate a first translation system that can be improved by monolingual techniques, such as denoising autoencoders or backtranslation.

Regarding the first step, and in order to obtain our cross-lingual word representations, we will rely on techniques pioneered by us to build cross-lingual embedding mappings (Artetxe, Labaka, and Agirre, 2017; Artetxe, Labaka, and Agirre, 2018a). Recent works in unsupervised word embedding mapping (Lample et al., 2018a) have managed to obtain results comparable to previous supervised techniques, which we managed to improve (Artetxe, Labaka, and Agirre, 2018b). However, existing methods are based on the geometric interpretation of the embedding space (e.g. minimizing the Euclidean distance, maximizing the cosine similarity...), which is unnatural and presumably suboptimal for machine translation. For that reason, we will

explore alternative interpretations. At the same time, while existing models are bilingual, we plan to extend them to the multilingual scenario, so we can exploit the relationship among several languages at the same time.

These cross-lingual word-embeddings contain enough bilingual information to generate a rudimentary word-by-word translation system, which can be improved by different techniques. Either by directly generating a neural translation system that makes use of these pre-trained embeddings and is trained using techniques such as denoising autoencoders or backtranslation, or, taking advantage of the SMT modular architecture, by using these embeddings to generate a phrase-table that can be combined with a language model.

In relation to Goal 2, the project will explore extensions of the previous approach to exploit cross-lingual signals of different degree when available, which would be used to improve the state-of-the-art in machine translation in different practical scenarios. In particular, we plan to:

- If a small parallel corpus is available, use it to fine-tune our model taking care not to overfit. This type of fine-tuning has already been shown to be effective in NMT in the case of domain adaptation (Chu, Dabre, and Kurohashi, 2017).
- If a comparable corpus is available, use our model to iteratively extract reliable parallel sentences from it and improve the model with these parallel sentences.

While the main scenario under consideration in this project is that of machine translation, we also plan to explore the application of the developed methods in cross-lingual transfer learning (Goal 3), where a model is trained in one language and used in a different one. This has a great practical interest, as it allows to leverage the largely available annotated data and resources in major languages (in particular, English), to other less resourced ones (e.g. Basque or even Spanish for some tasks like co-reference, sentiment analysis or named-entity recognition). For that purpose, we plan to extend the existing methods based on cross-lingual word embeddings to incorporate the entire encoder learned with our approach, so the resulting system does not only account for word-level relations but also for more complex phrase- or sentence-level relations.

4 *Current progress*

The first attempts which obtained promising results in standard machine translation benchmarks using monolingual corpora only (Artetxe et al., 2018; Lample et al., 2018a) build upon unsupervised cross-lingual embedding mappings, which independently train word embeddings in two languages and learn a linear transformation to map them to a shared space. The resulting cross-lingual embeddings are used to initialize a shared encoder for both languages, and the entire system is trained using a combination of denoising autoencoding, back-translation and, in the case of (Lample et al., 2018a), adversarial training.

During the first months of the project we have made progress mainly in our core Goal 1, where we show that the modular architecture of phrase-based SMT is more suitable for this problem. Our work (Artetxe, Labaka, and Agirre, 2018c), concurrent with (Lample et al., 2018b), adapted the same principles discussed above to train an unsupervised SMT model, obtaining large improvements over the original unsupervised NMT systems. More concretely, we learn cross-lingual n-gram embeddings from monolingual corpora based on the mapping method discussed earlier, and use them to induce an initial phrase-table that is combined with an n-gram language model and a distortion model. This initial system is then refined through iterative back-translation.

More recently (Artetxe, Labaka, and Agirre, 2019), we identifiy and address several deficiencies of existing unsupervised SMT approaches by exploiting subword information, developing a theoretically well founded unsupervised tuning method, and incorporating a joint refinement procedure. Moreover, we use our improved SMT system to initialize a dual NMT model, which is further fine-tuned through on-the-fly back-translation. Together, we obtain large improvements over the previous state-of-the-art in unsupervised machine translation. For instance, we get 22.5 BLEU points in English-to-German WMT 2014, 5.5 points more than the previous best unsupervised system, and 0.5 points more than the (supervised) shared task winner back in 2014.

Acknowledgments

UnsupNMT is a project funded by the Spanish Ministry of Economy, Industry and Competitiveness (TIN2017-91692-EXP).

References

Artetxe, M., G. Labaka, and E. Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

Artetxe, M., G. Labaka, and E. Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, October-November. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, April.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September.

Chen, Y., Y. Liu, Y. Cheng, and V. O. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935. Association for Computational Linguistics.

Chu, C., R. Dabre, and S. Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391. Association for Computational Linguistics.

He, D., Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., pages 820–828.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.

Lample, G., A. Conneau, L. Denoyer, and M. Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, April.

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 155-158 recibido 29-03-2019 revisado 07-05-2019 aceptado 17-05-2019

Project CAVIAR
CApturing VIewers' Affective Response

Inferencia de la respuesta afectiva de los espectadores de un vídeo

Fernando Fernández-Martínez[1], Zoraida Callejas[2], Ricardo Kleinlein[1], Cristina Luna Jiménez[1], Juan Manuel Montero[1], José Manuel Pardo[1]
[1]Universidad Politécnica de Madrid, Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Madrid, Spain
[2]Universidad de Granada, Department of Languages and Computer Systems, Granada, Spain
fernando.fernandezm@upm.es; Tel.: +34-915495700 (ext. 4228)

Abstract: In this project we propose the automatic analysis of the relation between the audiovisual characteristics of a multimedia production and the impact caused in its audience. With this aim, potential synergies are explored between different areas of knowledge including, among others: audiovisual communication, computer vision, multimodal systems, biometric sensors, social network analysis, opinion mining, and affective computing. Our efforts are oriented towards combining these technologies to introduce novel computational models that could predict the reactions of spectators to multimedia elements across different media and moments. On the one hand, we study the cognitive and emotional response of the spectators while they are watching the media instances, using neuroscience techniques and biometric sensors. On the other hand, we also study the reaction shown by the audience on social networks by relying on the automatic collection and analysis of different metadata related to the media elements, such as popularity, sharing patterns, ratings and commentaries.
Keywords: Opinion mining, aesthetics, multimedia information retrieval, metadata, video recommendation, indexation, biometrics, emotion, subjective response prediction

Resumen: Este proyecto propone el análisis de la posible dependencia entre el contenido audiovisual de una producción multimedia y el impacto causado por ésta en sus espectadores. Para ello, nos apoyamos en diferentes áreas de conocimiento tales como comunicación audiovisual, visión por computador, sistemas multimodales, sensores biométricos, análisis de redes sociales, análisis de opinión o computación afectiva, entre otras, con el objetivo de diseñar nuevos modelos computacionales que permitan predecir las reacciones de los espectadores de un vídeo de forma transversal a los medios y momentos en que éstas se producen. Trabajamos principalmente con dos tipos de respuesta: la respuesta cognitiva y emocional inmediata de los espectadores durante el visionado, que medimos utilizando técnicas de neurociencia y sensores biométricos, y la reacción expresada en redes sociales, cuyo impacto es cuantificado mediante el análisis automático de diferentes metadatos recabados para dichos vídeos, tales como popularidad, patrones de compartición, valoraciones y comentarios realizados en las redes.
Palabras clave: Análisis de opinión, estética, recuperación de la información multimedia, metadatos, recomendación e indexación de vídeo, biometría, emoción, predicción de respuesta subjetiva

ISSN 1135-5948. DOI 10.26342/2019-63-19 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

In a world where new technologies are progressively more related with multimedia information, it is essential to have tools that simplify the treatment of this type of data. In this sense, a problem that has recently attracted the interest of the scientific community is the development of models for objectivizing what is subjective, concretely, to measure the quality and impact of audiovisual productions (Luo, Wang, and Tang, 2011).

Project CAVIAR proposes the automatic analysis of the relation between the audiovisual characteristics of a multimedia production and the impact caused in its audience. With this aim, we have gathered together different knowledge areas including: audiovisual communication, computer vision, multimodal systems, biometric sensors, social network analysis, and affective computing. Our contribution will be based in combining these technologies to study thoroughly and predict the reactions of spectators to multimedia elements across different media and moments. We will work mainly with two response types: the immediate reaction while viewing the audiovisual, and the shown reaction on social networks.

On the one hand, we will study the cognitive and emotional response of the spectators while they are watching the audiovisual, using neuroscience techniques and biometric measures based on sensors. This will allow transitioning from the typical scenario only based on opinion surveys to a more precise model that will also allow measuring reactions to specific passages (not only the general reaction).

On the other hand, the natural reaction will also be studied focusing on the audience's behaviour on social networks by means of automatic analysis of the metadata corresponding to the multimedia elements watched. We will establish techniques to evaluate the impact on the users through the study of the popularity, sharing patterns, ratings and commentaries received, as a way of quantifying the effect of the audiovisual and its implications.

Regarding the automatic inference of the spectator's perception, the audiovisual content may provoke cognitive effects (e.g. attract attention) and affective responses (e.g. arouse a particular emotional state) in the audience, which are the typical objectives of every multimedia production. However, the factors that determine if they have a positive or negative effect (or no effect at all) on the audience's response are still unknown to a great extent, and their corresponding analysis is far from automatic. To achieve this, we will identify the relevant elements of the audiovisual production in two channels: video and sound, and will study their correspondence with the perception of the spectators in the lines described earlier, generating an automatic prediction system.

2 Project objectives

The main purposes of this project are:

- To present a novel approach based on different metadata (such as related opinions and comments) extracted from social networks (such as YouTube) for the automatic annotation of videos in terms of their impact or their expected or potential perceived value.
- To explore neuroscience techniques as an alternative and totally different automatic annotation solution by measuring biometric indicators of the cognitive and affective response of the audience to the video content stimuli.
- To expand upon existing research to investigate how audiovisual content can influence perception and emotions and use related findings for the construction of effective computational models.
- To develop technology for the automatic assessment of the effectiveness of a video and to provide novel automatic tools for multimedia document retrieval, focusing on the automatic analysis of the audiovisual content of a video as a mean to derive suitable principles for predicting the audience's perception.

3 Methodology and viability

3.1 Impact on social networks

CAVIAR researchers have already tested the feasibility of this proposal by means of a pilot study (Fernández-Martínez, Hernández-García, and de María, 2015) which has been conducted by choosing car commercials and related metadata downloaded from YouTube

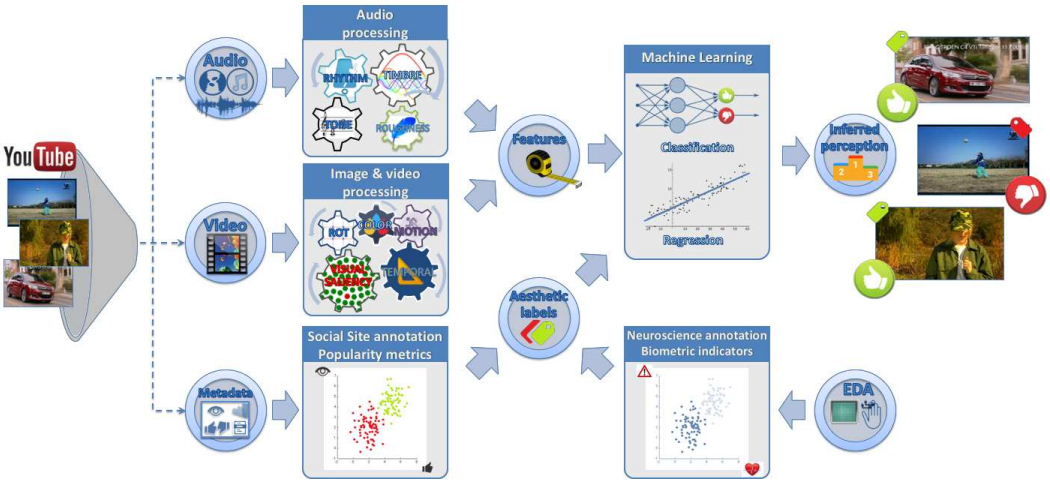


Figure 1: Overview of the proposed approach

as the evaluation domain for some of the suggested approaches (figure 1 shows the overall pipeline of the proposal).

To that end, a novel approach to automatically annotate a corpus composed of 138 videos is presented. Videos are grouped into 2 classes corresponding to different satisfaction levels, by means of a regular k-means algorithm applied to the video metadata related to users feedback. Evaluation results show that simple linear logistic regression models based on the 10 best visual descriptors and on the 10 best audio descriptors individually perform reasonably well, achieving a classification accuracy of roughly 70% and 75%, respectively. Combination of audio and visual descriptors (Hernández-García, Fernández-Martínez, and Díaz-de María, 2016) yields better performance, roughly 86% for the top-20 selected from the entire descriptor set, but tipping the balance in favor of the audio ones (i.e. 17 vs 3). Audio content bigger influence in this domain is also evidenced by a side analysis of the video comments (Fernández-Martínez et al., 2014).

3.2 Biometric response

Electrodermal activity, EDA, is a psychophysiological indicator which can be considered a somatic marker of the emotional and attentional reaction of subjects towards stimuli. EDA measurements are not biased by the cognitive process of giving an opinion or a score to characterize the subjective perception, and group-level EDA recordings integrate the reaction of the whole audience, thus reducing the signal noise.

Thanks to our collaboration agreement with Sociograph company (`www.sociograph.es`), a new corpus of audiovisual contents has already been annotated from recorded physiological signals (EDA) measured on 270 participants while they watched a concatenation of videos. These videos correspond to 136 short films selected from the Jameson Notodofilmfest Short Film Festival 2015.

In (Hernández-García, Fernández-Martínez, and Díaz-de María, 2017), we predict the levels of emotion and attention, derived from EDA records, by means of a small set of low-level visual descriptors computed from the video stimuli. Linear regression experiments show that our descriptors predict significantly well the sum of emotion and attention levels, reaching a coefficient of determination $R^2 = 0.25$. In (García-Faura et al., 2019) we extend the previous work by labeling short video clips according to the audience's emotion (high vs. low) and attention (increasing vs. decreasing), derived from EDA records. Here, we propose a set of low-level audiovisual descriptors and train binary classifiers that predict the emotion and attention with 75% and 80% accuracy, respectively.

Both works reinforce the usefulness of such low-level audiovisual descriptors to model video in terms of the induced affective response and set a promising path for further research on the prediction of emotion and attention from videos using EDA. In addition to the release of new datasets and annotation methods, the team plans to also take advantage of other data sets (Baveye et al., 2015; Demarty et al., 2016) from the literature, so that the advances obtained in CAVIAR can be compared with the state of the art and developments of other researchers.

4 Potential applications

From an applicability point of view, proposed solutions could pave the way for a new generation of recommendation, indexation and summarization systems that could change the way consumers interact with multimedia search engines by allowing them to actively use enhanced search choices inspired on audiovisual, social and/or biometric features that help the retrieved content to be more accurate and related to the affective response and more personalized. In addition, they could also help training inference models that could be far more representative and indicative of the viewers' real feelings and attitudes towards videos.

Anticipating the subjective value perceived by the viewers of any audiovisual content could also enable a more efficient multimedia content production. For instance, it could be a low-cost and reliable alternative to costly conventional test screening processes which are traditionally based on artificial setups that tend to bias the actual feelings or opinions of the participants.

Finally, the use of the proposed technology could also boost advertising by allowing us to anticipate the impact of a marketing campaign on the potential customers before it is released to the media, thus paving the way for the end of trial and error in advertising and saving up many costs. Aside, it could also help pointing out what audiovisual aspects would need to be changed for the campaign to be successful and effective.

5 Technology transfer

As a result of our recent efforts and advances, our automatic aesthetic assessment technology is currently being applied in the "ESITUR-UPM: Interactive Tourist Showcase" project, led by MOVILOK INTERACTIVIDAD MÓVIL S.L., and funded by the Spanish Ministry of Economy and Competitiveness under the "Retos de Colaboración 2016" call, which is oriented to the design and development of useful and efficient solutions for 'Smart tourism' to improve the tourism experience of its users.

Acknowledgments

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the ESITUR (MINECO, RTC-2016-5305-7), CAVIAR (MINECO, TEC2017-84593-C2-1-R), and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU used in this research.

References

Baveye, Y., E. Dellandrea, C. Chamaret, and L. Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*.

Demarty, C.-H., M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefèbvre. 2016. Predicting media interestingness task. In *MediaEval 2016*.

Fernández-Martínez, F., A. H. García, A. Gallardo-Antolín, and F. D. de María. 2014. Combining audio-visual features for viewers' perception classification of youtube car commercials. In *2nd International Workshop on Speech, Language, and Audio in Multimedia*, SLAM'14.

Fernández-Martínez, F., A. Hernández-García, and F. D. de María. 2015. Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials. *Expert Systems with Applications*.

García-Faura, A., A. Hernández-García, F. Fernández-Martínez, F. Díaz-de María, and R. San-Segundo. 2019. Emotion and attention: Audiovisual models for group-level skin response recognition in short movies. *Web Intelligence*, 17.

Hernández-García, A., F. Fernández-Martínez, and F. Díaz-de María. 2016. Comparing visual descriptors and automatic rating strategies for video aesthetics prediction. *Image Communications*.

Hernández-García, A., F. Fernández-Martínez, and F. Díaz-de María. 2017. Emotion and attention: Predicting electrodermal activity through video visual descriptors. In *International Conference on Web Intelligence*, WI '17.

Luo, W., X. Wang, and X. Tang. 2011. Content-based photo quality assessment. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 159-162 recibido 29-03-2019 revisado 08-05-2019 aceptado 17-05-2019

LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos

LINGUATEC: Development of linguistic resources to advance the digitisation of the languages of the Pyrenees.

Itziar Aldabe[6], Josu Aztiria[3], Francho Beltrán[1], Myriam Bras[2], Klara Ceberio[3], Itziar Cortes[3], Jean-Baptiste Coyos[4], Benaset Dazeas[5], Louise Esher[2], Gorka Labaka[6], Igor Leturia[3], Kepa Sarasola[6], Aure Séguier[5], Jean Sibille[2]
[1]ARAGON TURISMO [2]CNRS-UNIV TOULOUSE 2 [3]ELHUYAR
[4]EUSKALTZAINDIA [5]LO-CONGRÉS [6]UPV/EHU
Información de contacto: j.aztiria@elhuyar.eus
https://linguatec-poctefa.eu/

Resumen: El objetivo del proyecto es desarrollar, probar y difundir nuevos recursos, nuevas herramientas y aplicaciones lingüísticas innovadoras para mejorar el nivel de digitalización del aragonés, vasco y occitano. Resultados esperados: (1) Hoja de ruta para la digitalización del aragonés, (2) Nuevos recursos lingüísticos, (3) Herramientas lingüísticas desarrolladas (síntesis de voz occitana, aragonesa y vasca del País Vasco francés, detector de texto occitano y variantes del occitano, mejora de la traducción automática del francés al occitano, del castellano vasco, del castellano al aragonés, (4) Aplicaciones innovadoras desarrolladas en los idiomas de los Pirineos.
Palabras clave: Traducción automática, Idiomas con recursos limitados, Corpus bilingüe

Abstract: The goal of the project is to develop, test and disseminate new innovative linguistic resources, tools and solutions for a better digitalization level of the Aragonian, Basque and Occitan languages. As a result, we will obtain, among others, (1) a road map of Aragonian Digitalization, (2) new monolingual and bilingual lexicons and morphosyntactic and syntactic analysers for Occitan, (3) a Northern Basque speech recognition system, and several linguistic tools as well as (4) new innovative solutions for Aragonian, Basque and Occitan.
Keywords: Machine Translation, Languages with limited resources, Bilingual Corpora

1 *Introducción*

El proyecto LINGUATEC[1] se desarrolla dentro del marco de cooperación interregional POCTEFA (Programa INTERREG V-A España-Francia-Andorra) está financiado con fondos FEDER de la Comunidad Europea y cofinancia proyectos de cooperación transfronteriza diseñados y gestionados por actores de ambos lados de los Pirineos y de las zonas litorales que participan en el Programa desarrollo inteligente, sostenible e integrador del territorio.

Este proyecto también se enmarca perfectamente en los objetivos del informe del Parlamento Europeo aprobado en septiembre de 2018 sobre "La igualdad lingüística en la era digital"[2]. El informe fue aprobado por amplia mayoría y entre otras recomendaciones pide crear o ampliar proyectos de diversidad lingüística digital, que investiguen las necesidades digitales de todas las lenguas europeas, incluyendo desde aquellas con muy pocos hablantes hasta las que cuentan con gran número de hablantes, con el fin de combatir la brecha digital y contribuir a preparar a dichas lenguas para el futuro digital sostenible". También pide a las administraciones "que mejoren el acceso a los servicios e informaciones en línea en diferentes lenguas, especialmente en el caso de servicios en regiones transfronterizas". Con ello el informe pretende reducir las desigualdades entre lenguas y comunidades lingüísticas, fomentar el acceso equitativo a los servicios y estimular la movilidad en Europa de las empresas, los ciudadanos y los trabajadores, así como garantizar la creación de un mercado único digital multilingüe inclusivo".

Siguiendo los objetivos del marco POCTEFA y el informe del Parlamento Europeo el proyecto LINGUATEC está desarrollando nuevos productos de tecnología de la lengua para mejorar el

[1]https://linguatec-poctefa.eu/
[2]http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_ES.html

ISSN 1135-5948. DOI 10.26342/2019-63-20
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

nivel de digitalización del aragonés, vasco y occitano. Además de la creación de nuevos recursos y herramientas lingüísticas tales como léxico monolingüe y bilingüe, sistemas de análisis morfosintácticos y sintácticos, sistemas de traducción automática o sistemas de reconocimiento de voz. Dentro del proyecto se crearán 6 aplicaciones en las que esos recursos y herramientas se puedan utilizar, y a ser posible de forma coordinada para todas las lenguas.

Este artículo comienza presentando brevemente el estado de las lenguas con las que trabajaremos en el proyecto y con la presentación de los socios del consorcio. El tercer apartado describe los recursos y herramientas a desarrollar. El cuarto presenta las 6 aplicaciones informáticas que se desarrollarán aprovechando los recursos creados. Finalmente se describe el plan de diseminación para el año 2020 que será el último del proyecto.

2 Idiomas y Grupos Participantes

El consorcio agrupa tres socios que trabajan con el euskera, dos con el occitano y uno con el aragonés. La situación actual de estos tres idiomas no es la misma. La red META-NET elaboró en 2012 una serie de 32 "libros blancos" que mostraban grandes diferencias de desarrollo entre los diferentes idiomas con respecto a las tecnologías de la lengua[3]. Según esos estudios en un claro primer nivel se encuentra el inglés; el español y el francés aparecen en un segundo nivel. El libro blanco The Basque Language in the Digital Age concluye que el euskera necesita de mucha más investigación fundamental y recursos para su supervivencia digital (Hernáez et al., 2012). El occitano y el aragonés no llegaron a considerarse dentro del grupo de las 32 lenguas de estudio, lo cual confirma que su situación actual es aún más debil que la del euskera. Una hipótesis de partida en este proyecto es el pensar que la experiencia ontenida en la digitalización para el euskera puede ser útil a la hora de realizar pasos en la digitalización del occitano y del aragonés.

2.1 Elhuyar Fundazioa (líder del proyecto)

La Fundación Elhuyar es una entidad sin ánimo de lucro fundada en 1972 para la promoción y difusión del euskera y de la ciencia y la tecnología (libros revistas, diccionarios...). Desde 2002 su unidad de I+D ha participado en múltiples proyectos para desarrollar aplicaciones relacionadas con diccionarios y corpus, con correctores ortográficos y de estilo, tecnologías para la traducción, herramientas de extracción terminológica, gestión de la información y del conocimiento, tecnologías del habla y enseñanza de idiomas. En 2018 ha creado una aplicación para la gestión multilingüe en sistemas de contenidos (CMS) (Cortes et al., 2018), y un exitoso sistema de traducción neuronal (Etchegoyhen et al., 2018). También gestiona una app en código abierto de traducción automatica [4].

2.2 Lo Congrès Permanent de la Lenga Occitana

LO CONGRÈS es, junto con Elhuyar Fundazioa, la entidad de la que parte la idea y coordina el proyecto. LO CONGRÈS es el organismo interregional para la regulación del idioma occitano. Su objetivo es contribuir a la vitalidad y el desarrollo del occitano mediante la producción de herramientas computacionales (lexicografía, lexicología, terminología, neología, fonología , ortografía, gramática y toponimia).

2.3 UMR 5263 - CNRS (Toulouse)

El laboratorio CLLE-ERSS, UMR 5263 CNRS y Universidad Tolosa 2 Joan Jaurès. En el proyecto LINGUATEC aportará sus competencias en lingüística occitana y francesa y en el tratamiento automático de las lenguas. Cuenta con experiencia en la construcción de recursos para el occitano (Bernhard et al., 2018).

2.4 Aragón Turismo

La Sociedad De Promoción y Gestión del Turismo Aragonés, SLU, o TURISMO DE ARAGÓN, es la entidad encargada por el Gobierno de Aragón de la promoción turística y del patrimonio regional, incluyendo la lengua aragonesa. Tiene mucha experiencia en actividades de promoción y difusión. Ha desarrollado proyectos de nuevas tecnologías en colaboración con el Instituto Tecnológico Aragonés (ITA) [5].

2.5 Euskaltzaindia

La Academia de la Lengua Vasca Euskaltzaindia, entidad oficial encargada de cuidar y normalizar el uso del idioma vasco, tiene experiencia en bases de datos lexicográficas, corpus lingüísticos, en indexación y clasificación de textos para búsqueda, y además, con la ayuda de este tipo de tecnología, para difundir su servicio de consulta

[3] http://www.meta-net.eu/whitepapers/overview

[4] http://www.mitzuli.com/en/

[5] https://www.itainnova.es/es

sobre el uso del euskara en la parte norte del País Vasco.

2.6 Ixa Taldea (UPV/EHU)

El Grupo Ixa de la Universidad Del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU) viene trabajando en tecnologias de la lengua para el euskera desde 1988. Entre sus 23 productos registrados se destacan los siguientes: (i) el corrector ortográfico Xuxen[6], comercializado desde 1984; (ii) el sistema de traducción automática Opentrad-Matxin, en explotación por la empresa Elhuyar[7]; (iii) la base de conocimiento Multilingual Central Repository[8], que aglutina, entre otros, los wordnets del español, catalán, gallego y euskera; (iv) el corpus etiquetado de Ciencia y Tecnología, realizado en colaboración con Elhuyar[9]; (v) el corpusetiquetado de referencia para el EPEC[10]; (vi) la herramienta de desambiguación de sentidos y entidades basada en grafos UKB[11]; y (vii) las cadenas de procesamiento de texto Ixa-pipes[12] e Ixa-Kat[13].

Su papel en el proyecto se centra en el asesoramiento basado en la experiencia previa en el tratamiento de lenguas de escasos recursos (Alegria y Sarasola, 2017), en la mejora de los sistemas de traducción automática existentes en el proyecto y en la creación inicial de un sistema de traducción para el par de lenguas euskera-francés.

3 Desarrollo de recursos y herramientas lingüísticas para facilitar el tratamiento automatizado de las lenguas

Se pretende en primer lugar establecer un marco metodológico común para el desarrollo de la informatización en estas lenguas de los Pirineos. Para ello hemos realizado una puesta en común sobre las estrategias de digitalización de cada lengua y hemos acordado las prioridades para el desarrollo de recursos lingüísticos. Adicionalmente también estamos desarrollando herramientas lingüísticas para avanzar en la interoperabilidad de las lenguas de los Pirineos.

Como resultado de esta puesta en común se establecieron las siguientes prioridades, estableciéndose la tarea de cada socio:

Occitano:

- LO-CONGRÉS Detector textual del occitano
- LO-CONGRÉS Detector textual de variantes del occitano
- LO-CONGRÉS Mejora de la traducción automática en francés occitano
- LO-CONGRÉS Síntesis vocal en occitano,
- CNRS Léxico monolingüe occitano: Colección de formas flexionadas.
- CNRS Léxico bilingüe occitano/otras lenguas
- CNRS Análisis morfosintáctico
- CNRS Análisis sintáctico

Aragonés:

- TURISMO DE ARAGÓN Mejora de la traducción automática en español-aragonés
- TURISMO DE ARAGÓN Síntesis vocal en aragonés

Euskera:

- ELHUYAR Síntesis de voz en euskera de Iparralde (Iparrahotsa 2.0)
- ELHUYAR Reconocimiento de voz en euskera: reconocer y clasificar palabras
- ELHUYAR Mejora de la traducción automática para el par español-euskera
- UPV/EHU Mejora de la traducción automática español-euskera
- UPV/EHU + ELHUYAR Investigación sobre traducción euskera-francés

4 Desarrollo de Aplicaciones Innovadoras en el Ámbito Lingüístico

Este proyecto pretende desarrollar una serie de aplicaciones innovadoras en el ámbito lingüístico. Con ello queremos facilitar la interoperabilidad entre estas lenguas de los Pirineos. Las aplicaciones previstas y el socio responsable de cada una de ellas son las siguientes:

- LO CONGRES Barra descargable de traducción automática para sitios web,
- ELHUYAR Aplicación de Traducción Automática para CMS,

[6] http://xuxen.eus/es/xuxen5
[7] http://matxin.elhuyar.eus/
[8] http://adimen.si.ehu.eus/web/MCR
[9] http://www.ztcorpusa.eus/aurkezpena.htm
[10] http://www.ixa.eus/epec-dep-deskarga
[11] http://ixa2.si.ehu.eus/ukb
[12] http://ixa2.si.ehu.eus/ixa-pipes/
[13] http://ixa2.si.ehu.eus/ixakat/

- ELHUYAR App de traducción automática entre las lenguas de los Pirineos: Euskara-Francés, Euskara-Español, Francés-Occitano y Español-Aragones
- ELHUYAR App de traducción automática entre las lenguas de los Pirineos: Euskara-Francés, Euskara-Español, Francés-Occitano y Español-Aragones
- EUSKALTZAINDIA Manual del Vasco Unificado: .Euskara Eskuz Esku Digitala",
- TURISMO DE ARAGÓN Diccionario Online del Aragonés.
- LO CONGRÉS Buscador Semántico Multilingüe.
- Medición de la Vitalidad del occitano, euskara y aragonés.

5 *Difusión de resultados*

Durante el año 2020 se llevarán a cabo actividades con el objeto de dar a conocer los resultasos a entidades, organizaciones, empresas, profesionales e investigadores del área POCTEFA y de otras regiones europeas, que trabajan en el ámbito de las lenguas y las tecnologías lingüísticas o que realizan aplicaciones multilingües para servicios de e-administración, e-salud, e-justicia, e-educación o e-cultura.

Aparte de varios talleres que cada socio organizará a nivel local, también se organizará un seminario Europeo sobre Tecnologías de la Lengua y contribución al Desarrollo Económico. Su objetivo será la ayuda práctica a otros agentes ajenos al proyecto para que puedan incorporar la tecnología obtenida. Esperamos firmar acuerdos Universidad–Empresa para la comercialización de soluciones.

La colaboración transfronteriza permitirá transferir conocimientos y desarrollar soluciones lingüísticas con potencial de mercado, que beneficien a profesionales de las lenguas y faciliten el acceso público multilingüe a contenidos. Estamos dando pasos para avanzar en el desarrollo de un clúster transfronterizo de tecnologías lingüísticas.

Reconocimientos

La investigación llevada a cabo en este proyecto se lleva a cabo como parte del proyecto "LINGUATEC: Desarrollo de la cooperación transfronteriza y transferencia de conocimiento en tecnologías de la lengua"(POCTEFA EFA227/16, FEDER), financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER).

Bibliografía

Alegria, I. y K. Sarasola. 2017. Language technology for language communities: An overview based on our experience. En *Çommunities in Control: Learning tools and strategies for multilingual endangered language communities", CinC 2017. October 19-21 2017, Alcanena / Portugal. Full corrected version: FEL XXI Alcanena 2017 Communinities in Control. Foundation for Endangered Languages, DIDLeS, SOAS World Languages Institute and Mercator Research Centre. pp. 91-97, ISBN: 978-0-9560210-9-0.*

Bernhard, D., A.-L. Ligozat, F. MARTIN, M. Bras, P. Magistry, M. Vergez-Couret, L. Steible, P. Erhart, N. Hathout, D. Huck, C. Rey, P. Reynés, S. Rosset, J. Sibille, y T. Lavergne. 2018. Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. En *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, Mayo.

Cortes, I., I. Leturia, I. Alegria, A. Astigarraga, y K. S. eta Manex Garaio. 2018. Massively multilingual accessible audioguides via cell phones. En *EAMT 2018 (Project/Product Track)*.

Etchegoyhen, T., E. Martinez, A. Azpeitia, I. Alegria, G. Labaka, A. Otegi, K. Sarasola, I. Cortes, A. Jauregi, I. Ellakuria, E. Calonge, y M. Martin. 2018. Quales: Estimación automática de calidad de traducción mediante aprendizaje automático supervisado y no-supervisado. *Procesamiento del Lenguaje Natural, vol. 61, pp. 143-146. ISSN: 1135-5948*.

Hernáez, I., E. Navas, I. Odriozola, K. Sarasola, A. Diaz de Ilarraza, I. Leturia, A. Diaz de Lezana, B. Oihartzabal, y J. Salaberria. 2012. *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 163-166 recibido 27-03-2019 revisado 06-05-2019 aceptado 15-05-2019

Prevención de enfermedades infecciosas basada en el análisis inteligente en RRSS y participación ciudadana

Prevention of infectious diseases based on intelligent analysis in social networks and citizen participation

Óscar Apolinario[1,2], José Medina-Moreira[1], Harry Luna-Aveiga[1], José Antonio García-Díaz[3], Rafael Valencia-García[3] , José Ignacio Estrade-Cabrera[4]
[1] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil, Ecuador
[2] VIAMATICA S.A., Edif. San Francisco 300, Córdova y Av. 9 de Octubre, 090313,Guayaquil, Ecuador
[3] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Murcia, España
[4] DANTIA Tecnología S.L., Parque Empresarial de Jerez 10, Calle de la Agricultura, 11407, Jerez de la Frontera, Cádiz, España
{oscar.apolinarioa, jose.medinamo.es@ug.edu.ec, harry.lunaa}@ug.edu.ec
{joseantonio.garcia8, valencia}@um.es
jiestrade@dantia.es, oapolinario@viamatica.com

Resumen: Este proyecto consiste en el desarrollo una plataforma inteligente de monitorización de enfermedades infecciosas a partir de la monitorización de redes sociales, fuentes de datos oficiales y la participación ciudadana. Esta información estará disponible para las autoridades sanitarias y gubernamentales a través de un panel de mandos personalizable para que puedan detectar zonas calientes en las que exista algún tipo de brote o focos de infección. También estará disponible un sistema de alertas para avisar a los ciudadanos cuándo se ha detectado cierto nivel de alarma en radio cercano a donde se encuentran. Este proyecto está siendo desarrollado por la empresa de Ecuador VIAMÁTICA, algunos docentes de la Universidad de Guayaquil, la empresa española DANTIA y las Universidades de Murcia y Carlos III de Madrid. La parte española del proyecto está financiada mediante una convocatoria de proyectos unilaterales del CDTI.
Palabras clave: Análisis de sentimientos, Infodemiología, Enfermedades Infecciosas, BlockChain

Abstract: This project consists in the development of an intelligent platform for the monitoring of infectious diseases based on: text written in natural language on social networks, official data sources and citizen participation. This information will be available to the health and governmental authorities through a customizable control panel so they can detect hot areas in which there is some type of outbreak. An alert system will also be available to notify citizens when a certain level of alarm is detected in a nearby radius of where they are. This project is being developed by VIAMÁTICA (Ecuador), the University of Guayaquil (Ecuador), DANTIA (Spain) and the Universities of Murcia and Carlos III of Madrid (Spain). The Spanish part of the project is financed through the unilateral CDTI projects call.
Keywords: Sentiment analysis, Infoveillance, Infectious Diseases, BlockChain

1 *Introducción*

Los brotes epidémicos son uno de los problemas más graves a los que se enfrenta la especie humana. En los últimos años se han producido cerca de cinco alertas sanitarias internacionales graves: el ZIKA, la Gripe Aviar, el

ISSN 1135-5948. DOI 10.26342/2019-63-21 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

virus del Ébola, la Gripe Tipo A y el SARS. Aunque ninguno de ellos tuvo una elevada mortalidad ni comprometió la existencia humana en su conjunto, si que supusieron un grave perjuicio para la sociedad en general y para las instituciones en particular; repercutiendo, además, negativamente en la economía. Además, el virus del Zika, muy presente en América Latina, tiene millones de casos infectados y miles de bebés están siendo afectados con trastornos neurológicos.

Para mitigar los efectos de estos brotes infecciosos, la Organización Mundial de la Salud recomienda diseñar estrategias de detección temprana. El diseño de estas estrategias requiere disponer de suficientes evidencias para predecir, con suficiente probabilidad, un brote epidémico. Aunque la mayoría de estos estudios se realiza a partir de datos clínicos recolectados en hospitales, la comunidad científica está explorando métodos fiables alternativos para la adquisición de información a través de Internet. Sin embargo, estos sistemas tienen todavía un amplio margen de mejora ya que todavía hacen interpretaciones inadecuadas de los datos (Choi et al., 2016).

El presente proyecto consiste en el desarrollo de un sistema de monitorización de enfermedades infecciosas que detecte zonas calientes en las que existe algún tipo de brote o foco de infección y comunique dicha información, de forma comprensible y aprovechable, tanto a los ciudadanos como a las autoridades sanitarias y gubernamentales.

A nivel técnico, esta plataforma extrae datos de tres categorías de fuentes de datos: 1) datos estructurados, a partir de la extracción del conocimiento en fuentes de datos oficiales publicadas en la web; 2) evidencias, a partir de la participación directa de la ciudadanía identificando y notificando directamente casos encontrados por ellos y, finalmente, a partir de 3) datos no estructurados, procedentes de textos escritos en lenguaje natural procedente de redes sociales públicas, tales como Twitter o Facebook.

Con la consecución de los objetivos del proyecto se pretende, por un lado, reducir los costes sociales y económicos derivados del tratamiento de estos brotes infecciosos y, por otro lado, mejorar la percepción de la ciudadanía en temas de salud pública transparentes y participativos, permitiendo el empoderamiento de la ciudadanía[1].

Figura 1: Arquitectura del sistema

2 *Arquitectura del sistema*

Esta plataforma está formado por cinco módulos: 1) Módulo de monitorización de redes sociales, 2) Módulo Crawler de extracción de datos oficiales, 3) Aplicación multidispositivo de participación ciudadana, 4) Módulo de sistema de filtrado y confianza basado en BlockChain y, 5) Módulo para la monitorización inteligente de enfermedades en tiempo real (ver Figura 1).

Una característica clave de esta plataforma es el desarrollo de algoritmos de evaluación de la confianza basados en la credibilidad de la fuente, con objeto de descartar falsos positivos y de priorizar hallazgos relevantes.

Una vez la información ha sido filtrada, es presentada a los usuarios finales mediante un panel de mandos configurable, donde pueden indicar que indicadores de rendimiento (KPIs) necesitan, con el fin de ayudarles a establecer medidas estratégicas acerca de cómo actuar ante posibles nuevos brotes o epidemias.

A continuación, se describen brevemente cada uno de los módulos de la plataforma haciendo mayor hincapié en los relacionados con tecnologías del lenguaje humano.

2.1 Módulo de monitorización de redes sociales

Este módulo se encarga de extraer información escrita en lenguaje natural en redes sociales. En pocas palabras, el sistema funciona de la siguiente manera: En primer lugar, se obtienen todos los mensajes que contienen palabras clave, además de otros indicadores relacionados, de la enfermedad objetivo.

[1] https://www.who.int/csr/labepidemiology/projects/earlywarnsystem/en/

En segundo lugar, se emplean técnicas de reconocimiento de entidades (Ruiz-Martínez et al., 2012) y técnicas de análisis de sentimientos basado en aspectos (Peñalver-Martínez et al., 2014) para obtener la información subjetiva relacionada con los tópicos que tratan, así como son los síntomas, las causas o los medicamentos. Para ello, se ha definido una ontología que describe el vocabulario principal de las enfermedades infecciosas, sus síntomas, posibles focos de infección, medicamentos, etc. Esta ontología será ampliada durante el desarrollo del proyecto incluyendo más información de enfermedades infecciosas ya que actualmente se ha centrado en las enfermedades infecciosas transmitidas por mosquito como dengué, malaria, zika y chuchugua.

Con respecto al análisis de sentimientos, se ha compilado hasta el momento un corpus balanceado inicial de 8.966 tuits y se han realizado pruebas con clasificadores binarios a partir de un modelo lingüístico basado en el análisis y extracción de variables psicolingüísticamente relevantes específica para textos escritos en castellano y en el español de América latina obteniendo unos buenos resultados iniciales (García-Díaz et al., 2018).

2.2 Módulo Crawler de extracción de datos oficiales

Este módulo se encarga de la obtención de datos estructurados a partir de un sistema de extracción de conocimiento (crawler). Para ello, se emplean técnicas de reconocimiento de entidades, extracción de términos y construcción automática de ontologías (Valencia-García et al., 2008) con el fin de extraer información relacionada con el dominio de las enfermedades infecciosas. Debido a que la información se encuentra en la red en distintos formatos, tales como documentos web, imágenes o documentos de procesadores de texto, es necesario el desarrollo de diversos procedimientos específicos capaces de extraer información para cada formato.

2.3 Aplicación multidispositivo de participación ciudadana

Este módulo consiste en una aplicación multidispositivo de participación ciudadana para la identificación activa de factores de riesgo relacionados con enfermedades infecciosas. Los ciudadanos también etiquetarán sus contribuciones y participación mediante etiquetas obtenidas de las ontologías del dominio.

Esta aplicación se ha diseñado con un enfoque basado en Aplicaciones Web Progresivas por dos motivos. En primer lugar, para permitir a los usuarios identificar evidencias en zonas sin conexión a Internet, realizando el envío de datos una vez que recupere la conexión. En segundo lugar, para aumentar el periodo de autonomía del dispositivo a través de minimizar el consumo de su batería (Malavolta et al., 2017). Además, esta aplicación incluye diversas técnicas de gamificación para fomentar la participación ciudadana a través de un sistema de reputación y recompensas.

2.4 Módulo de sistema de filtrado y confianza basado en BlockChain

Este módulo se encarga del valorar la calidad y la confianza de las distintas evidencias encontradas por los módulos anteriormente descritos. Al tratarse de información sensible, se ha optado por usar la tecnología de cadena de bloques (Blockchain) para garantizar la privacidad de la información médica recolectada. Esta aplicación de las cadenas de bloques ya ha sido estudiada en (Zyskind, Nathan, y others, 2015), donde los autores describen un sistema descentralizado encargado de manejar información personal sensible.

Para asignar un nivel de confiabilidad a los datos extraídos en redes sociales, se analizan características tales como si la cuenta del usuario ha sido validada o número de publicaciones relevantes realizadas. Por otro lado, para valorar las evidencias recogidas por los usuarios a través de la aplicación de participación ciudadana, se tienen en cuenta parámetros como el historial del usuario o la frecuencia de las mediciones.

2.5 Módulo para la monitorización inteligente de enfermedades en tiempo real

Este módulo se encarga de la gestión y monitorización de los parámetros relacionados con las enfermedades infecciosas a través de un proceso guiado por ontologías. En concreto, este sistema se encarga de cuatro apartados principales: 1) Configuración, que permite a los operarios del sistema crear áreas de interés, a partir de especificar un área geográfica y las enfermedades objetivo a monitorizar; 2) dashboard, que consiste en el desarrollo de un panel de mandos genérico y configurable donde los operarios podrán indicar y confi-

gurar los KPIs que estimen convenientes; 3) KPIs, que son indicadores de desempeño independientes capaces de leer datos de manera periódica y de mostrar la información en distintos formatos, tales como tablas o gráficas; 4) Sistema de alertas, capaz de configurar y notificar automáticamente a los usuarios suscritos en el momento en que los datos superen cierto umbral establecido para que se tomen medidas preventivas.

3 Trabajo futuro

El presente proyecto se encuentra todavía en una fase temprana y su desarrollo termina en 2020. Para cada uno de los módulos se están realizando estudios básicos para comprobar la viabilidad de ciertas operaciones, especialmente los relativos al módulo de módulo de sistema de filtrado y confianza basado en BlockChain, descrito en la 2.4sección 2.4.

Con respecto a las técnicas de análisis de sentimientos, se han llevado a cabo estudios a la hora de analizar la polaridad de textos en castellano y en español de México. Estos sistemas lingüísticos se han probado para clasificar el sentimiento a nivel general y se está trabajando en la extracción de conceptos a partir de una ontología para poder realizar un análisis más minucioso basado en aspectos.

Durante el tiempo que queda de desarrollo del proyecto está previsto la mejora de las tecnologías de extracción de conocimiento y minería de opiniones basados en aspectos. Por otro lado, se desarrollará la aplicación de participación ciudadana que permitirá etiquetar las contribuciones con vocabulario del dominio, además que se analizarán también sus contribuciones en texto.

Por último, se realizará una integración de los datos estructurados y no estructurados dentro del primer prototipo de la plataforma global. En este sentido, se planificarán distintas pruebas de campo para comprobar la calidad de las mediciones por parte de los usuarios en entornos sin conexión a Internet.

En la segunda anualidad también se desarrollarán las tecnologías de análisis inteligente de datos guiado por ontologías en el que se permitirá seleccionar distintos conceptos de las ontologías para así solamente realizar el análisis inteligente sobre los datos relacionados con esos conceptos.

Agradecimientos

Este trabajo está siendo financiado por el CDTI dentro del proyecto con referencia IDI-20180989 dentro de la convocatoria de proyectos unilaterales.

Bibliografía

Choi, J., Y. Cho, E. Shim, y H. Woo. 2016. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*, 16(1):1238.

García-Díaz, J. A., O. Apolinario-Arzube, J. Medina-Moreira, J. O. Salavarria-Melo, K. Lagos-Ortiz, H. Luna-Aveiga, y R. Valencia-García. 2018. Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. En *International Conference on Technologies and Innovation*, páginas 229–239. Springer.

Malavolta, I., G. Procaccianti, P. Noorland, y P. Vukmirović. 2017. Assessing the impact of service workers on the energy efficiency of progressive web apps. En *Proceedings of the 4th International Conference on Mobile Software Engineering and Systems*, páginas 35–45. IEEE Press.

Peñalver-Martínez, I., F. García-Sanchez, R. Valencia-García, M. A. Rodriguez-García, V. Moreno, A. Fraga, y J. L. Sánchez-Cervantes. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995–6008.

Ruiz-Martínez, J. M., R. Valencia-García, R. Martínez-Béjar, y A. Hoffmann. 2012. Bioontoverb: A top level ontology based framework to populate biomedical ontologies from texts. *Knowledge-Based Systems*, 36:68–80.

Valencia-García, R., J. T. Fernández-Breis, J. M. Ruiz-Martínez, F. García-Sánchez, y R. Martínez-Béjar. 2008. A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. *Expert Systems*, 25(3):314–334.

Zyskind, G., O. Nathan, y others. 2015. Decentralizing privacy: Using blockchain to protect personal data. En *2015 IEEE Security and Privacy Workshops*, páginas 180–184. IEEE.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 167-170 recibido 26-03-2019 revisado 03-05-2019 aceptado 15-05-2019

Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del Plan de Impulso de las Tecnologías del Lenguaje

Report on reusable documents as language resources in Spain, under the Government Plan for Language Technologies

Antonio Moreno[1], Doroteo Torre[1], Ana Valverde[1,3], Leonardo Campillos[1,2]
[1] Universidad Autónoma de Madrid
[2] LIMSI-CNRS (Orsay, Francia)
[3]Real Academia Nacional de Medicina
{antonio.msandoval, doroteo.torre, leonardo.campillos}@uam.es, avalverde@ranm.es

Resumen: Este estudio ha sido realizado dentro del ámbito del Plan de impulso de las Tecnologías del Lenguaje (Plan TL) con financiación de la Secretaría de Estado para el Avance Digital y Red.es. Los objetivos centrales son realizar un censado de recursos de las diferentes administraciones públicas que puedan ser convertidos en RL, así como proponer un plan de acción para abordar su conversión en RL. Se ha elaborado una metodología específica para el censado y evaluación de la madurez de los datos. Se han generado dos listados, uno preliminar compuesto por 101 recursos, del que se han seleccionado 24 para su análisis detallado y evaluación. El informe también incluye un repaso de estudios similares en otros países. Concluye con unas recomendaciones genéricas, así como estrategias concretas para los recursos seleccionados. El informe final y los listados están disponibles públicamente en Red.es y la página del Plan TL.

Palabras clave: recursos lingüísticos, Plan TL, administraciones públicas, datos abiertos, reutilización de información del sector público, RISP.

Abstract: This report was carried out within the Spanish administration-driven initiative Language Technologies Plan (Plan TL), funded by Secretaría de Estado para el Avance Digital and Red.es. The main goals are collecting from Spanish public administrations a listing of provided resources and open data that can be transformed to language resources, as well as proposing an action plan to process and distribute them. We designed a specific methodology for listing and evaluating the degree of maturity of the considered data. We created two listings: a preliminary collection of 101 resources, and 24 resources and data repositories selected from the first list for a detailed analysis and evaluation. This report also features a comparative analysis of similar initiatives and studies conducted abroad. We conclude with generic recommendations and detailed strategies for the selected resources. The report and listings are publicly available at Red.es and the Plan TL. website.
Keywords: language resources, Plan TL, Spanish public administrations, Open Data, Public Sector Information, PSI.

1 *Contexto del estudio*

El Plan TL tiene como objetivo general desarrollar la industria del Procesamiento del Lenguaje Natural (PLN), la Traducción Automática (TA) y los sistemas conversacionales (SSCC) en España, y especialmente, en lengua española y lenguas cooficiales, aprovechando, además, estas novedosas tecnologías para mejorar el servicio público.

En el marco de este Plan, ya se han realizado otros informes (Bel y Rigau 2015; Soroa et al. 2017) sobre el estado de las Tecnologías del

ISSN 1135-5948. DOI 10.26342/2019-63-22 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Lenguaje (TL) en España. El objetivo del presente estudio es localizar conjuntos de datos o documentos de distintos organismos públicos que resulten de interés desde el punto de vista de las TL y que sean, además, factibles para su conversión en RL.

2 Metodología empleada

Para la elaboración y evaluación del censo de recursos, se generó una ficha técnica para la recogida de información, compuesta por diferentes campos (Tabla 1)

1. **Identificación del recurso:** nombre, url, clasificación, lenguas, licencia, etc.
2. **Persona de contacto u organización responsable**
3. **Creación del recurso:** proveedor, proyecto financiador.
4. **Descripción del recurso:** variedad de lengua, niveles de anotación, estándares, tamaño, unidad, formato, dominio, etc.
5. **Otros recursos relacionados.**
6. **Grado de madurez de datos conforme al modelo de la tabla 2.**
7. **Posibles aplicaciones del futuro recurso lingüístico.**

Tabla 1: Estructura y principales contenidos de la ficha técnica

Adicionalmente, se elaboró una ficha específica para la evaluación de la madurez (Tabla 2), teniendo en cuenta factores técnicos requeridos por los RL usados en PLN y aspectos legales para su (re)utilización.

3 Censado de recursos

3.1 Listado de documentos analizados por dominio

Los 101 recursos recogidos en la primera búsqueda se clasificaron en cuatro áreas temáticas que son de interés prioritario para el Plan TL:

- **Sanidad: 41** (textos, documentos multimedia, entidades nombradas, memorias bilingües, o terminología, de ámbito español o latinoamericano).
- **Justicia: 5** (textos y documentos multimedia de ámbito español).
- **Inteligencia competitiva: 7** (entidades nombradas, terminología y corpus paralelos de ámbito español o latinoamericano).
- **Cultura, Turismo y otros: 35** (textos, documentos multimedia, entidades nombradas y corpus paralelos de ámbito español).

Puntos que considerar para evaluar el grado de madurez	**Valores: 0, 1, 2**	**Observaciones**
Aspectos técnicos:		
1. Digitalización (conversión a formato procesable).		
2. Transcripción (ortográfica, fonológica, …).		
3. Alineación vídeo/sonido y texto.		
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1).		
5. Anotación morfológica y/o sintáctica.		
6. Anotación de entidades nombradas.		
7. Otros tipos de anotación (semántica, pragmática, palabras clave, …).		
8. Revisión de aspectos formales (ortografía, formato de etiquetado, …).		
9. Revisión de contenido (incoherencias, redundancia de datos, …).		
10. Anotación conforme a estándares PLN.		
11. Presencia de metadatos.		
Aspectos legales:		
12. Necesidad de anonimización de datos personales.		
13. Necesidad de solicitud de permiso de uso.		
TOTAL		

Tabla 2: Plantilla para la evaluación de la madurez como RL de un recurso

De este primer censo, se seleccionaron 24 recursos para su análisis exhaustivo, según los criterios de interés (calidad, cantidad y disponibilidad de los datos), plurilingüismo[1], estado de la propiedad intelectual, variedad temática, grado de madurez y tipología del RL (ver Tabla 3).

4 Conclusiones preliminares sobre la madurez

Como era de esperar, **la mayoría de los conjuntos de datos analizados** en este informe se quedan en los estadios de **madurez baja** o **media**. Esto es comprensible y esperable, dado

[1] Se han considerado únicamente recursos en las cuatro lenguas cooficiales del Estado.

que los requisitos para ser considerado un recurso *maduro* son muy estrictos: solo los ya procesados y en formatos directamente usables por los investigadores de PLN (por ejemplo, XML o TMX) pueden ser considerados propiamente RL.

Listado de Recursos Analizados
1. Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM).
2. Patentes multilingües digitalizadas en PATSTAT de European Patent Office (EPO).
3. Diccionarios terminológicos del Centro de Terminología (TERMCAT).
4. Padrón: Relación de municipios del Instituto Nacional de Estadística.
5. Topónimos del Instituto Geográfico Nacional (IGN).
6. Grabaciones de vídeo de RTVE a la carta.
7. Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España.
8. Índices de clasificación de los catálogos de la BNE.
9. Publicaciones periódicas digitalizadas de la Hemeroteca Digital.
10. Documentos digitalizados de la Biblioteca Digital Hispánica.
11. Publicaciones en repositorio SciELO (Scientific Electronic Library Online).
12. Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII).
13. Banco de datos de enfermedades raras y medicamentos huérfanos de OrphaData.
14. Guías de práctica clínica (GPC) del portal Guía Salud.
15. Vídeos del portal web de TV del Gobierno Vasco relacionados con el tema de Salud.
16. Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).
17. Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA).
18. Textos de Jurisprudencia del CENDOJ.
19. Textos del Boletín Oficial del Estado (BOE) Diario.
20. Textos de códigos electrónicos del Boletín Oficial del Estado (BOE).
21. Textos sobre Legislación del Boletín Oficial del Estado (BOE).
22. Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP).
23. Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.
24. Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial.

Tabla 3: Recursos analizados en el estudio

5 Estudios e iniciativas similares en otros países

El estudio contempla iniciativas de creación de RL en países hispanoamericanos y en Europa, con algunas incursiones en EEUU y Canadá. Otros países destacables (China, Japón o India) quedaron fuera de este análisis.

Un plan de acción similar para la reutilización de datos públicos o de contenidos de páginas webs en administraciones públicas para desarrollo de RL es el portal de la European Language Resource Coordination,[2] que proporciona una lista de los recursos existentes en Europa de carácter abierto.

España se encuentra entre los países de cabeza en cuanto a RL en formato libre y disponible para TL. A nivel europeo, Francia y Reino Unido son los referentes, si bien la Comisión Europea ha advertido del retraso con respecto al inglés y a las compañías americanas, que han *capitalizado* el uso de los *Big Data* lingüísticos, permitiendo que las grandes multinacionales tecnológicas como Google, Microsoft, Amazon o IBM ofrezcan servicios lingüísticos en muy diferentes dominios y lenguas.

Pese a la apertura y disponibilidad aparente de recursos, es importante destacar que **la comunidad investigadora en PLN en España necesita RL de calidad y en cantidad suficiente para desarrollar aplicaciones competitivas en el mercado internacional**, con la amenaza potencial de que otras empresas de fuera de nuestras fronteras ocupen el espacio del PLN en las lenguas del Estado.

De ahí la necesidad de que la conversión a RL siga estándares técnicos internacionales para garantizar la interoperabilidad entre datos y procesadores (por ejemplo, los requisitos definidos por CLARIN).

6 Recomendaciones

Otro de los objetivos del estudio es describir un plan de acción a corto y medio plazo, con el fin de establecer prioridades.

6.1 Recomendaciones genéricas

Se aplican a todos los tipos de RL:

1. Garantizar la disponibilidad y el acceso universal a los datos abiertos para RL en todas las lenguas del Estado a través de un portal común y único.
2. Mejorar la visibilidad de los conjuntos de datos en cuanto a su disponibilidad y madurez.

[2] www.lr-coordination.eu

3. Facilitar la descarga masiva de grandes ficheros en formatos apropiados (texto plano, XML, CSV, JSON, RDF).
4. Impulsar la conversión progresiva de los millones de páginas digitalizadas en PDF o imagen en texto plano: los documentos en PDF, EPUB o las imágenes digitalizadas a formato TXT, CSV o XML.
5. Proporcionar la transcripción de ficheros multimedia. Las cadenas públicas de radio y televisión y las grabaciones de vistas orales de los juicios son una valiosa fuente de datos para distintos ámbitos de procesamiento de habla.
6. Estimular la adopción de licencias de libre uso y acceso a los datos. El empleo de licencias de tipo Creative Commons resulta muy práctico y, en términos generales, agiliza la obtención de los RL.
7. Estimular la reutilización de datos organizando competiciones tecnológicas.
8. Facilitar el acceso a capacidad de cómputo y almacenamiento, ya que la evolución de la tecnología en los ámbitos del PLN se apoya, cada vez más, en estos aspectos.

6.2 Estrategias concretas para recursos seleccionados

El informe destaca una serie de conjuntos como líneas prioritarias en el Plan de Acción:

1. Los **recursos digitales de la Biblioteca Nacional de España (BNE)**: tanto la Hemeroteca Digital como la Biblioteca Digital Hispánica poseen conjuntos de datos de enorme interés lingüístico y cultural, de gran variedad temática, temporal y geográfica. Además, están ya digitalizados en PDF, son de acceso libre, no tienen impedimentos legales de copyright, y su uso no requiere la solicitud expresa de un permiso escrito.
2. **Archivo de RTVE** y **RTVE a la carta**: datos (en formato vídeo y audio) de gran interés lingüístico, cultural y tecnológico, con gran variedad temática. Contienen cerca de 100.000 horas de televisión de alta calidad, con más de 10.000 informativos.
3. **Patentes del ámbito iberoamericano** o registradas en la **OEPM** y **PATSTAT**. Si bien el grado de madurez de estos recursos (patentes, modelos de utilidad e informes técnicos) puede ser mejorado para las tareas de PLN, el potencial y el interés es enorme.
4. **Recursos de dominio médico**: destacan algunos recursos de un grado alto de madurez (los bancos de datos de OrphaData y el Nomenclátor de prescripción del CIMA), con un modelo de licencia, distribución y riqueza de anotaciones. Habría un proceso de conversión más costoso para las publicaciones científicas de ISCIII o la AEMPS (junto con sus alertas médicas y farmacológicas), o las guías de práctica clínica del portal GuíaSalud.

7 Conclusiones

Los patrimonios documental, jurídico, cultural y lingüístico del país están muy exhaustivamente representados en diferentes conjuntos de datos abiertos y la comunidad I+D está esperando con fundadas expectativas que se conviertan en RL. Su disponibilidad debería permitir a la industria española de TL poder competir con las empresas internacionales en un Mercado Único Digital basado en el tratamiento multilingüe. De esta manera, España se convertiría en un referente en TL a nivel europeo junto a Francia y Reino Unido, si se potencia el desarrollo y las infraestructuras de RL.

Agradecimientos

Este informe ha sido financiado por la Secretaría de Estado para el Avance Digital (SEAD) y Red.es. Agradecemos las observaciones realizadas por David Pérez, Sonia Castro, Pilar Polo, Inés Rodríguez y Doaa Samy.

Bibliografía

Bel, N. y G. Rigau, (eds.) 2015. *Informe sobre el estado de las Tecnologías del Lenguaje en España*. Accesible en https://www.plantl.gob.es

Soroa, A., G. Rigau, J. Porta, J. Atserias, y X. Gómez Guinovart. 2017. *Plataformas y sistemas de procesamiento lingüístico de alto rendimiento*. Accesible en https://www.plantl.gob.es

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 171-174 recibido 30-03-2019 revisado 25-04-2019 aceptado 14-05-2019

EXTRAE: EXTRacción de Asociaciones entre Enfermedades y otros conceptos médicos

EXTRAE: EXTRaction of Associations between Diseases and Other Medical Concepts

Lourdes Araujo[1], Juan Martinez-Romo[1], Andrés Duque[1], Fernando López-Ostenero[1], Ricardo Sanchez de Madariaga[2], Adolfo Muñoz Carrero[2], Mario Pascual Carrasco[2]
[1]Univ. Nacional de Educación a Distancia (UNED), 28040 Madrid
[2]Inst. de Salud Carlos III (ISCIII) 28029 Madrid
lurdes,juaner@lsi.uned.esaduque@scc.uned.es,fernando@lsi.uned.es
ricardo.sanchez@isciii.es adolfo.munoz@isciii.es mario.pascual@isciii.es

Resumen: El propósito de este proyecto es investigar en la mejora de las técnicas de extracción de Reglas de Asociación (RA) entre enfermedades, o entre enfermedades y otros conceptos médicos. Estas reglas permiten representar el conocimiento médico subyacente a un conjunto de Historias Clínica Electrónica (HCE). Concretamente nos planteamos explorar técnicas semisupervisadas que nos permitan alcanzar resultados equiparables a los de las técnicas supervisadas con una mínima supervisión. El proyecto se propone realizar avances significativos en la selección de reglas de asociación relevantes en el dominio de la salud, que pueden tener una alta aplicabilidad en la ayuda al diagnostico y en la prevención de enfermedades.
Palabras clave: dominio médico, extracción de información, reglas de asociación

Abstract: This project aims to improve the techniques for extracting Association Rules (AR) between diseases, or between diseases and other medical concepts. These rules allow the representation of medical knowledge underlying a set of Electronic Medical Records (EHR). Particularly, we plan to explore semi-supervised techniques that allow us to achieve similar results to those obtained using supervised techniques, but requiring minimal supervision. The project intends to make significant progress in the selection of relevant AR, which may be applied in the health domain for developing diagnostic help systems, or for disease prevention.
Keywords: medical domain, information extraction, association rules

1 Introducción

Los profesionales de la salud disponen en la actualidad de acceso a la Historia Clínica Electrónica (HCE) de los pacientes. La disponibilidad de información precisa, completa y estructurada permite mejorar considerablemente la toma de decisiones. Sin embargo, cada vez es más difícil tomar estas decisiones dado el gran volumen de datos que ha de considerarse. Este volumen dificulta encontrar manualmente relaciones que pueden ser utilizadas en la extracción de conocimiento. En este proyecto nos proponemos diseñar algoritmos que ayuden a la identificación de relaciones relevantes entre distintas enfermedades. Esta información es muy útil para realizar nuevos diagnósticos, probar nuevos tratamientos o fármacos, para prever la posible evolución de la enfermedad, etc. En la actualidad los médicos tienen que basarse en su experiencia para encontrar estas relaciones. El problema se hace prácticamente intratable cuando el especialista quiere abordar no sólo su área de especialización, sino también otras. Por esta razón sería muy útil disponer de un sistema que realice una preselección de las relaciones entre enfermedades y se las proponga a los especialistas en salud, para su consideración. Muchas enfermedades comparten uno, o varios aspectos, como síntomas, evolución, tratamiento, etc., pero esto no siempre significa que exista una relación entre ellas. Por ello, lo que propo-

ISSN 1135-5948. DOI 10.26342/2019-63-23 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

nemos es un sistema capaz de detectar relaciones entre enfermedades que se pueden considerar significativas. La significatividad vendrá dada por la coincidencia de aspectos más allá de la casualidad que se capturará definiendo un modelo estadístico apropiado. Las relaciones entre distintas enfermedades se pueden establecer en base a distintos patrones, separada o conjuntamente: aparición conjunta, síntomas comunes, similitudes de tratamientos, etc. Estas relaciones entre enfermedades se pueden codificar como Reglas de Asociación (RA), que se pueden considerar formas de representar el conocimiento médico subyacente en el conjunto de HCE almacenadas en el repositorio de información clínica. Sin embargo, la extracción de RA no resulta un proceso evidente o inmediato a partir del repositorio de HCE, por dos razones fundamentales. Primero, como se ha apuntado antes, la producción de relaciones entre enfermedades o RA por parte de los algoritmos de Aprendizaje Máquina (AM) no significa que estas relaciones sean significativas, pues pueden haber sido producidas por el ruido o la casualidad. Por otra parte, el conocimiento médico codificado en forma de RA debe ser ratificado por la experiencia empírica de los profesionales médicos. Este es un proceso extremadamente largo, complejo y costoso, si se tiene en cuenta que estos profesionales disponen de recursos de tiempo y esfuerzo limitados. Por lo tanto, es imprescindible la proposición a la comunidad científica médica de RA que tengan la máxima probabilidad a priori de ser validadas y añadidas al corpus de conocimiento médico establecido. En este proyecto estamos diseñando algoritmos semisupervisados (con la mínima intervención médica inicial posible, ya que este es un recurso muy costoso) para producir listas de RA plausibles ordenadas por la máxima probabilidad a priori de ser ciertas, de forma que a los profesionales médicos se les presenten únicamente las nuevas RA mejor colocadas en esta ordenación óptima, y así también se optimice el uso de un recurso tan preciado como el tiempo y el esfuerzo de estos profesionales.

2 Grupos Involucrados

Los grupos implicados en el proyecto son:

- Grupo NLP&IR[1] de la Universidad Nacional de Educación a Distancia (UNED). Dispone de una amplia experiencia en Acceso Inteligente a la Información y Adquisición y Representación de Conocimiento Léxico, Gramatical y semántico. Tiene una amplia trayectoria en la realización de proyectos de investigación.
- Grupo del Instituto de Salud Carlos III (ISCIII), con amplia experiencia en aprendizaje máquina de HCE estandarizada (norma UNE-EN ISO 13606), participación en proyectos de investigación y publicaciones científicas.

El grupo del Instituto de Salud Carlos III cuenta con datos de historiales clínicos (HCE) anonimizados con los que trabajar que suponen una excelente fuente de información para poder abordar el proyecto. Los extractos de HCE normalizados han sido transferidos por cinco hospitales españoles que han colaborado o colaboran en la actualidad con el ISCIII. Estos extractos han sido normalizados según el estándar UNE-EN ISO 13606 usando arquetipos diseñados por profesionales médicos de esos hospitales. Esta información médica está estructurada y sistematizada según el conocimiento médico de esos arquetipos. Esta fuerte estructuración y organización de la información médica la hace óptima para extraer de ella las reglas de asociación obtenidas por los algoritmos semisupervisados de AM. Este grupo cuenta así mismo con experiencia en la definición de reglas de asociación básicas.

Por su parte el grupo de la UNED cuenta con experiencia en métodos de aprendizaje automático que han aplicado en distintos problemas, y en particular en el dominio biomédico. Cuentan así mismo con experiencia en técnicas estadísticas que permitirán asignar probabilidades y establecer un ranking en las reglas de asociación que se quieren generar de forma automática.

3 Antecedentes

Uno de los modelos más estudiados en el ámbito de la minería de datos es el de las reglas de asociación (Agrawal, Imieliński, y Swami, 1993; Witten, Frank, y Hall, 2011). Este modelo se utiliza habitualmente para descubrir hechos que ocurren de forma común dentro de un mismo conjunto de datos. Por ejemplo, Witten, Frank, y Hall (2011) emplearon las reglas de asociación para descu-

[1] http://nlp.uned.es/

brir las relaciones entre los datos recopilados a gran escala en los sistemas de terminales de punto de venta (TPV) de unos supermercados. De esta forma, se extraían relaciones que indicaban por ejemplo, que si un usuario compraba cebollas y verduras al mismo tiempo, era muy probable que la carne también estuviera incluida en su cesta de la compra. Estas reglas de asociación representan las relaciones ocultas entre los datos de un conjunto. Estas son implicaciones del tipo $X \rightarrow Y$, donde X e Y son conjuntos de ítems disjuntos. Dentro de las reglas de asociación se manejan dos medidas que determinan la intensidad de la relación: el soporte y la confianza. El soporte de una regla se define como la frecuencia de dicha regla dado un conjunto de ítems, mientras que la confianza mide la frecuencia con la que los ítems del conjunto Y aparecen en las transacciones que contienen X. Estas dos medidas se utilizan con el fin de minimizar el hecho de encontrar patrones por casualidad. Para que la regla sea considerada fuerte, debe sobrepasar los umbrales establecidos de soporte y confianza, determinados por el experto para el conjunto de datos. Uno de los primeros trabajos (Stilou et al., 2001) que emplearon este tipo de reglas lo hicieron sobre una base de datos médica con información de pacientes diabéticos con el objetivo de identificar nuevos patrones hasta entonces no contemplados. Los autores consideraron los resultados como prometedores y sentaron las bases de la investigación de estos métodos en medicina.

Existen trabajos como el de Rashid, Hoque, y Sattar (2014) en el que tratan de encontrar co-ocurrencias entre las enfermedades padecidas por un paciente analizando un repositorio de informes clínicos. El sistema creado era capaz de predecir las co-relaciones entre enfermedades primarias (aquellas por las que el paciente visita al doctor) y secundarias, que son otras enfermedades asociadas a las primarias y padecidas por el mismo paciente. En la misma línea se encuentra el trabajo de Kang'ethe y Wagacha (2014) en el que sobre una base de 98.000 informes clínicos de pacientes de Estados Unidos y usando las reglas de asociación obtuvieron un nivel mínimo de confianza del 56-76 % y una co-ocurrencia del 90 % en la búsqueda de patrones de comorbilidad. El autor también uso la codificación ICD-9, algo que ayudó en la tarea de limitar el conjunto de posibles grupos de diagnóstico. La validación de reglas de asociación por profesionales médicos es un problema complejo y prolongado en el tiempo que requiere una cuidada selección de las potenciales reglas candidatas. Esto se debe al alto coste del tiempo y del esfuerzo de estos profesionales y de los estudios y ensayos clínicos. Por lo tanto es indispensable la optimización de los resultados producidos por los algoritmos de aprendizaje máquina aplicados a los repositorios de HCE normalizada. Para ello estamos explorando algoritmos de aprendizaje semi-supervisados, de forma que la intervención previa de los profesionales sanitarios sea minimizada. Estos algoritmos serán a su vez optimizados para que las listas de reglas de asociación que produzcan estén ordenadas según las probabilidades mayores de ser aceptadas por la comunidad médica internacional.

4 Objetivos

El principal objetivo del proyecto es la extracción de relaciones entre enfermedades que puedan ser inducidas a partir de los datos recogidos en informes médicos (HCE estandarizada). Estás relaciones deben estar ponderadas con probabilidades de forma que puedan indicar a los profesionales de la salud si existe un riesgo alto de observar una enfermedad futura en un paciente. Los objetivos específicos del proyecto son:

- Preparación de los datos de trabajo. Esto incluye la preparación de los textos, eliminando erratas, mayúsculas, etc. Se incluye también la generación de reglas de asociación utilizando algoritmos clásicos del área, como FP-growth (Han, Pei, y Yin, 2000), para generar las reglas candidatas. Parte de estas reglas candidatas han sido evaluadas manualmente por un doctor.
- Diseño e implementación de técnicas de aprendizaje automático semi-supervisadas para la extracción de relaciones entre enfermedades. Se incluye aquí la exploración de algoritmos no supervisados y supervisados que nos guíen en el diseño del algoritmo semisupervisado.
- Evaluación, comparativa y optimización de los resultados obtenidos con distintas técnicas.

- Obtención, evaluación y validación de resultados.

5 Situación actual

El proyecto, de dos años de duración, se encuentra ya en un estado avanzado de desarrollo. Contando con la colaboración de un doctor, se han anotado manualmente 1000 reglas de asociación, indicando si son ciertas o falsas. Estas reglas han permitido evaluar sistemas basadas en distintos enfoques. Se ha diseñado un sistema no supervisado basado en la significatividad estadística de las reglas. Se han construido también sistemas supervisados utilizando las reglas anotadas manualmente para entrenar. Finalmente se está diseñando un sistema semi-supervisado combinando propiedades del sistema no supervisado y del supervisado.

Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos EXTRAE (IMIENS 2017) y PROSA-MED (TIN2016-77820-C3-2-R).

Bibliografía

Agrawal, R., T. Imieliński, y A. Swami. 1993. Mining association rules between sets of items in large databases. En *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, páginas 207–216, New York, NY, USA. ACM.

Han, J., J. Pei, y Y. Yin. 2000. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, Mayo.

Kang'ethe, S. M. y P. W. Wagacha. 2014. Extracting diagnosis patterns in electronic medical records using association rule mining. *International Journal of Computer Applications*, 108(15):19–26, December. Full text available.

Rashid, M. A., M. T. Hoque, y A. Sattar. 2014. Association rules mining based clinical observations. *CoRR*, abs/1401.2571.

Stilou, S., P. D. Bamidis, N. Maglaveras, y C. Pappas. 2001. Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. En *MEDINFO 2001 - Proceedings of the 10th World Congress on Medical Informatics, September 2-5, 2001, London, UK*, páginas 1399–1403.

Witten, I. H., E. Frank, y M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edición.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 175-178 recibido 17-03-2019 revisado 17-04-2019 aceptado 14-05-2019

Enriched communication across the lifespan

Comunicación enriquecida a lo largo de la vida

Martin Cooke[1], **Simon King**[2], **Valerie Hazan**[3], **Yannis Stylianou**[4], **Esther Janse**[5], **Deniz Baskent**[6], **Volker Hohmann**[7], **Axel Winneke**[8], **Inma Hernaez**[9]
[1]Ikerbasque, Basque Science Foundation, Spain
[2]University of Edinburgh, UK
[3]University College London, UK
[4]University of Crete, Greece
[5]Radboud University Nijmegen, The Netherlands
[6]University Medical Center Groningen, The Netherlands
[7]Hörzentrum GmbH, Germany
[8]Fraunhofer Institute for Digital Media Technology, Germany
[9]Universidad del País Vasco, Spain
m.cooke@ikerbasque.org

Abstract: Speech is a hugely efficient means of communication: a reduced capacity in listening or speaking creates a significant barrier to social inclusion at all points through the lifespan, in education, work and at home. Hearing devices and speech synthesis can help address this reduced capacity but their use imposes greater listener effort. The goal of the EU-funded ENRICH project is to modify or augment speech with additional information to make it easier to process. Enrichment reduces listening burden by minimising cognitive load, while maintaining or improving intelligibility. ENRICH investigates the relationship between cognitive effort and natural and synthetic speech. Non-intrusive metrics for listening effort will be developed and used to design modification techniques which result in low-burden speech. The value of various enrichment approaches will be evaluated with individuals and cohorts with typically sub-optimal communication ability, such as children, hearing- or speech-impaired adults, non-native listeners and individuals engaged in simultaneous tasks.
Keywords: Intelligibility, listening effort

Resumen: El habla es un medio de comunicación sumamente eficiente: la capacidad reducida para oir o hablar crea barreras importantes para la inclusión social a lo largo de la vida en la educación, en el trabajo o en el hogar. Los audífonos y la síntesis del habla pueden ayudar a abordar esta capacidad reducida, pero su uso impone un mayor esfuerzo por parte del oyente. El objetivo del proyecto europeo ENRICH es modificar o aumentar el habla con información adicional resultando así más fácil de procesar. El enriquecimiento reduce el esfuerzo de escucha minimizando la carga cognitiva, mientras se mantiene o mejora la inteligibilidad. ENRICH investigará la relación entre el esfuerzo cognitivo y las diferentes formas de habla natural y sintética. Se desarrollarán métricas no intrusivas para el esfuerzo de escucha que se utilizarán para diseñar modificaciones que resulten en un habla de baja carga. El valor de los diversos enfoques de enriquecimiento se evaluará con individuos y cohortes con habilidades de comunicación típicamente subóptimas, como niños, adultos con problemas de audición o de habla, oyentes no nativos e individuos que realizan tareas simultáneas.
Palabras clave: Inteligibilidad, esfuerzo de escucha

ISSN 1135-5948. DOI 10.26342/2019-63-24 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Speech is highly variable in terms of its clarity and intelligibility. Conversational speech is more difficult to understand than formal speech (e.g., broadcast speech) because speakers tend to reduce many sounds in casual conversations. Furthermore, speech may be difficult to understand when it is produced by non-native speakers or young children, or when it is produced in a very noisy or reverberant environment. Much of the focus in robust speech research has been on developing speech processing techniques aimed at improving speech intelligibility in challenging conditions and many advances have been made in this regard. For example, it is possible to improve the intelligibility of speech heard in noisy conditions via modifications carried out on speech before it is transmitted to a listener's ears (Sauert and Vary, 2006; Skowronski and Harris, 2006; Taal, Jensen, and Leijon, 2013; Zorila, Kandia, and Stylianou, 2012; Cooke et al., 2013; Godoy and Stylianou, 2013; Tang, Fazenda, and Cox, 2018). Other research has shown that manipulating the speech produced by adults with dysarthria can improve its intelligibility (Kain et al., 2007).

Because of these advances, improving intelligibility *per se* is no longer a sufficient goal. Understanding less-than-optimal speech also imposes a high cognitive load (Rönnberg et al., 2010; Pals, Sarampalis, and Baskent, 2013; Hornsby, 2013), and the resulting increase in listener effort may have a significant cost, in terms of comprehension lagging behind the speech, and in terms of its effect on an individual's well-being and ability to multitask successfully (e.g., listening to speech while driving). Less-than-optimal speech can be expected to be particularly troublesome for those listening in a non-native language, or with immature auditory-cognitive systems, or early hearing loss, or older adults facing age-related cognitive decline, or users who rely on hearing prostheses. Nearly all listeners fall into one or more of these categories at various points during their lifetime.

Given the success of speech modification in improving speech intelligibility in noise, the extent to which speech can be further transformed to increase its effectiveness in reducing cognitive load is an open question. Enrichment denotes both modifications to the speech signal itself as well as augmentation, for instance by embedding additional contextual information such as text or visual cues into the speech signal, to be decoded at the receiver.

The goal of ENRICH, a Marie Skłodowska-Curie European Training Network, is to equip researchers with the capacity to produce innovative advances in the new field of speech enrichment. This pursuit is naturally inter-disciplinary, involving cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups, and will require future researchers who possess a broad range of skills as well as an appreciation of business practices and the needs of end-users. ENRICH members represent complementary academic disciplines with strong links to hospitals and schools, as well as businesses focused on the development of assistive technologies in speech and hearing.

ENRICH addresses the unmet need for multi-skilled practitioners and engineers in this rapidly growing sector currently facing a serious workforce shortage. Through a comprehensive training programme driven by the needs of industry and clinical practice, ENRICH is equipping 14 early-stage researchers (predoctoral fellows) with not just the necessary cross-disciplinary knowledge and research techniques, but also with experience of entrepreneurship and technology transfer so they can translate research findings into meaningful products and services that will facilitate spoken language communication in the coming decades.

2 Objectives

1. To provide world-class research training in a multidisciplinary/multi-sectorial setting
2. To generate new scientific findings in speech perception and production
3. To develop novel algorithms for speech enrichment
4. To demonstrate the benefit of enrichment on cognitive effort for listener groups and individuals
5. To pilot the application of speech enrichment in assistive technology

3 Partners

The ENRICH consortium consists of 16 beneficiaries and partners from academia, industry and clinical practice, collectively providing infrastructure for investigating spoken communication and for applying innovations to end-user populations.

Beneficiaries:

- University of the Basque Country (UPV-EHU), Spain (Coordinator)
- University of Edinburgh, UK
- University College London (UCL), UK
- Radboud University Nijmegen, Netherlands
- University of Crete, Greece
- University Medical Center Groningen, Netherlands
- Horzentrum, Germany
- Fraunhofer Institute for Digital Media Technology, Germany

Partners:

- Voxygen, France
- Toshiba, UK
- Basque Center for Cognition, Brain and Language, Spain
- FORTH, Greece
- Hacettepe University, Turkey
- Sonova, Switzerland
- Chania General Hospital, Greece
- University of Oldenburg, Germany

The consortium provides expertise in computational modelling of human speech perception in noise; metrics for intelligibility prediction; speech synthesis; casual and formal speech differences; relating variability in language performance across the adult life span to individual characteristics such as cognitive and hearing abilities; speech signal processing and modification; speech perception with hearing impairment and hearing devices, in relation to cognitive effort; commercial hearing-aid evaluation.

4 Research themes

The proposed work programme revolves around three main themes (Fig. 1).

Theme 1 explores the dimensions of cognitive effort in listening, and its relation with different forms of speech, natural and artificial. Building on this understanding, non-intrusive metrics for effort based on information in the speech signal are being developed

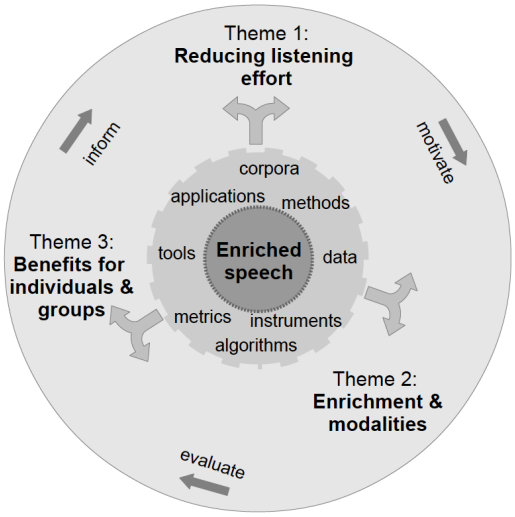


Figure 1: ENRICH research themes

to inform the design of effort-reducing enrichment techniques.

Theme 2 examines practical aspects of speech enrichment and explores issues such as how the speech signal can be augmented to contain further information useful to certain classes of listeners, or devices used by the hearing impaired. It also examines what forms of enrichment are suitable to different communication modes (e.g., passive listening, remote two- or multiparty conversations, near- and far-field broadcast).

Variability in the response of different individuals or specific populations to distinct forms of speech provides the setting for theme 3. This theme investigates the impact of specific enrichments on individuals and groups in order to better understand perceptual and cognitive processes. Key questions covered in this theme include: What can we learn from highly-trained listeners such as musicians or interpreters? Can enrichment ameliorate the consequences of gradual hearing deterioration over the long-term? To what extent does speech enrichment transfer across languages?

Theme 1 provides data and ideas to support the development of approaches to speech enrichment in theme 2. The value of proposed techniques for different classes of listeners is being evaluated in theme 3, which subsequently informs the design of further behavioural studies in theme 1. ENRICH benefits from network-wide speech data, experimental methods, instrumentation, tools and knowledge. This common infrastructure is being augmented with novel metrics for measuring

cognitive effort, algorithms for speech modification, evaluation frameworks and real-life technology applications.

5 Selected highlights

Studies to date have demonstrated that poor quality synthetic speech is more effortful to process (Govender and King, 2018) and that existing intelligibility enhancement techniques can also reduce cognitive load (Simantiraki, Cooke, and King, 2018). Intelligibility enhancing modifications can also be learnt by example (Shifas, Tsiaras, and Stylianou, 2018) or these can also occur on the perceptual side, for example, due to music training (Kaplan, Wagner, and Baskent, 2018).

More information about ENRICH is available at `http://www.enrich-etn.eu` and via the ENRICH video `https://youtu.be/_2W52Y3IE_Y`.

Acknowledgements

ENRICH has received funding from the EU H2020 research and innovation programme under MSCA GA 675324.

Bibliografía

Cooke, M., C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang. 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55:572–585.

Godoy, E. and Y. Stylianou. 2013. Increasing speech intelligibility via spectral shaping with frequency warping and dynamic range compression plus transient enhancement. In *Proc. Interspeech*, pages 3572–3576.

Govender, A. and S. King. 2018. Using Pupillometry to Measure the Cognitive Load of Synthetic Speech. In *Proc. Interspeech 2018*, pages 2838–2842.

Hornsby, B. 2013. The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear Hearing*, 34:523–534.

Kain, A., J.-P. Hosom, X. Niu, J. van Santen, M. Fried-Oken, and J. Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49:743–759.

Kaplan, E., A. Wagner, and D. Baskent. 2018. Are musicians at an advantage when processing speech on speech? In *Proceedings of ICMPC15/ESCOM10. Graz, Austria: Centre for Systematic Musicology, University of Graz.*

Pals, C., A. Sarampalis, and D. Baskent. 2013. Listening effort with cochlear implant simulations. *J. Speech Hearing Language Res.*, 56:1075–1084.

Rönnberg, J., M. Rudner, T. Lunner, and A. A. Zekveld. 2010. When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, 12:263–269.

Sauert, B. and P. Vary. 2006. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proc. ICASSP*, pages 493–496, Toulouse, France, May.

Shifas, M., V. Tsiaras, and Y. Stylianou. 2018. Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model. In *Proc. Interspeech 2018*, pages 1868–1872.

Simantiraki, O., M. Cooke, and S. King. 2018. Impact of Different Speech Types on Listening Effort. In *Proc. Interspeech 2018*, pages 2267–2271.

Skowronski, M. D. and J. G. Harris. 2006. Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558.

Taal, C. H., J. Jensen, and A. Leijon. 2013. On optimal linear filtering of speech for near-end listening enhancement. *IEEE Signal Proc. Let.*, 20(3):225–228.

Tang, Y., B. Fazenda, and T. Cox. 2018. Automatic speech-to-background ratio selection to maintain speech intelligibility in broadcasts using an objective intelligibility metric. *Applied Sciences*, 8:59, 01.

Zorila, T., V. Kandia, and Y. Stylianou. 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Proc. Interspeech*, pages 635–638.

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 179-182 recibido 15-03-2019 revisado 09-05-2019 aceptado 14-05-2019

Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante

Social-Univ 2.0: Human Languages Tecnologies applied to the omnichannel monitoring of the University of Alicante's digital social environment

Isabel Moreno[1,2], Javi Fernández[2], Yoan Gutiérrez[1]
[1]Instituto Universitario de Investigación Informática, Universidad de Alicante
[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{imoreno, javifm, ygutierrez}@dlsi.ua.es

Resumen: Este proyecto viene motivado por la necesidad de conocer el ecosistema digital social universitario de aquellas universidades con presencia en redes sociales, y más concretamente, de la Universidad de Alicante. Para ello, las Tecnologías del Lenguaje Humano juegan un papel fundamental, ya que se utilizan para extraer meta-datos sobre comentarios de las redes sociales y representar tanto perfiles de actores sociales (empresas, instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo. El resultado de este proyecto es facilitar el estudio de fenómenos como: las interacciones sociales y sus matices; analizar grupos sociales y sus comportamientos (i.e. patrones de su crecimiento y decrecimiento); identificación de actores sociales relevantes e influyentes; identificar temas comunes en determinados grupos sociales; y la medición de campañas y eventos a través de los medios sociales.
Palabras clave: Tecnologías del Lenguaje Humano, Universidad, Social, Relaciones

Abstract: This project is motivated by the need to know the digital social university ecosystem of those universities present in social networks. More specifically, the focus is the University of Alicante. To that end, Human Language Technologies (TLH) play a fundamental role. TLH are employed not only to extract meta-data from comments from social networks, but also to represent profiles of social actors (companies, institutions, people, etc.) as well as the social relation that exists between them in a certain period of time. The outcome of this project is to ease the study of phenomena such as: social interactions and its nuances; analyse social groups and their behaviours (i.e. their growth and decline patterns); identification of relevant and influential social actors; identify common themes in certain social groups; and measurement of campaigns and events through social media.
Keywords: Human Language Tecnologies, University, Social, Relations

1 Introducción

Actualmente, Internet cuenta con más de 4000 millones de usuarios conectados[1]. La Web 2.0 (o Web social) es uno de los escenarios de Internet donde los usuarios juegan un papel primordial, pues ya no sólo consumen información, si no que los usuarios también forman parte activa en la creación de contenidos. Estos pueden participar, interactuar e intercambiar información con otros usuarios en redes sociales. Por ejemplo si se analizan los datos de Twitter, una de las redes sociales más conocidas, observamos que cuenta con más de 344 millones de usuarios activos[2], que generan 500 millones de tweets al día[3].

Dicha información textual se encuentra en diversas fuentes de información de distinta naturaleza y en distintos idiomas. Estos fac-

[1]`https://www.internetworldstats.com/stats.htm` (Febrero 2019)

[2]`http://www.internetlivestats.com/watch/twitter-users/` (Febrero 2019)

[3]`http://www.internetlivestats.com/twitter-statistics/` (Febrero 2019)

ISSN 1135-5948. DOI 10.26342/2019-63-25
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

tores, unidos a la redundancia y a la información contradictoria, suponen que los usuarios inviertan mucho más tiempo de lo deseado analizando la información para seleccionar aquella que sea de su interés.

Por lo tanto, es necesario reducir el tiempo que los usuarios invierten en analizar grandes cantidades de información formuladas en lenguaje natural. Esto se consigue gracias a las Tecnologías del Lenguaje Humano (TLH) que se encargan de procesar automáticamente el lenguaje humano de manera eficiente y efectiva.

Para su correcto funcionamiento se ha mejorado la búsqueda, recuperación y extracción de información (Irfan et al., 2015; Dalvi, Cohen, y Callan, 2012; El-Helw, Farid, y Ilyas, 2012) o la detección y minería de opiniones (Gutiérrez, Tomás, y Fernández, 2015; Ravi y Ravi, 2015; Montoyo, Martínez-Barco, y Balahur, 2012; Mihalcea, Banea, y Wiebe, 2012; Thelwall y Buckley, 2013), junto con los procesos intermedios involucrados en cada una, como el análisis semántico (Li y Joshi, 2012; Gutiérrez, Vázquez, y Montoyo, 2017).

A pesar de esa mejora, es evidente la necesidad de integrar todos los procesos de TLH anteriores en una sola plataforma. De esta forma, la información de redes sociales que necesita el usuario, será recuperada y procesada para presentársela de manera adecuada y flexible para su posterior análisis.

2 Estado del arte

En la actualidad existen infraestructuras analíticas que de alguna forma incluyen TLH. Algunos ejemplos serían: (i) Atribus[4] rastrea, busca, recoge, filtra y devuelve todo lo que se está diciendo de un cliente en la red a partir de las palabras clave en tiempo real; (ii) Natural Opinions[5] analiza todo lo que se está diciendo en cada momento en Internet sobre una persona, una marca, una institución o un producto, y detecta automáticamente las entidades, conceptos y opiniones más relevantes; (iii) Textalytics[6] extrae elementos con significado de cualquier contenido y lo estructura para que procesarlo y gestionarlo fácilmente; (iv) Sentiment viz[7] estima y visualiza el sentimiento asociado a textos cortos e incompletos; (v) Tweet Reach[8] permite obtener informes estadísticos a partir de analizar twitter; (vii) SocialBro[9] gestiona y analiza comunidades de Twitter para que los profesionales del Marketing analicen a fondo sus contactos y definan sus estrategias consecuentemente; y (viii) SumAll[10] obtiene estadísticas sobre seguidores de varias redes sociales como Facebook o Instagram y genera gráficas de intervalos de tiempo (días, semanas, meses) con números de me gustas, mensajes, localidades, etc.

Sin embargo, si nos enfocamos en el uso de las TLH a día de hoy, no encontramos infraestructuras analíticas que analicen el impacto que las universidades tienen en las redes sociales. Por tanto, si deseáramos conocer el ecosistema digital social en el entorno de una universidad, sería necesario emplear cuantiosos recursos, humanos y temporales, en entender cada noticia y comentario que se emite sobre esa universidad. Después, podríamos crear perfiles e identificar roles de todos aquellos actores y eventos que pudieran incidir en su reputación a nivel social y global.

3 Propuesta

Ante la falta de tecnologías que resuelvan esta problemática, el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI) se plantea este proyecto, donde se persigue como objetivo estudiar, a través del uso de las TLHs, el entorno digital-social universitario. Por este motivo, se elige como caso de estudio la Universidad de Alicante (UA) y entidades relacionadas con esta institución. El objetivo principal se centra en el uso de las TLHs para extraer meta-datos sobre comentarios de las redes sociales y representar tanto perfiles de actores sociales (empresas instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo. La Figura 1 nos proporciona una idea gráfica del resultado esperado.

Para alcanzar nuestras metas requerimos del procesamiento de grandes cantidades de información proveniente de las redes sociales, por lo que se hace uso de plataformas ya existentes en el GPLSI. Entre ellas, destacamos la plataforma Social Analytics (Fernández et al., 2017), con su ayuda podemos obtener

[4] `http://www.atribus.com` (Febrero 2019)
[5] `http://www.bitext.com` (Febrero 2019)
[6] `https://www.meaningcloud.com` (Febrero 2019)
[7] `http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app` (Febrero 2019)
[8] `http://tweetreach.com` (Febrero 2019)
[9] `http://es.socialbro.com` (Febrero 2019)
[10] `https://sumall.com` (Febrero 2019)

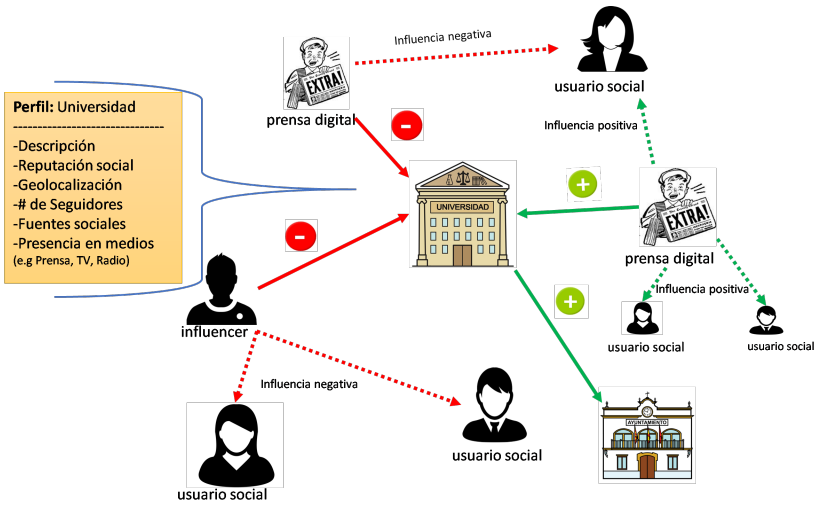


Figura 1: Perfiles y relaciones sociales en la Web 2.0

contenido de la Web Social y extraer metadatos sobre cualquier entidad digital, en este caso de estudio la UA, desde un punto de vista analítico, tras un intenso procesamiento de mensajes de usuarios sociales. A continuación, el análisis de sentimientos y emociones intervienen como las claves para valorar las diferentes posturas (valoración, opinión, etc.) que cada actor tiene respecto a otros. Finalmente se diseñan y aplican algoritmos que permiten obtener modelos de representación tanto de perfiles de actores sociales (empresas, instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo.

Para que el proyecto se lleve a cabo con éxito, son precisas las siguientes tareas:

1. Análisis y definición de tecnologías: Esta tarea tiene como objetivo analizar los requerimientos de este proyecto, estudiando en profundidad los tipos de aplicaciones, recursos, herramientas y tecnologías necesarias, así como sus disponibilidades y viabilidades en el desarrollo de ellas. Se definen también las guías y formato a seguir en el desarrollo de TLH necesarias. Además, se establece el modo de visualización para presentar los datos.

2. Recuperación y extracción de información: En esta tarea se reutilizan y desarrollan TLHs que permitan rastrear, recuperar y extraer información de distintas fuentes (webs, foros, blogs, redes sociales, etc.), formatos (JSON, XML, HTML, etc.) e idiomas (inicialmente inglés y español).

3. Minería y procesamiento de textos: En esta tarea reutilizan y desarrollan las TLHs necesarias para que, a partir de textos escritos en lenguaje humano, estos sean analizados y posteriormente clasificados y estructurados para ser almacenados considerando tecnologías de Minería de Datos. La clasificación resultante es interpretada según: el significado de la unidad de información; los dominios al que se refiere el contexto; la determinación si es o no una opinión (información objetiva versus subjetiva); polaridad emocional implicada; términos relevantes detectados y contabilizados; la geolocalización de contenido generado por usuarios y de autores; etc. Además, dentro de esta tarea se tienen en cuenta qué aspectos pueden influir en la calidad del contenido de la información, como por ejemplo, información redundante, contradictoria, inconsistente, etc., y que pueden afectar al texto generado.

4. Presentación del contenido resultante: En esta tarea se analizan, reutilizan y desarrollan enfoques que permitan representar el contenido procesado automáticamente de un modo útil y sencillo para ser interpretados por usuarios finales. Es decir, representar tanto perfiles de actores sociales como la relación social que existe entre los actores en un determinado período de tiempo.

5. Integración de tecnologías: En esta tarea los mejores enfoques de investigación analizados en las tareas previas se inte-

gran en un único proceso. Esto permite tener como entrada unas necesidades de información, y generar como salida, un paquete con información anotada a distintos niveles, que será la que dé lugar al contenido finalmente mostrado a usuarios finales.

6. Evaluación de los resultados: La evaluación es un aspecto clave y crucial para el proyecto. Se realizan dos tipos de evaluaciones: (i) una evaluación intrínseca de cada una de los enfoques analizados y propuestos para cada tarea involucrada en el proyecto en base a sistemas de referencia, y (ii) una evaluación extrínseca para medir la utilidad de las tecnologías desarrolladas como un todo. Para ambas se utilizan métodos cuantitativos, basados en las medidas tradicionales de precisión y cobertura, y cualitativos, que permitan analizar criterios para los que no existen todavía medidas automáticas (por ejemplo, coherencia de los resultados mostrados).

4 Resultados

El resultado tecnológico de este proyecto es una herramienta llamada *Social-Univ 2.0* que involucra las tecnologías antes mencionadas. En un futuro, dicha herramienta podría adaptarse a aplicaciones móviles y el UA-Cloud[11] de la UA. *Social-Univ 2.0* requerirá para su funcionamiento del uso de tecnologías de minería de textos sociales provenientes del sistema Social Analytics.

Con las tecnologías resultantes de este proyecto es posible estudiar el entorno social de las entidades sociales digitales, en concreto la universidad de Alicante. Esto incluye: estudiar las interacciones sociales y sus matices; analizar grupos sociales y sus comportamientos (i.e. crecimiento, decrecimiento); identificación de actores sociales relevantes e influyentes; identificar temas comunes en determinados grupos sociales; y la medición de campañas y eventos a través de medios sociales.

Agradecimientos

Este proyecto con referencia ENCARGOINTERNO5-19EN: "Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante", está financiado por la Universidad de Alicante.

[11] `https://cvnet.cpd.ua.es/uacloud/` (Febrero 2019)

Bibliografía

Dalvi, B. B., W. W. Cohen, y J. P. Callan. 2012. Websets: extracting sets of entities from the web using unsupervised information extraction. En *WSDM*.

El-Helw, A., M. H. Farid, y I. F. Ilyas. 2012. Just-in-time information extraction using extraction views. En *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, páginas 613–616, New York, NY, USA. ACM.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58:141–148.

Gutiérrez, Y., D. Tomás, y J. Fernández. 2015. Benefits of using ranking skip-gram techniques for opinion mining approaches. En *eChallenges e-2015 Conference*, páginas 1–10. IEEE.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2017. Spreading semantic information by word sense disambiguation. *Knowledge-Based Systems*, 132:47 – 61.

Irfan, R., C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, y et al. 2015. A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(2):157–170.

Li, Y. y K. D. Joshi. 2012. The state of social computing research: A literature review and synthesis using the latent semantic analysis approach. En *AMCIS*.

Mihalcea, R., C. Banea, y J. Wiebe. 2012. Multilingual subjectivity and sentiment analysis. En *Tutorial Abstracts of ACL 2012*, ACL '12, páginas 4–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Montoyo, A., P. Martínez-Barco, y A. Balahur. 2012. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675 – 679.

Ravi, K. y V. Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46.

Thelwall, M. y K. Buckley. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 185-188 recibido 01-04-2019 revisado 16-04-2019 aceptado 17-05-2019

Asistente Conversacional Móvil para la Rehabilitación de la Tartamudez

Mobile Conversational Interface for Stuttering Treatment

David Griol[1], Javier Font[1], Zoraida Callejas[2]
[1]Univ. Carlos III de Madrid. Avda. de la Universidad, 30. Leganés (España)
[2]Univ. de Granada. Calle Periodista Daniel Saucedo Aranda, s/n. Granada (España)
dgriol@inf.uc3m.es, 100291364@alumnos.uc3m.es, zoraida@ugr.es

Resumen: Los dispositivos móviles se han convertido en una herramienta de uso diario que ha provocado cambios en las formas de interactuar de las personas gracias a la accesibilidad que ofrecen a grandes contenidos de información y a servicios de comunicación de forma continua y prácticamente ubicua. La combinación de estos dispositivos y los interfaces conversacionales ha facilitado el desarrollo de aplicaciones en el campo de la salud móvil (mHealth en inglés) cada vez más avanzadas. En este trabajo presentamos una app que utiliza el potencial de los interfaces conversacionales para ayudar a mejorar la fluidez en el habla en personas que padecen tartamudez.
Palabras clave: Tartamudez, salud móvil, interfaces conversacionales, reconocimiento de emociones

Abstract: Mobile devices have become a tool for daily use that has caused changes in people's ways of interacting, thanks to the accessibility they offer to large contents of information and communication services continuously, regardless of the place and the moment. Thanks to the development of mobile applications, advanced apps have been developed in the area of mobile health (mHealth), a field of eHealth in which the practice of medicine and public health is supported by mobile devices. In this paper we present an app that uses the potential of conversational interfaces for the development of mobile applications that help improve speech fluency in people who suffer from stuttering.
Keywords: Stuttering, mobile health, conversational interfaces, emotion recognition

1 Introducción

Uno de los trastornos de la comunicación más conocidos es la tartamudez o disfemia, que está asociada a las interrupciones involuntarias del habla. Se trata de un trastorno de la comunicación y no del lenguaje que, según la Fundación Española de la Tartamudez, afecta a un 2 % de adultos y a un 5 % de niños. Estos datos se traducen aproximadamente en 800.000 personas en España (Fundación Española de la Tartamudez, 2019).

Los síntomas de la tartamudez se manifiestan, sobre todo, en situaciones en las que un paciente tiene que interactuar en público. Sin embargo, si el paciente se encuentra en un entorno agradable y aislado, puede que sea capaz de hablar de forma fluida. Por esta razón, existen técnicas empleadas en Logopedia, que consisten en engañar al cerebro, de tal forma que el paciente piense que está estableciendo una conversación (Jackson, Rodgers, y Rodgers, 2019). Estas técnicas emplean dispositivos electrónicos que tienen la capacidad de aplicar un tiempo de retardo a la señal de voz y de modificar la frecuencia fundamental de esta señal, cuyo uso provoca una reducción de la velocidad en el habla y la prolongación de vocales, minimizando los efectos que produce la tartamudez (Jaramillo y Gil-Lozada, 2014; Salgado-Ruíz, 2005).

Los avances en las capacidades y sensores de los dispositivos móviles han posibilitado generar aplicaciones (Apps) destinadas a la salud y a los servicios sociosanitarios que han supuesto una revolución en los últimos años gracias a que pueden ofrecer funcionalidades como las mencionadas anteriormente, basadas en la retroalimentación auditiva con modificaciones del tono, y con cierto retardo de la señal de voz, ayudando de forma posi-

ISSN 1135-5948. DOI 10.26342/2019-63-26 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

tiva a minimizar los efectos de la tartamudez a través de un dispositivo móvil.

En este artículo presentamos una aplicación para dispositivos móviles con sistema operativo Android basada en un interfaz conversacional hablado (McTear, Callejas, y Griol, 2016; Griol et al., 2014) que incorpora herramientas para mejorar la fluidez verbal de personas que padecen tartamudez. Se ha desarrollado un sistema con un funcionamiento sencillo e intuitivo que fomenta la comunicación verbal a través del habla. El interfaz conversacional es el elemento central de la aplicación, pues a partir de él, el usuario tendrá que interactuar de forma oral para poder ejecutar las diferentes funcionalidades facilitadas por la aplicación. Además, la aplicación dispone de un módulo basado en plantear una serie de preguntas que permitan evaluar el estado emocional del usuario. Esta información se facilita al agente conversacional para realizar recomendaciones al usuario entre el conjunto de actividades integradas en la aplicación, además de considerar las actividades completadas por el usuario en interacciones anteriores con la aplicación y los resultados obtenidos.

A partir de la búsqueda en Google Play Store del texto "Tartamudez", podemos encontrar un número muy pequeño de aplicaciones dedicadas a su tratamiento. Partiendo de la información sobre las funcionalidades que ofrecen, se puede comprobar que la mayoría de estas aplicaciones ofrecen funcionalidades basadas en la retroalimentación auditiva retardada con la posibilidad de modificar el tono de la señal de voz. Otras aplicaciones simplemente contienen artículos o permiten la realización de un test para la detección temprana de la tartamudez. La aplicación desarrollada, en cambio, proporciona las siguientes funcionalidades:

- Un módulo de aprendizaje con consejos y contenidos multimedia didácticos para mejorar la fluidez verbal.
- Un módulo de juegos para mejorar tanto la fluidez verbal, como el control de la respiración de una forma más práctica y entretenida.
- Un módulo de entrenamiento para la lectura de textos en voz alta, a partir de la selección de un sistema de reconocimiento, grabación y síntesis de de voz, o de un sistema de retroalimentación auditiva de la señal de voz con un tiempo de retardo.
- Un módulo de relajación que facilita mejorar el estado de ánimo.

Con la intención de conseguir una aplicación que consuma el menor número de recursos disponibles del dispositivo móvil, se ha integrado una base de datos en la nube, que almacena el mayor número de contenidos posibles que conforman la interfaz gráfica de la aplicación (por ejemplo, los textos que se muestran en pantalla). Además, esta base de datos almacena las consultas solicitadas al usuario, los diferentes estados de control y resultados procedentes de las actividades realizadas por el usuario, necesarias para que el interfaz conversacional pueda adaptar la interacción dependiendo de la evolución, preferencias y requisitos de cada usuario.

2 App desarrollada

La aplicación desarrollada está conformada por un total de siete módulos principales. En el módulo inicial es necesario permitir el acceso al micrófono para grabar los audios de los usuarios. En caso de que el usuario utilice la aplicación por primera vez, se le solicita unos datos para su registro en la base de datos de la aplicación en la plataforma para el desarrollo de aplicaciones web y aplicaciones móviles Firebase[1].

El segundo módulo de la aplicación está conformado por un agente conversacional diseñado en la plataforma Dialogflow[2], que interactúa con el usuario en función de los valores almacenados en la base de datos, la actividad previa del usuario y última fecha de uso de la aplicación, su estado emocional y el modelo de diálogo diseñado para que el agente interactúe con el usuario según el tipo de actividad propuesta por la aplicación.

El módulo de estado emocional se encuentra conectado al servicio web de analítica de textos proporcionado por Meaningcloud[3] para que el agente conversacional pueda estimar el estado de ánimo del usuario. El estado emocional se obtiene a partir del análisis de las respuestas proporcionadas por el usuario para una serie de preguntas extraídas del

[1] https://firebase.google.com/
[2] https://dialogflow.com/
[3] https://www.meaningcloud.com/es

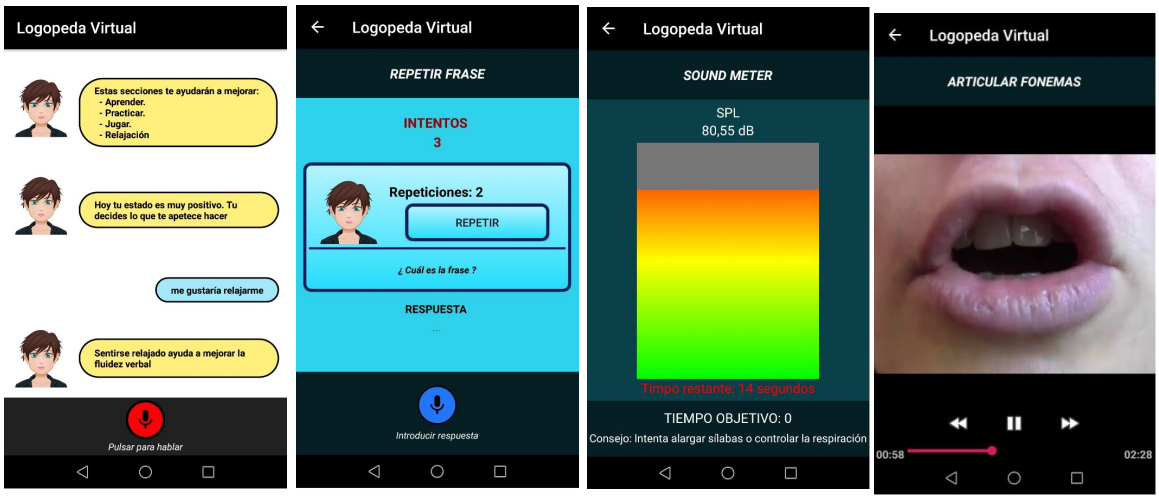


Figura 1: Capturas de pantalla de la app (I): asistente conversacional y módulo de práctica

Figura 2: Capturas de pantalla de la app (II): módulo de relajación, módulo de juegos y módulo de aprendizaje

nodo correspondiente en la base de datos. El usuario debe ir contestando a cada una de las preguntas por medio de la voz. Para garantizar que el sistema de transcripción de la voz ha funcionado de forma correcta, la app posibilita visualizar la transcripción y confirmar la respuesta antes de pasar a la siguiente pregunta. El resultado final de evaluación del estado de ánimo del usuario se extrae de una función interna de la aplicación, que se encarga de realizar un sistema de puntuaciones en función de la valoración del texto generada por Meaningcloud. A partir de los resultados obtenidos del test emocional, el asistente virtual recomendienda al usuario acciones adecuadas a su estado de ánimo.

Para los casos en los que el estado emocional sea negativo o muy negativo, el asistente conversacional inicia de forma automática el módulo de relajación, con el objetivo de mejorar el estado de ánimo del usuario. Este módulo está conectado al nodo de control de actividad y facilita ejercicios de relajación de corta duración basados en audios y vídeos. Una vez terminada la actividad, se redirige al usuario a la pantalla de interacción con el asistente conversacional.

El módulo de juegos está conformado por juegos que ayudan a mejorar la respiración y la fluidez verbal. Estos juegos extraen los enunciados y las soluciones del nodo de la base de datos que administra los contenidos de los mismos. Además, este bloque está conectado al nodo de control de actividad del

usuario que también, va a almacenar los resultados obtenidos tras su correspondiente finalización. Se han integrado tres tipos de actividades: ordenar palabras para formar una frase con sentido, sonómetro para solicitar al usuario mantener la presión sonora durante un determinado tiempo, repetición de la frase recitada por el asistente virtual. Para este tipo de actividades se han definido ejercicios de tres grados de dificultad (baja, media y alta). En todos ellos, el usuario interactúa con la aplicación a través de la voz.

El módulo de práctica permite almacenar textos introducidos por el usuario en la base de datos a través del nodo encargado de la gestión de textos. Estos textos se utilizan posteriormente en un sistema de grabación de voz o en un sistema de retroalimentación retardada, de tal modo que el usuario pueda realizar prácticas de lectura en voz alta para mejorar la fluidez verbal. En el modo de grabación, el usuario puede grabar su voz para posteriormente reproducirla. Cada vez que se utilice este modo se genera un archivo multimedia con formato 3GP que se almacena en el el dispositivo móvil, pero que puede ser reemplazado cada vez que se realice una nueva grabación del mismo texto, con el objetivo de no consumir grandes cantidades de memoria que limiten el dispositivo. En el modo de retroalimentación auditiva, el usuario podrá ir escuchando su propia voz con un tiempo de retardo mientras realiza la grabación de su mensaje. Para que este modo sea de utilidad, el usuario deberá de utilizar unos auriculares durante la práctica de lectura en voz alta.

Finalmente, el módulo de aprendizaje está conformado por actividades con contenidos multimedia que permitan al usuario mejorar su fluidez verbal y controlar la respiración. Este módulo también se encuentra conectado a un nodo de la base de datos que contiene una lista de consejos y con el nodo de la base de datos encargado de controlar la actividad del usuario. Se han incluido tres tipos de actividades: articulación de fonemas (contiene archivos de vídeo que muestran la articulación de los diferentes sonidos vocálicos y consonánticos que conforman el abecedario a través de ejemplos prácticos), respiración (contiene archivos de vídeo cuyos contenidos enseñan al usuario los beneficios de practicar una respiración correcta y controlada) y consejos prácticos. Las Figuras 1 y 2 muestran diferentes capturas de pantalla de los módulos de la app descritos en esta sección.

3 Conclusiones y trabajo futuro

En este artículo hemos descrito una aplicación móvil cuya funcionalidad es ayudar a mejorar los problemas de comunicación relacionados con la fluidez verbal debidos a la tartamudez o disfemia. Con respecto a las diferentes aplicaciones relacionadas con la tartamudez que se pueden encontrar en Google Play Store, la aplicación desarrollada incorpora funcionalidades novedosas y distintivas de las existentes, destacando la implementación de un agente conversacional con la capacidad de interactuar, motivar y basarse en el estado de ánimo del usuario para recomendar actividades de relajación, práctica de juegos y actividades de aprendizaje y mejora. Como trabajo futuro, además de extender el número de actividades de la aplicación, queremos validar la aplicación con usuarios que padecen tartamudez, pues hasta ahora solamente hemos realizado una evaluación preliminar con cuatro pacientes.

Bibliografía

Fundación Española de la Tartamudez. 2019. https://www.fundacionttm.org/la-tartamudez/mas-informacion/. (Consultado: Marzo 2019).

Griol, D., Z. Callejas, R. López-Cózar, y G. Riccardi. 2014. A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer Speech and Language*, 28(3):743–768.

Jackson, E., N. Rodgers, y D. Rodgers. 2019. An exploratory factor analysis of action responses to stuttering anticipation. *Journal of Fluency Disorders*, 60:1–10.

Jaramillo, J. H. y Y. Gil-Lozada. 2014. Efectos de la retroalimentación auditiva retardada en los patrones de tartamudez. *Revista Ciencias de la Salud*, 12(2):243–251.

McTear, M., Z. Callejas, y D. Griol. 2016. *The conversational interface: Talking to smart devices.* Springer.

Salgado-Ruíz, A. 2005. *Manual práctico tartamudez.* Síntesis.

Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 189-192 recibido 26-03-2019 revisado 07-05-2019 aceptado 17-05-2019

Plataforma Web para la evaluación de la comprensión lectora aplicada a tareas de búsqueda

Web Platform for the evaluation of reading comprehension applied to search tasks

Javier Perea de la Casa, Fernando Javier Martínez Santiago, Miguel Ángel García Cumbreras, Manuel Carlos Díaz Galiano
Universidad de Jaén
{jperea, dofer, magc, mcdiaz}@ujaen.es

Resumen: Presentamos una plataforma web que constituye una herramienta para poder evaluar la comprensión lectora a través de dos experimentos: un test de fluidez lectora que permitirá inferir el grado de fluidez lectora del participante y un sistema de recuperación de información interactivo, donde el participante explorará y evaluará un conjunto de documentos como relevantes o no relevantes. La cuestión que se plantea es la incidencia del nivel de fluidez lectora en la búsqueda de información.
Palabras clave: Comprensión y fluidez lectora, recuperación de información interactiva

Abstract: We present a web platform that constitutes a tool to evaluate reading comprehension through two experiments: a reading fluency test that will allow inferring the degree of participant's reading fluency and an interactive information retrieval system where the participant will explore and evaluate a set of documents as relevant or not relevant. The question to be answered is the impact of the level of reading fluency regarding with performance in tasks of searching for information.
Keywords: Reading comprehension, reading fluency, interactive information retrieval

1 Introducción

Desde principios del siglo XX, educadores, pedagogos y psicólogos han tenido en cuenta la importancia de la comprensión lectora, intentando explicar qué sucede cuando una persona comprende un texto. Una correcta comprensión lectora implica que el lector pase por todos los niveles de lectura (Langer, 1995), con el fin de conseguir una comprensión global, elaborando una interpretación y reflexionando sobre el contenido y la estructura de un texto. Hay que destacar, además, el uso de conocimiento o experiencia previa, cuando se habla de comprensión lectora.

Se puede definir la fluidez lectora como la capacidad de leer palabras, pseudopalabras y textos, sin cometer errores de manera expresiva y a un ritmo adecuado, de forma que tanto el lector como un eventual interlocutor puedan comprender lo leído (Hudson et al., 2005; Kuhn y Stahl, 2003; National Reading Panel, 2000).

En este artículo se presenta una plataforma web como una herramienta de evaluación de la comprensión lectura, a partir de dos experimentos: un test de fluidez lectura y un sistema de recuperación de información interactivo. La cuestión que se plantea es la incidencia del nivel de fluidez lectora cuando el participante está realizando tareas específicas de búsqueda de información mediante el uso de un sistema de Recuperación de Información.

El resto del artículo está organizado de la siguiente forma. El capítulo 2 muestra el corpus utilizado y los juicios de relevancia. En el capítulo 3 se describen los casos de uso y la interfaz de la plataforma web desarrollada. Por último, el capítulo 4 muestra los experimentos realizados y el análisis de los resultados obtenidos.

2 Corpus y juicios de relevancia

El corpus generado está formado por 20 consultas predefinidas de diversas temáticas. Para cada una se han recopilado 20

ISSN 1135-5948. DOI 10.26342/2019-63-27 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

documentos, utilizando como fuentes de información NEWSELA y Google.

El corpus de NEWSELA[1] incluye miles de artículos de noticias profesionalmente adaptados a diferentes complejidades de lectura divididas en cinco niveles (0: más complejo – 4: menos complejo). Conforme aumentamos el nivel, más simplificada estará la información respectiva a dicho documento, es decir, la información pasa a ser un subconjunto del nivel anterior.

Los documentos procedentes de Google se han recuperado manualmente a través del motor de búsqueda lanzando cada una de las 20 consultas.

Respecto a los juicios de relevancia que intervienen en la complejidad cualitativa del texto, se plantea la siguiente cuestión: ¿cómo llegan a evaluar los distintos usuarios los documentos en contraste con la evaluación de la relevancia proporcionada por los expertos?

En un primer momento, nos centramos en las características cualitativas del texto, de manera que el nivel de la dificultad del texto no es calculado de ninguna forma, pero sí juzgado y contrastado por los expertos. Para este fin, usamos el corpus NEWSELA como se ha descrito previamente.

Tras esto, evaluamos el rendimiento del usuario cuando se compara la evaluación de la relevancia proporcionada por los expertos y la evaluación procedente de cada participante de los experimentos.

La pregunta es si es posible encontrar diferencias significativas en el grado de concordancia entre los expertos y los participantes acorde a la complejidad del texto.

3 *Casos de uso y plataforma web desarrollada*

A continuación, presentamos casos de uso donde se reflejarán las acciones de diversos usuarios sobre la plataforma web desarrollada.

En primer lugar, el participante deberá completar un test de fluidez lectora donde al comienzo, se le informará de una serie de prerrequisitos y se le asignará un ID para tener acceso al siguiente experimento. La Figura 1 muestra la página inicial del test, en la plataforma web.

Figura 1: Página inicial del test de fluidez lectora

A continuación, el participante deberá completar un test de entrenamiento previo al test oficial de fluidez lectora para así orientar la interacción del participante con el experimento. Todas las preguntas han de responderse correctamente para pasar al test oficial de fluidez lectora. La Figura 2 muestra la plataforma web para este segundo test.

Figura 2: Test oficial de fluidez lectora

Por último, el usuario completará un formulario de datos personales que nos garantizará una mayor flexibilidad a la hora de segregar los resultados de los experimentos en varias poblaciones.

El siguiente experimento consiste en un sistema de recuperación de información interactiva (PIR, del inglés *Personal Information Retrieval*), donde el participante evaluará un conjunto de documentos como relevantes o no relevantes en base a la consulta lanzada.

Al acceder a la página inicial, el participante se encontrará con los correspondientes prerrequisitos y un campo de inicio de sesión que deberá completar con el ID asignado en el

[1] Disponible en https://NEWSELA.com/data/ (última visita: 24/03/2019)º

experimento previo. La Figura 3 muestra la página de entrada al sistema PIR.

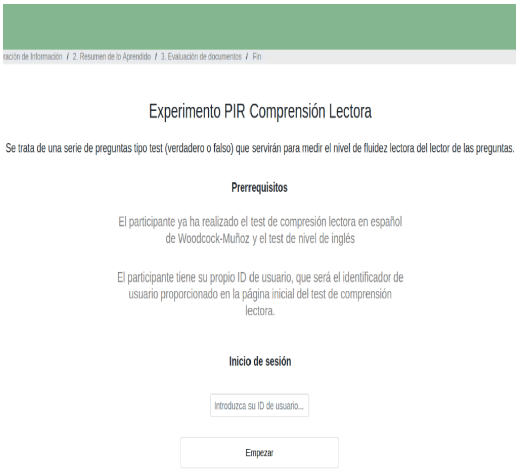


Figura 3: Página inicial del sistema PIR

A continuación, se iniciará una nueva sesión en la que el participante deberá seleccionar una de las veinte consultas disponibles e iniciar la búsqueda de información. Tras lanzar la consulta, el sistema recuperará el conjunto de veinte documentos que el participante explorará de manera totalmente libre. La Figura 4 muestra la pantalla de los documentos del sistema PIR.

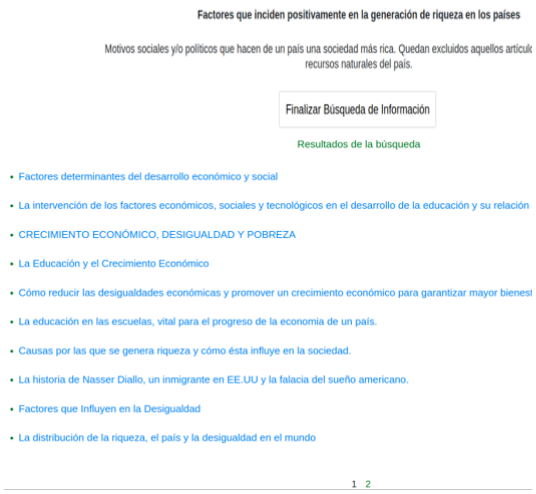


Figura 4: Ventana del sistema de recuperación de información

Tras este paso el participante debe redactar un breve resumen de lo aprendido a lo largo de la fase de búsqueda de información. Esta fase recibirá el nombre de *resumen de lo aprendido*.

Una vez finalizado el resumen, pasamos al proceso central del sistema, en el que mostrará el conjunto de documentos a evaluar como relevantes o no relevantes en base a la consulta lanzada.

Finalmente, una vez que el participante evalúe todos los documentos, finalizaría la sesión de evaluación, y, en consecuencia, se le permitirá al mismo iniciar una nueva sesión descartando del conjunto de posibles consultas a lanzar, ya sí, la consulta que ha evaluado.

4 *Experimentación realizada*

El test de fluidez lectora se realizó sobre un conjunto de 46 participantes (descartando los participantes que presentaron valores atípicos, valores mayores que 1.5 veces el valor del rango intercuartil más allá de los cuartiles), principalmente compuesto por alumnos de grado y master, con edades comprendidas entre los 22 y 29 años (x=24,2, std=3) y en su mayor parte varones (41 hombres 5 mujeres). De estos 46, 14 participantes han realizado el segundo experimento, concerniente a la tarea de búsqueda y evaluación de documentos.

4.1 Experimento 1: Test de fluidez lectora

Las Figuras 5 y 6 muestran las respuestas obtenidas en el test de fluidez lectora, sobre un total de 46 participantes. La Tabla 1 muestra los datos estadísticos obtenidos con este primer test.

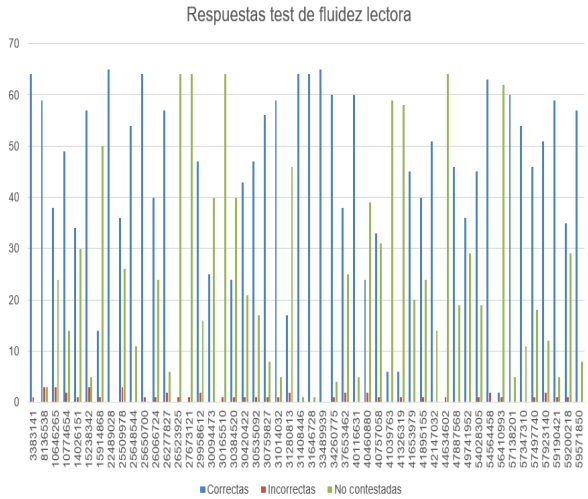


Figura 5: Respuestas obtenidas con el test de fluidez lectora

	Promedio	Desviación típica
Correctas	44,76	17,15
Incorrectas	1,02	0,93
No contestadas	19,22	17,08

Tabla 1: Datos estadísticos del test

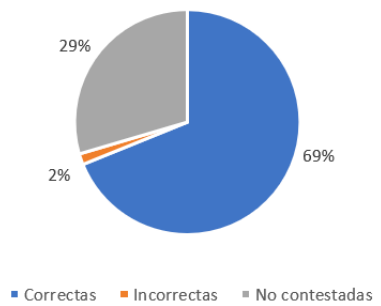


Figura 6: Proporción de respuestas recopiladas

Como podemos observar, el desempeño obtenido por los participantes por lo general ha sido satisfactorio en la medida que el promedio de respuestas correctas (44,76) se sitúa por encima del valor medio del total de respuestas posibles a registrar puesto que el test cuenta con 65 afirmaciones a evaluar. El valor medio de las respuestas recopiladas está entre 32,5 y 33.

4.2 Experimento 2 : Aplicación PIR

Dado que los documentos de NEWSELA presentan diversos niveles de complejidad lectora, es probable que para una misma consulta el participante llegue a evaluar el mismo documento de NEWSELA con diferente grado de dificultad. Por ello, para profundizar en el análisis de la comprensión lectora del participante, se analiza si hay diferencia estadísticamente significativa entre los resultados si distinguimos entre documentos fáciles y difíciles. La Tabla 2 muestra los datos obtenidos al realizar este análisis.

	Media	Desviación típica
Difíciles	7,2	6,3385
Fáciles	7,3	5,2925

Tabla 2: Datos estadísticos por dificultad de los documentos

Como se puede observar, no existe diferencia estadísticamente significativa entre estos resultados, indicando que los participantes han fallado y acertado en la misma medida, independientemente de la complejidad del texto, y que, por lo tanto, la comprensión lectora de estos participantes no se ha visto influida por la dificultad de los textos.

Sin embargo, podemos observar que el número de documentos evaluados correctamente fluctúa más (desviación típica) cuando tienen una mayor complejidad lectora que cuando son más fáciles de leer: 6,34 (difíciles) frente a 5,29 (fáciles).

5 *Conclusiones*

Para comprobar si realmente existe una correlación entre el desempeño del participante y su nivel de fluidez lectora calculamos los coeficientes de correlación entre los niveles de acierto y error en los juicios de relevancia realizados por cada participante y su nivel de fluidez lectora:

$$\mathbf{R}_{\text{Correctas}} = 0{,}245 \qquad \mathbf{R}_{\text{Incorrectas}} = -0{,}3024$$

Los resultados obtenidos demuestran que existe una correlación débil entre el desempeño del participante y su fluidez lectora. Esto lo justifica que ambos coeficientes de correlación tengan valores próximos a cero y, en consecuencia, no se puede afirmar con determinación que una mejor fluidez lectora conlleva una mayor habilidad a la hora de establecer los juicios de relevancia de un documento dado.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto del Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto REDES (TIN2015-65136-C2-1-R).

Bibliografía

Hudson, R.F., Lane, H.B. y Pullen, P.C. 2005. Reading fluency assessment and instruction: What, why and how? *The Reading Teacher*, 58, 702-714.

Kuhn, M.R. y Stahl, S.A. 2003. Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95, 3-21.

Langer, J.A. 1995. *Envisioning literature: Literary understanding and literature instruction*. New York: Teachers College Press.

National Reading Panel. 2000. *Teaching children to read: An evidence-based assessment of scientific research literatura on reading and its implications for reading instruction: reports of subgroups*. Bethesda, MD: National Institute of Child Health and Human Development.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 193-196 recibido 31-03-2019 revisado 06-05-2019 aceptado 15-05-2019

Corpus Viewer: NLP and ML-based Platform for Public Policy Making and Implementation

Corpus Viewer: una plataforma basada en PLN y Aprendizaje Automático para diseño e implementación de política pública

David Pérez-Fernández[1], **Jerónimo Arenas-García**[1,2], **Doaa Samy**[1,3]
Antonio Padilla-Soler[4], **Vanesa Gómez-Verdejo**[2]
[1]Secretaría de Estado de Avance Digital, Spain
[2]Universidad Carlos III de Madrid, Spain
[3]Instituto de Ing. del Conocimiento (IIC), Madrid, Spain and Cairo University, Egypt
[4]Ministerio de Industria Comercio y Turismo, Spain
dperezf@mineco.es, jeronimo.arenas@uc3m.es, doaa.samy@iic.uam.es

Abstract: Corpus Viewer is a production service developed by the State Secretary for Digital Advancement (SEAD) within the framework of the National Language Technologies Plan (Plan TL), promoted by the same State Secretary. Corpus Viewer relies on Natural Language Processing (NLP), Machine Learning (ML) and Machine Translation (MT) to analyze structured metadata and unstructured textual data in large document corpora. The platform allows the decision maker and the policy implementer the possibility of analyze R&D&i information space (mainly patents, scientific publications and public aids) for evidence and knowledge-based policy making and implementation. In this paper, we describe the main functionalities of the platform and enumerate the techniques it is based on, which include a variety of methods like document topic modeling and graph analysis.
Keywords: Topic modeling, Latent Dirichlet Allocation (LDA), graph analysis, Document Similarity, Automatic Classification, dynamic topic analysis

Resumen: Corpus Viewer es un servicio en producción desarrollado por la Secretaría de Estado del Avance Digital dentro del marco del Plan de Impulso de Tecnologías del Lenguaje (Plan TL). Se basa en técnicas de Procesamiento del Lenguaje Natural (PLN) y Aprendizaje Automático para analizar datos estructurados y no-estructurados en grandes colecciones de documentos como las patentes, las publicaciones científicas de acceso abierto, los proyectos europeos, etc. El objetivo es ofrecer al decisor político y al gestor la posibilidad de navegar en el espacio de la información teniendo una visión de conjunto que le ayude a tomar decisiones basadas en conocimiento y evidencias. En este artículo, se describen las funcionalidades básicas de la plataforma enumerando las técnicas empleadas que incluyen, entre otros, modelados de tópicos y análisis de grafos.
Palabras clave: Modelado de Tópicos, Latent Dirichlet Allocation (LDA), Análisis de Grafos, Similitud entre Documentos, Classificación Automática, Modelado Dinámico de Tópicos.

1 Introduction

In this paper, we present Corpus Viewer, a production service developed by the State Secretary for Digital Advancement (SEAD) within the framework of the National Language Technologies Plan, promoted by the same State Secretary. The development of this service started in 2016 and is still progressing based on the collaboration of several subcontracted University research groups and companies. Corpus Viewer on its 1.0 version, is currently used by three public administrations: SEAD (Ministry of Economy), the Spanish Foundation for Science and Technology (FECYT) and the State Secretary for University and Research, Development and Innovation (SEUIDI) at the Spanish Ministry of Science.

Corpus Viewer relies on Natural Language Processing (NLP), Machine Learning (ML),

ISSN 1135-5948. DOI 10.26342/2019-63-28 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

and Machine Translation(MT) to analyze structured metadata and unstructured text data in large quantities of documents. Although a generic platform that can be exploited with virtually any collection of text documents, the current deployment of the platform mainly hosts R&D related text corpora, such as patents, scientific publications and projects funded at national, European (Cordis) and international level (NSF, NIH). These data sources are processed to assist in the definition and implementation of R&D&i public policies through a set of functionalities allowing to:

1. compare R&D&i funding and knowledge areas in different geographic regions,
2. identify competitive advantages between countries, regions, organizations,
3. identify R&D&i knowledge areas, as well as their emergence, evolution and even hybridization with other knowledge areas (it provides also metadata aggregation and BI type dashboard visualization),
4. R&D agent (organization, researcher and firm) profiling and,
5. assist in the assessment of the impact of public policies by tracking the outputs of grants, short and long term outcomes in terms of lead-lag.

Corpus Viewer also provides tools for policy implementation, in particular for the selection of evaluators or the retrieval of relevant documents (patents, scientific publications, R&D aid grants and proposals, etc) for innovation evaluation. Furthermore, it is used for plagiarism detection, identification of cases of double funding and fraud in aid grants and proposals submitted for national funding.

In this paper, we briefly describe the basics of the Corpus Viewer platform and announce also some of the newest functionalities that have already been completed and will be incorporated to version 1.5 to be deployed this year. In the next section we briefly describe the main functionalities offered by the platform, whereas Section 3 enumerates some of the most relevant techniques that constitute the artificial intelligence core of the platform.

2 *Corpus Viewer functionalities*

Current version of Corpus Viewer provides the following functionality for public policy design and implementation:

1. Scalable NLP pipeline and Machine Translation of large volumes of documents.
2. Automatic classification of documents according to available taxonomies using deep learning networks.
3. Topic modeling analysis of document collections and topic inference for new documents.
4. Information retrieval system based document similarity. This function is used to provide identify highly similar documents for plagiarism detection or to avoid project reevaluation.
5. Optimized columnar and textual indexing for efficient searches and query metadata filtering.
6. Thematic correlation and main semantic area detection. Semantic grouping, using semantic graph algorithms, by area of knowledge.
7. Tracking of semantically alike documents (semantic clusters) (emergence, evolution and hybridization with other clusters.
8. Dynamic topic analysis and temporal thematic evolution. Temporal analysis by areas of knowledge, lead-lag between different types of document corpus (Figure 1).
9. Topic-enhanced dashboards to analyze R&D&i distribution by metadata, including geographic area (Figure 2).
10. Automatic profiling and disambiguation of R&D&i key players (researchers, institutions, groups) based on their R&D production.
11. Analysis of collaboration networks between key players in the R&D&i production space.

3 *Facilitating techniques*

In order to provide the previous functions, the platform heavily relies on a number of NLP and ML techniques:

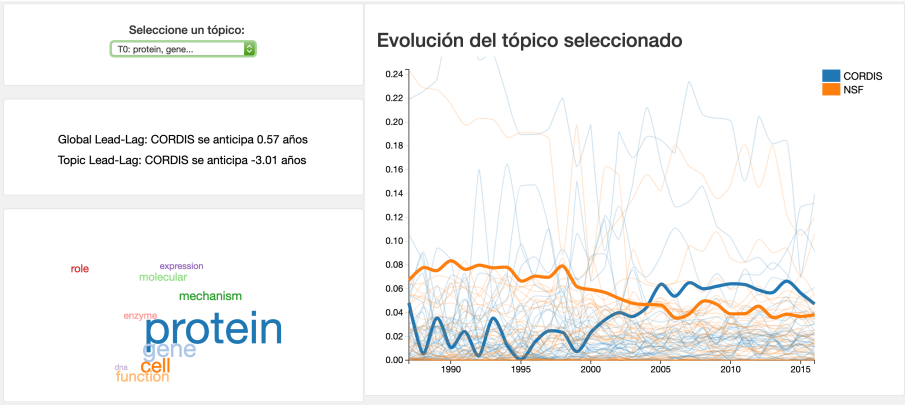


Figure 1: Time evolution and comparison between number of granted projects by topic in FP7/H2020 and NSF.

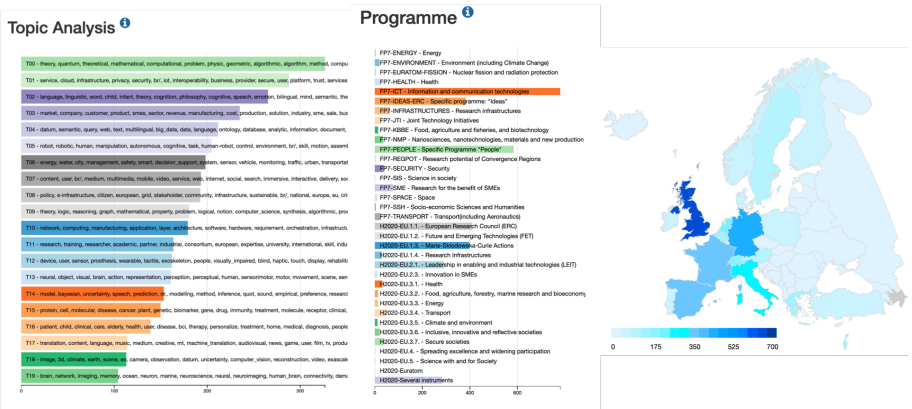


Figure 2: Topic-enhanced dashboard for analysis of the distribution of Cordis projects in the area of Artificial Intelligence.

1. Scalable tokenization, PoS tagging, lemmatization, disambiguation and wikification for English and Spanich languages.
2. Automatic translation (ES-EN), NMT based.
3. Topic modeling, including: 1) static models, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) and Correlated Topic Models (CTM) (Blei and Lafferty, 2005), 2) dynamic models using the Dynamic Topic Model (DTM) of (Blei and Lafferty, 2006), and a new tool for time analysis developed within the scope of the project and based on the work (Greene and Cross, 2017), and 3) hierarchical LDA and recursive LDA, developed within the scope of the project, to allow multi-level navigation in large corpora.
4. Textual search and document similarity implemented for topic, bag of words, and word Embeddings document representations (Mikolov et al., 2013).
5. Analysis of graphs; modularity, distances between clusters and centrality calculation

Such analysis techniques are based on underlying topics, conceptual indexing of documents, word Embeddings and other techniques of documentary representation. The relations of semantic similarity among the document representations (scientific literature, patents and funded project proposals and grants) equip this space with a metric that favors representation and navigation. Similar documents concentrate in semantic aggregates (document clouds are stable under metric modifications). Semantic aggregates can be obtained using graph analysis tools –see Fig. 3.

Unlike traditional statistical description methods, our approach does not require predefined taxonomies or controlled vocabular-

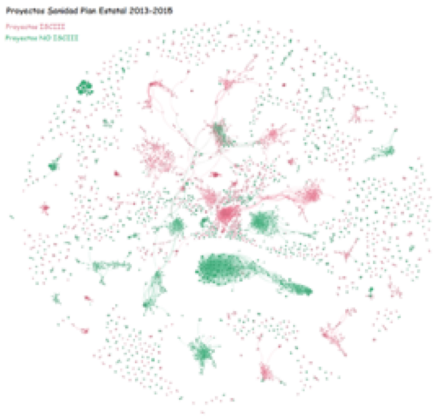


Figure 3: Graph of the Spanish health R&D project proposals and grants based on topic analysis, topic vector interdistances and graph community detection. Different colors represent proposals funded by ISCIII (a health-funding institution, red) and other Spanish R&D funding organizations (green).

ies. Among the advantages of this content-based approach are: increased flexibility for comparing documents from different corpora (without common taxonomies), no need for predefined taxonomies (especially important in wide and dynamic R&D sectors), and the possibility to represent and characterize R&D agents (researchers, research centers, firms) using their R&D production.

4 Use cases examples

In this Section we provide examples of Corpus Viewer use cases and visualizations:

Use-case 1: Comparing the funding destined to certain thematic fields in different countries and its evolution within a certain time frame is of high importance to R&D&i policy makers. Figure 1 illustrates an example of what Corpus Viewer can offer in this use-case. This would provide data-based evidence to answer questions such as:

- How many projects were funded in a certain knowledge area (previously automatically detected) through the framework programmes FP7 and H2020?
- How many projects in the same area were funded by the National Science Foundation within the same time frame?

Use-case 2: Given a certain knowledge field, e.g. Artificial Intelligence, the policy maker would like to know better the investment, and answer questions such as:

- What are the field subareas and how does the invested money distribute among them?
- What countries receive the largest funding in total and in each detected subarea (see Fig. 2).

Use-case 3: Figure 3 represents a semantic graph of R&D health projects funded by Instituto de Salud Carlos III (ISCIII) and by other institutions. Using this approach, Corpus Viewer can help answering the following questions:

- How does the thematic distribution of applications compare to the granted projects?
- Are different institutions (or countries, regions)) funding the same subfields? Are there some fields that receive insufficient funding?

Acknowledgments

This work has been carried out in the framework of the Spanish State Plan for Natural Language Technologies. We would like to acknowledge the different organizations that have contributed to the project under the SEAD-SEUIDI-FECYT agreement for Competitive Intelligence: UPM, IIC, UPF, IXA UPV, Elhuyar.

The work of J. Arenas-García and V. Gómez-Verdejo has been partly funded by MINECO projects TEC2014-52289-R and TEC2017-83838-R.

References

Blei, D. M. and J. D. Lafferty. 2005. Correlated topic models. In *Proc. NIPS*.

Blei, D. M. and J. D. Lafferty. 2006. Dynamic topic models. In *Proc. ICML*.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. In *Proc. NIPS*.

Greene, D. and J. P. Cross. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25:77–94.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 197-200 recibido 29-03-2019 revisado 08-04-2019 aceptado 14-05-2019

Atención automatizada a estudiantes en el proceso de matriculación en la Universidad de Alicante

Automatic Students Service in the Enrolment Process at the University of Alicante

Isabel Moreno[1,2], Yoan Gutiérrez[1], Andrés Montoyo[1,2]
[1]Instituto Universitario de Investigación Informática, Universidad de Alicante
[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{imoreno, ygutierrez, montoyo}@dlsi.ua.es

Resumen: Esta demo presenta un chatbot diseñado para automatizar la atención a los estudiantes durante el proceso de matricula en la Universidad de Alicante (UA). Hemos creado un chatbot en la plataforma de mensajería instantánea Telegram mediante tecnologías del lenguaje humano e inteligencia artificial. El principal obstáculo ha sido ofrecer el chatbot en los dos idiomas oficiales de la Comunidad Valenciana, por lo que se ha utilizado traducción automática como componente intermedio. Este chatbot ofrece numerosos beneficios a la UA para el proceso de matriculación, pero destacamos el refuerzo que representa para la atención del estudiante, en términos de reducción de costes y asistencia 24/7. Esta tecnología generará un ambiente de confianza entre la UA y sus potenciales estudiantes.
Palabras clave: Tecnologías del lenguaje humano, universidad, chatbot

Abstract: This demonstration presents a chatbot desinged to automate students service in the enrolment process at the University of Alicante (UA). A chatbot in the instant messaging service Telegram, with human lenguage tecnologies and artificial inteligence, has been implemented. The main obstacle has been to offer this chatbot in the two oficial languages of Comunidad Valenciana, so automatic translation has been used as a middleware component. This chatbot offers numerous advantages to UA for the enrollment process, however we emphasize it represents a reinforcement of the student service, in terms of human cost reduction and assistance 24/7. This technology generates a trusted environment between UA and its potential students.
Keywords: Human language tecnologies, university, chatbot

1 *Introducción*

La Universidad de Alicante[1] (UA) es una universidad pública española que se encuentra en la Comunidad Valenciana. Es por ello que debe ofrecer la enseñanza en ambos idiomas. La UA tiene un total de 25.500 estudiantes matriculados en el curso actual[2], de los cuales un 6,4 % demandan docencia en Valenciano[3].

Los estudiantes generan una gran demanda de información universitaria en ambos idiomas que, en la gran mayoría de los casos, es repetitiva. Aunque la demanda es continua a lo largo del año, en la UA hay un periodo en el que se necesita responder a muchas preguntas de los estudiantes sobre las titulaciones en las que se van a matricular. Ese periodo es el de Preinscripción y Matrícula.

Así, con el objeto de atender las necesidades de información de estudiantes de la UA y sus familias en dicho periodo, el Vicerrectorado de Estudiantes y Empleo organiza un servicio específico de atención telefónica con un amplio horario continuado y atendido con líneas múltiples[4]. El servicio está en funcionamiento desde el 7 de junio hasta finales de septiembre de cada año (excepto el mes agosto) y se realiza actualmente por el personal de atención telefónica de la UA. En este servicio se atienden consultas sobre: (i) Resul-

[1] `https://www.ua.es/` (Marzo 2019)
[2] `https://utc.ua.es/es/documentos/ua-en-cifras/anual/matricula-grado.pdf` (Marzo 2019)
[3] `https://utc.ua.es/es/documentos/ua-en-cifras/anual/0956-demanda-de-valenciano-por-centro.pdf` (Marzo 2019)
[4] `https://web.ua.es/es/estudia-ua/preinscripcion-y-matricula.html` (Marzo 2019)

ISSN 1135-5948. DOI 10.26342/2019-63-29
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

tados pruebas de acceso; (ii) Preinscripción en la Comunidad Valenciana; (iii) Matrícula; (iv) Becas; (v) Procedimientos académicos y administrativos; y (v) Servicios de la Universidad de Alicante (alojamiento, transporte, deportes, etc.).

Ante esta problemática, nos proponemos desarrollar un ChatBot para automatizar la atención a los estudiantes de la UA durante el proceso de matriculación en los dos idiomas oficiales de la Comunidad Valenciana. Así, estos robots, establecerán conversaciones con los estudiantes para informar, resolver problemas y orientar su proceso de matriculación como si fuera un ser humano. Para ello es fundamental el Procesamiento del Lenguaje Natural (PLN) que permitirá que el ChatBot entienda las peticiones realizadas por los estudiantes. De esta manera, no estarán perdidos ante posibles variaciones e interpretaciones que personas hacen del lenguaje. Debido a la falta de espacio, dirigimos al lector interesado profundizar sobre los Chatbots a los análisis de Abdul-Kader y Woods (2015) y McTear, Callejas, y Griol (2016).

2 Descripción del Sistema

El ChatBot de la UA, es capaz de atender dudas sobre el proceso de matriculación a los estudiantes en la aplicación de mensajería instantánea Telegram mediante BotFather[5], que administra robots en esta aplicación.

Los estudiantes requieren una cuenta en Telegram para iniciar la consulta enviando mensajes. El acceso al bot se realiza o bien buscándolo por su nombre identificativo, *@UAOficialChatBot*, o bien directamente a través de su URL[6]. Esto permite contactar con el bot tanto desde el navegador como desde la aplicación de Telegram, y guardar el contacto para futuras consultas.

La Figura 1 muestra como el ChatBot de la UA se divide en cuatro componentes:

- **intermediario**: este componente esta implementado en Node.js[7] con la API de Telegram[8]. Se encarga de manejar la sesión de los diferentes usuarios y el idioma de cada uno de ellos. De esta forma, el bot es capaz de mantener conversaciones con varios usuarios simultáneamente en distintos idiomas.
- **robot**: Dialogflow[9] ha sido seleccionado para construir el ChatBot de la UA porque soporta muchas plataformas de mensajería, entre ellas Telegram, lo que nos permitiría añadir nuevas plataformas en el futuro. Está basado en Inteligencia Artificial y, además, dispone de interfaz gráfica y de aplicaciones. Para ello se crea un agente que transforma las peticiones de los usuarios en datos comprensibles por la máquina.
- **traductor**: este componente entra en funcionamiento cuando el usuario elige ser atendido en valenciano. Necesitamos un traductor porque Dialogflow incluye al castellano entre los idiomas soportados, pero no al valenciano[10]. Para superar este reto, empleamos Apertium (Forcada et al., 2011)[11], un traductor basado en reglas de código abierto desarrollado en la UA. El objetivo es traducir las preguntas de valenciano (lengua origen) a castellano (lengua meta), el idioma que entiende Dialogflow, cuando el módulo intermediario lo solicite para un usuario valenciano-parlante.
- **contenido**: las respuestas que debe ofrecer el ChatBot son facilitadas por el Secretariado de Acceso y Orientación Universitaria. En DialogFlow, ese contenido se debe estructurar de forma manual en '*intents*', que indican el objetivo del usuario (ej. anular matricula) junto con la respuesta que se debe ofrecer. Cada '*intent*' tiene dos versiones de cada respuesta, una por idioma. De manera que, además de la pregunta realizada por el usuario, debemos recibir el idioma escogido para comunicarse con nosotros, para responderla en la lengua esperada.

3 Caso de Estudio: Matrícula - Documentación

La Figura 2a nos muestra el mensaje de bienvenida del ChatBot en los dos idiomas oficia-

[5] `https://core.telegram.org/bots#6-botfather` (Marzo 2019)

[6] `https://telegram.me/UAOficialChatBot` (Marzo 2019)

[7] `https://nodejs.org/` (Marzo 2019)

[8] `https://github.com/yagop/node-telegram-bot-api` (Marzo 2019)

[9] `https://dialogflow.com/docs` (Marzo 2019)

[10] `https://cloud.google.com/dialogflow-enterprise/docs/reference/language` (Marzo 2019)

[11] `https://www.apertium.org/index.cat.html?dir=spa-cat#translation` (Marzo 2019)

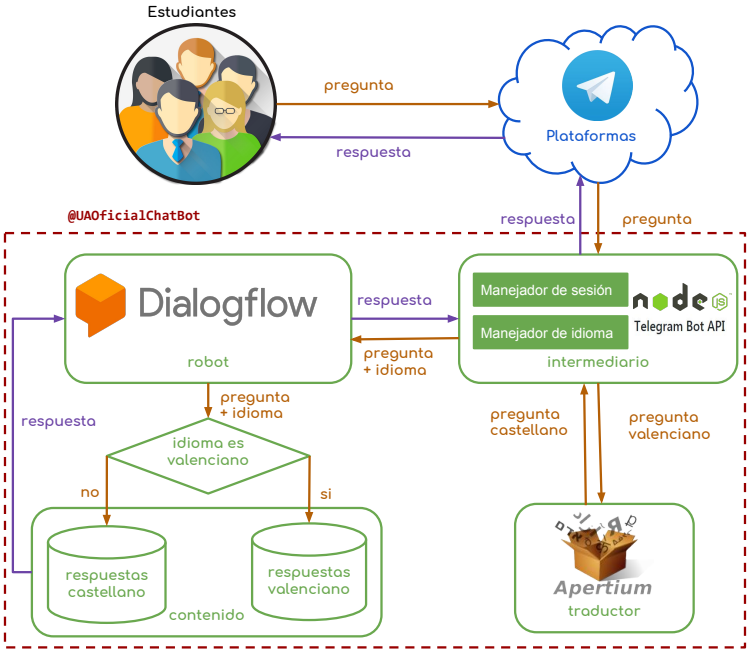


Figura 1: Arquitectura del ChatBot para la atención en matriculación

les de la Comunidad Valenciana. El estudiante primero debe seleccionar el idioma, en este ejemplo se elije castellano.

Después, puede elegir entre tres opciones: MATRICULACIÓN, HORARIOS Y NOTAS DE CORTE. En la Figura 2b se observa que la opción escogida es matriculación y que, a su vez, se divide en 7 categorías. El bot es preguntado por la documentación que deben entregar los estudiantes de nuevo ingreso y, más concretamente, aquellos que acceden a través de la Prueba de Acceso a la Universidad (PAU). La duda queda solventada, como evidencia la Figura 2c, pues el ChatBot propuesto nos especifica todos los documentos que debemos aportar.

Así la Figura 2 nos muestra una conversación breve, por razones de espacio, que permite ilustrar el correcto funcionamiento del ChatBot propuesto. Si aún tuviéramos más dudas, podríamos continuar preguntando.

4 *Conclusiones*

Hemos presentado un ChatBot cuyo propósito es informar y ayudar a los estudiantes de la UA en el proceso de matriculación. Este ChatBot está en desarrollo y todavía no está abierto al uso de todos los estudiantes.

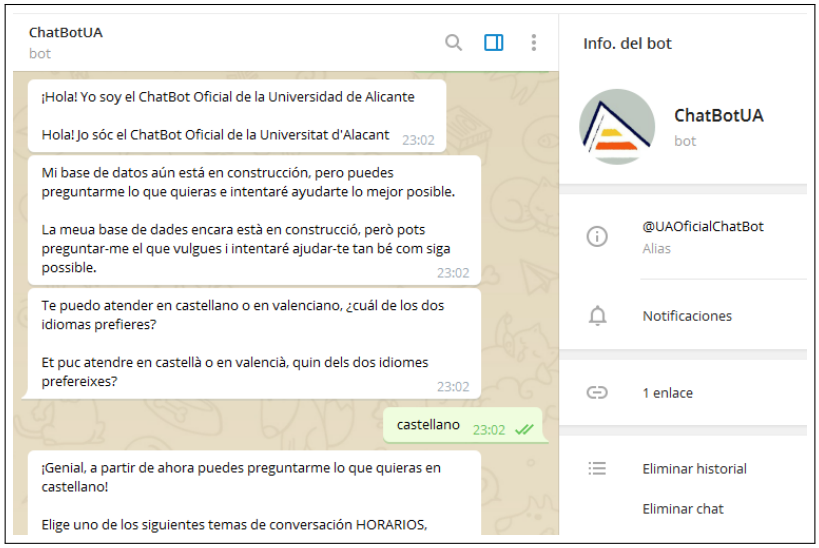
Tras una primera fase de pruebas, nuestros próximos objetivos son tres: (i) pulir las respuestas ofrecidas al usuario para hacerlas más naturales, (i) mejorar su usabilidad para hacerlo más intuitivo, y (iii) obtener la valoración del estudiante tras cada conversación, para conocer si ha sido satisfactoria la experiencia. Este último objetivo es de vital importancia para evaluar el trabajo realizado, para ello pensamos involucrar a alumnos reales y al servicio de atención al alumnado.

Con ello podremos cuantificar los numerosos beneficios que la UA puede alcanzar por tener un ChatBot para el proceso de matriculación, pero destacamos el gran refuerzo para la atención del estudiante. Puesto que sea cual sea la hora y el día del año en que el estudiante solicite información sobre nuestros títulos y su matriculación, el ChatBot podrá atenderles con total solvencia. Finalmente, no debemos olvidar que la inclusión del ChatBot puede reducir drásticamente los costos de la atención humana.

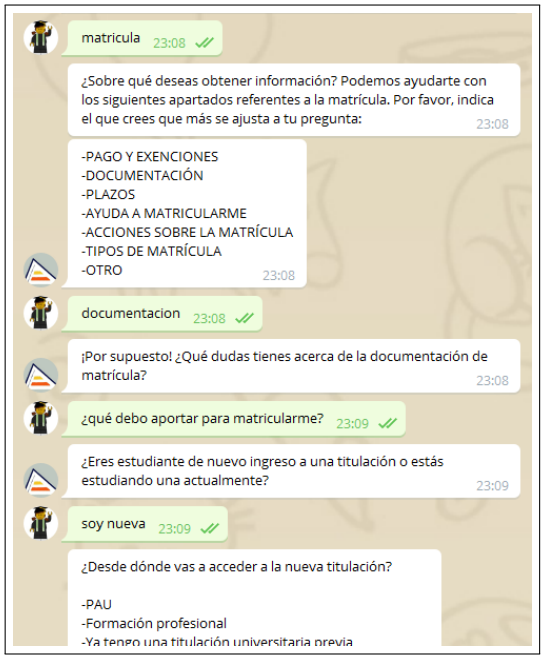
Agradecimientos

Aplicación desarrollada por el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI)[12] de la UA. Financiada por la UA, la GVA y el gobierno de España a través de los proyectos “Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante (ENCARGOINTERNO5-19EN)”, “SIIA: Tecnologías del Lenguaje Humano para una Sociedad Inclusiva Igualitaria y Accesible (PROMETEU/2018/089)”, “MODELA-

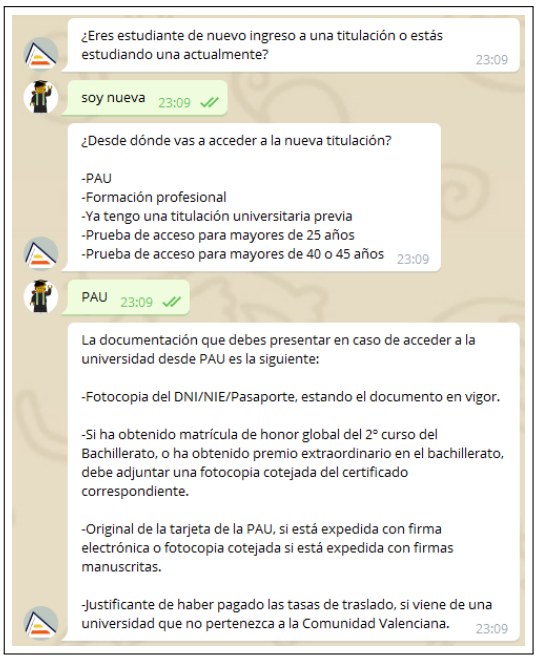
[12] `https://gplsi.dlsi.ua.es/` (Marzo 2019)



(a) Bienvenida del ChatBot en dos idiomas



(b) Solicitud de información sobre documentación a presentar en la matricula de nuevos estudiantes



(c) Solicitud de información sobre documentación para nuevos estudiantes accediendo con la PAU

Figura 2: Ejemplo que ilustra el funcionamiento en Telegram

DO DEL COMPORTAMIENTO DE ENTIDADES DIGITALES MEDIANTE TECNOLOGIAS DEL LENGUAJE HUMANO (RTI2018-094653-B-C22)" e "INTEGER: Intelligent Text Generation, GENERACION INTELIGENTE DE TEXTOS (RTI2018-094649-B-I00)". Agradecer a Nuria Salazar García, Víctor Manuel Belén Rivera, Sergio Conejero Vicente y Mateo Linás Ghaín su apoyo técnico (PRA-18/2018-2019; PRA-28/2018-2019).

Bibliografía

Abdul-Kader, S. A. y D. J. Woods. 2015. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, 6(7).

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, y F. M. Tyers. 2011. Free/open-source machine translation: preface. *Machine Translation*, 25(2):83–86.

McTear, M., Z. Callejas, y D. Griol, 2016. *Conversational Interfaces: Past and Present*, páginas 51–72. Springer International Publishing, Cham.

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 201-204 recibido 28-03-2019 revisado 07-05-2019 aceptado 14-05-2019

Social-Univ 2.0: Monitorización del entorno digital social universitario

Social-Univ 2.0: Monitoring the Digital and Social Environment of the University

Isabel Moreno[1,2], Javi Fernández[2], Yoan Gutiérrez[1]
[1]Instituto Universitario de Investigación Informática, Universidad de Alicante
[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{imoreno, javifm, ygutierrez}@dlsi.ua.es

Resumen: Social-Univ 2.0 es una aplicación web que monitoriza el ecosistema digital de universidades con presencia en redes sociales y, más concretamente, se centra en analizar el entorno de la Universidad de Alicante. Usa las tecnologías del lenguaje humano para extraer meta-datos sobre comentarios en Twitter para representar perfiles de actores sociales, así como su relación social en un período de tiempo. Esta aplicación, en su primera versión: facilita tanto el estudio de interacciones sociales y sus matices, como el análisis de grupos sociales y sus comportamientos; identifica a actores sociales relevantes e influyentes, así como los temas comunes en determinados grupos sociales; y favorece la medición del impacto de campañas y eventos.
Palabras clave: Tecnologías del lenguaje humano, universidad, relaciones sociales

Abstract: Social-Univ 2.0 is a web application that monitors the university digital ecosystem in social networks and, specifically, it focuses on analysing the University of Alicante environment. It uses human language technologies to extract meta-data from tweets and represents actors' profiles, as well as their social relations in a time period. This application, in its first version: eases the study of social interactions and its nuances, as well as the analysis of social groups and their behaviour; identifies relevant and influential social actors as well as common topics in groups; and helps to measure impacts from campaigns and events.
Keywords: Human language technologies, university, social relationships

1 Introducción

Millones de personas de todo el mundo participan en las redes sociales. Las interacciones de los usuarios ya no pasan únicamente por consumir información, si no que también expresan su opinión sobre dicha información. Si analizamos Twitter, una de las redes sociales más conocidas, observamos que dispone de más de 344 millones de usuarios activos[1], que generan 500 millones de tweets al día[2].

Existen infraestructuras analíticas, como Atribus[3] o SumAll[4], que automáticamente recuperan y procesan información de las redes sociales para ser analizada por usuarios. Sin embargo, si nos enfocamos en las universidades, a día de hoy, no encontramos infraestructuras que analicen el impacto que estas tienen en las redes sociales. Por tanto, si deseáramos identificar los roles de todos aquellos actores y eventos que pudieran incidir en la reputación de una universidad a nivel social y global, necesitaríamos hacerlo de forma (semi-)manual y emplear para ello cuantiosos recursos, humanos y temporales.

Ante la falta de tecnologías que resuelvan esta problemática, la herramienta *Social-Univ 2.0* tiene como objetivo principal el uso de Tecnologías del Lenguaje Humano (TLH) para extraer comentarios de las redes sociales, así como sus meta-datos, y representar tanto perfiles de actores sociales (empresas, instituciones, personas, etc.) como la relación social que existe entre ellos y una universidad en un determinado período de tiempo.

2 Descripción del Sistema

Social-Univ 2.0 analiza comentarios de la red social Twitter y relaciona actores mediante los mensajes en los que intervienen. Pa-

[1] `https://bit.ly/1pIwiJ8` (Marzo 2019)
[2] `https://bit.ly/1pJzREY` (Marzo 2019)
[3] `http://www.atribus.com` (Febrero 2019)
[4] `https://sumall.com` (Febrero 2019)

ISSN 1135-5948. DOI 10.26342/2019-63-30
© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

ra iniciar el estudio social, el sistema requiere a un actor de partida (i.e. una universidad) que se debe proporcionar según la forma autóctona de identificador en redes sociales (ej. @UA_Universidad) o nombre (ej. Universidad de Alicante).

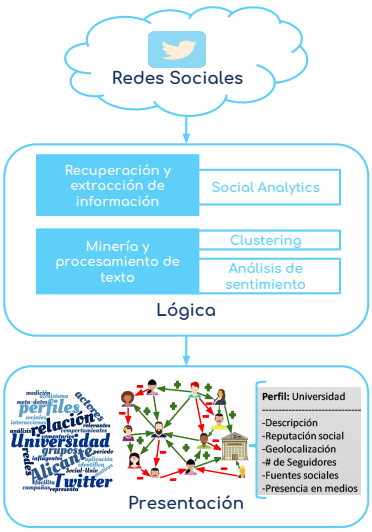


Figura 1: Arquitectura de *Social-Univ 2.0*

La Figura 1 muestra como *Social-Univ 2.0* se divide en dos capas. Por un lado, la capa lógica, detallada en la Sección 2.1, se encarga de descargar comentarios de Twitter y analizarlos semánticamente para valorar las diferentes posturas que cada actor tiene respecto a otros, así como su influencia en el entorno digital-social. Por otro lado, la capa de presentación, detallada en la Sección 2.2, muestra al usuario la información social analizada mediante grafos, perfiles y nubes de palabras en una interfaz que permite filtrar el análisis por periodos inferiores a 6 meses.

2.1 Capa de lógica de la aplicación

Esta capa hace uso de las TLH y, como puede verse en la Figura 1, se divide en dos módulos:

El **módulo Recuperación y extracción de información** se encarga de descargar el contenido generado por el usuario en Twitter, en inglés y español, que mencione a la universidad a estudiar. Este proceso se realiza reutilizando la plataforma Social Analytics (Fernández et al., 2017) que permite recuperar el contenido de la Web 2.0 y extraer meta-datos sobre cualquier entidad digital.

El **módulo Minería y procesamiento de texto** analiza el contenido generado por el usuario para determinar: (i) Cuáles son los términos relevantes y contabilizarlos con técnicas de clusterización usando el algoritmo *Lingo*, con los parámetros por defecto, de la herramienta Carrot2[5] (Stefanowski y Weiss, 2003). (ii) Cuál es la geolocalización de contenido y del autor, buscando la población de origen de los usuarios en la base de datos de Geonames[6], de forma similar a Fernández et al. (2017). Y (iii) si es una opinión, cuál es su polaridad, basándonos en la aproximación de Fernández, Gómez, y Martínez-Barco (2014) y Fernández et al. (2017). Esto último es clave para valorar las diferentes posturas (valoración, opinión, etc.) que cada actor tiene de otros y su influencia en un entorno social digital. Finalmente, se obtienen modelos de representación de perfiles de actores sociales y su relación en un período de tiempo.

2.2 Capa de presentación

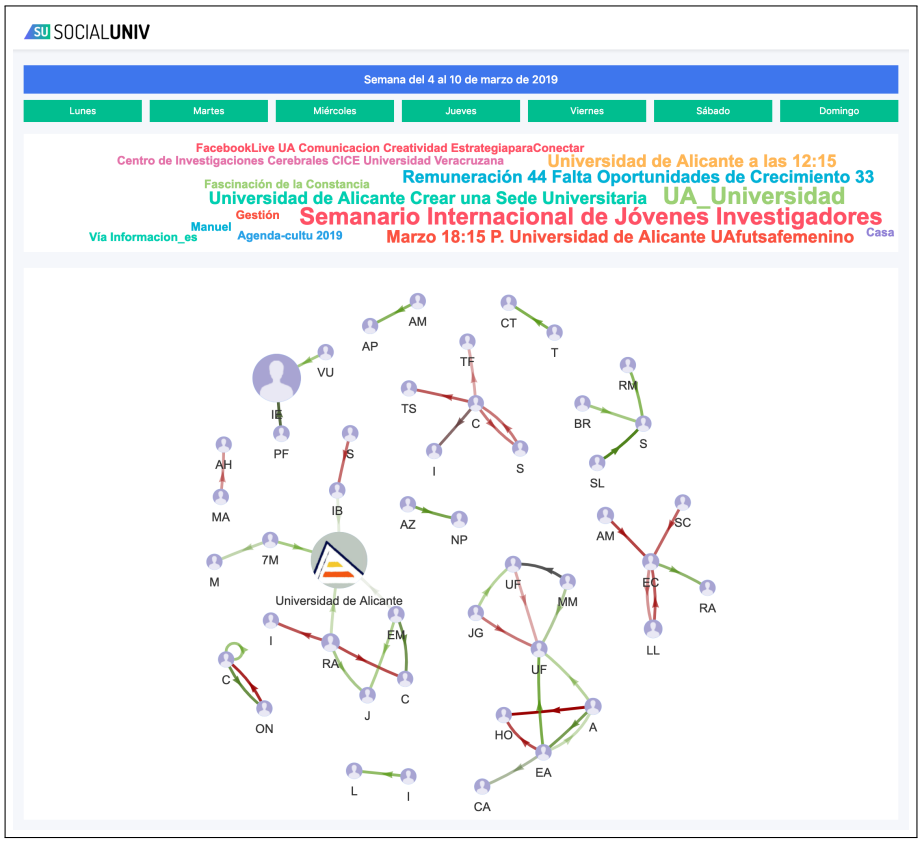
Esta capa genera un informe mediante HTML, PHP y Javascript que visualiza el análisis de la capa lógica (Sección 2.1) de un modo útil y sencillo para ser interpretado por usuarios finales (gestores universitarios). El informe contiene:

1. Selector de fechas para analizar periodos de tiempo (día, semana o mes).
2. Nube de palabras relevantes clusterizadas que se muestra con D3-Cloud[7].
3. Grafos sociales y conjuntos, donde se representan los actores sociales y sus relaciones mediante la herramienta Vis[8]. Las relaciones se establecen a través de los mensajes recuperados, que se clasifican y ponderan como positivas o negativas según el contenido del mensaje.
4. Perfil de cada actor con: descripción, geolocalización, páginas web relacionadas, número de seguidores en Twitter (de su API pública[9]) y reputación social en una escala de 1 a 10 para el periodo seleccionado.

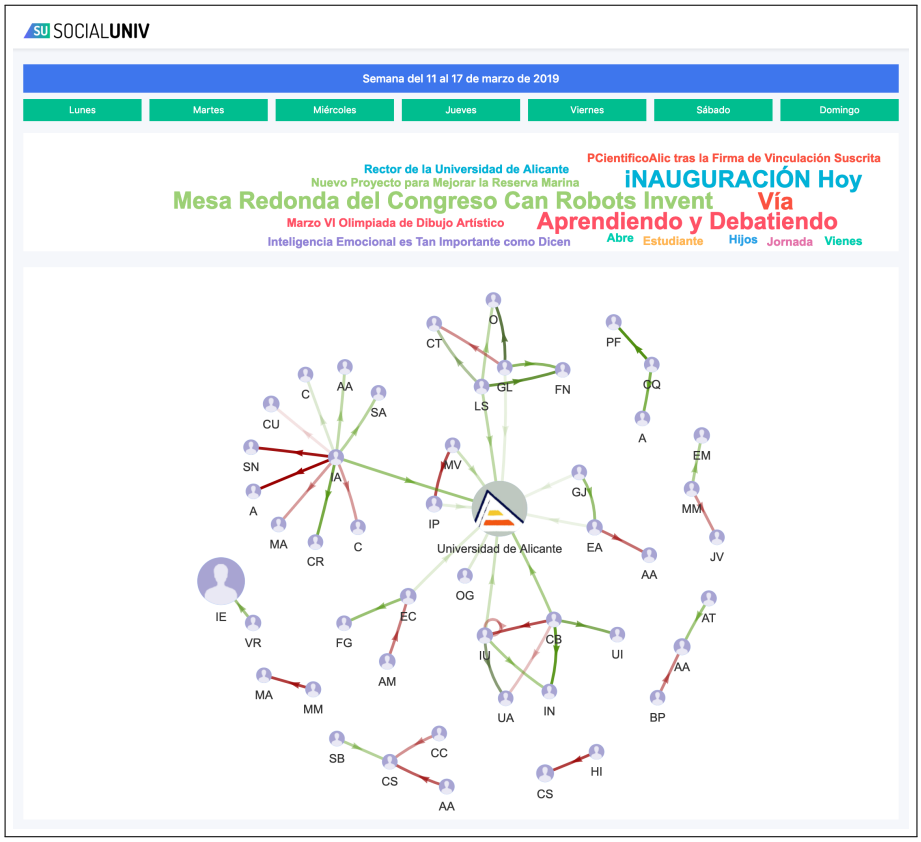
3 *Caso de Estudio: Universidad de Alicante*

Una vez elegida la universidad a monitorizar, comienza a realizarse el seguimiento en

[5] `https://project.carrot2.org` (Marzo 2019)
[6] `https://www.geonames.org/` (Marzo 2019)
[7] `https://github.com/jasondavies/d3-cloud` (Marzo 2019)
[8] `https://github.com/almende/vis` (Marzo 2019)
[9] `https://developer.twitter.com` (Marzo 2019)



(a) Grafo y nubes de palabras de la semana 10 al 17 de marzo de 2019



(b) Grafo, nubes de palabras de la semana 4 al 10 de marzo de 2019

Figura 2: Ejemplo anonimizado que ilustra el funcionamiento de SocialUniv 2.0

Figura 3: Perfil de la Universidad de Alicante

las redes sociales. En este caso, se ha creado una instancia de la aplicación para la Universidad de Alicante[10](UA). A medida que se van encontrando diferentes publicaciones mencionando a la UA, se genera una nube de palabras que recoge la forma de expresarse de los actores involucrados durante un periodo de tiempo. Por ejemplo, en la Figura 2a vemos que, del 4 al 10 de marzo de 2019, el "*Semanario Internacional de Jóvenes Investigadores*" recibió atención los días previos a su celebración (2 de abril de 2019). Otro ejemplo, durante el periodo 11 al 17 de marzo de 2019, lo encontramos en la Figura 2b y vemos que se produjo el "*congreso Can Robots Invent* and Create", organizado por la Facultad de Derecho de la UA, del 14 al 15 de marzo.

También se construye un grafo social que relaciona a los actores relevantes en un lapso de tiempo. Continuando con la Figura 2b, se observa como se produce una influencia positiva (flecha verde) con todos los actores sociales que interactúan con la UA; o como el Equipo de fútbol sala femenino de la UA (usuario UF) tiene tanto influencias negativas (flechas rojas) con 2 usuarios, como positivas con 3, como se aprecia en la Figura 2a. Por último, si hacemos clic sobre un actor del grafo, accedemos a su perfil. La Figura 3 muestra el perfil de la UA con 97.2 mil seguidores y una reputación de 8.5 sobre 10.

Todo ello permite analizar a la UA, a sus órganos, y cualquier otro actor que interactué con ella. Además, posibilita observar la evolución social digital de la UA para que, como actores sociales o representantes de dicha entidad, consigamos visualizar grafos cada vez más poblados (i.e. más actores debaten sobre la UA) y donde el color dominante sea el verde (i.e. mejor reputación social).

4 Conclusiones

Hemos presentado la primera versión de *Social-Univ 2.0*, una herramienta que extrae comentarios de la red social Twitter sobre una universidad y sus meta-datos. Con esta tecnología, hemos estudiado el entorno social de la UA y las entidades sociales digitales con las que se relaciona. Mediante el grafo social generado por *SocialUniv*, hemos analizado las interacciones sociales entre grupos de usuarios y sus comportamientos en diferentes momentos o detectado a los actores sociales relevantes e influyentes. Además, la nube de palabras nos ha permitido identificar la expresión social en los comentarios analizados. Por último, el perfil ha permitido caracterizar al actor principal de este estudio, la UA.

El próximo paso es evaluar extrínsecamente Social-Univ 2.0 para medir su utilidad como un todo. En un futuro, podría adaptarse a otras redes sociales, a aplicaciones móviles y al campus virtual de universidades españolas.

Agradecimientos

Aplicación desarrollada por el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI)[11] de la UA. Financiada por la UA a través del proyecto "Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante (ENCARGOINTERNO5-19EN)", conjuntamente con la Generalitat Valenciana (PROMETEU/2018/089) y el gobierno de España (RTI2018-094653-B-C22 y RTI2018-094649-B-I00).

Bibliografía

Fernández, J., J. M. Gómez, y P. Martínez-Barco. 2014. A supervised approach for sentiment analysis using skipgrams. En *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, páginas 30–36. Association for Computational Linguistics.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58:141–148.

Stefanowski, J. y D. Weiss. 2003. Carrot 2 and language properties in web search results clustering. En *International Atlantic Web Intelligence Conference*, páginas 240–249. Springer.

[10] `https://www.ua.es/` (Marzo 2019)

[11] `https://gplsi.dlsi.ua.es/` (Marzo 2019)

Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 205-208 recibido 31-03-2019 revisado 08-05-2019 aceptado 13-05-2019

Terminological and language resources for developing a virtual patient dialogue system in Spanish

Recursos lingüísticos y terminológicos para desarrollar un sistema de diálogo con un paciente virtual en español

Leonardo Campillos-Llanos[1,2], **Catherine Thomas**[1,2], **Éric Bilinski**[1], **Pierre Zweigenbaum**[1], **Sophie Rosset**[1]
[1]LIMSI - CNRS
[2]SATT Paris Saclay
{campillos, bilinski, pz, rosset}@limsi.fr

Abstract: This article reports the terminology and language resources used to develop a dialogue system for the medical domain in Spanish. The system simulates the medical history-taking step with a virtual patient, and is aimed mainly at medical students, who can train their communication and anamnesis skills in e-learning contexts. The conversational agent was first developed in French, and only the French system was evaluated by potential end-users. Herein, we provide an overall description of the system with a special focus on the Spanish components.
Keywords: Sistemas de diálogo, Terminología, PLN en Medicina

Resumen: Este artículo presenta los recursos lingüísticos y terminológicos empleados para desarrollar un sistema de diálogo de dominio médico en español. El sistema simula la etapa de entrevista clínica con un paciente virtual, y se destina principalmente a estudiantes de Medicina, quienes pueden entrenar sus destrezas comunicativas y de anamnesis en contextos de enseñanza a distancia. El agente conversacional se desarrolló principalmente en francés, y solo la versión francesa fue evaluada por potenciales usuarios. En esta contribución, describimos de manera global el sistema, con un foco especial en los componentes para la lengua española.
Palabras clave: Dialogue systems, Terminology, Medical NLP

1 *Introduction*

Dialogue systems for health applications are a vibrant field of research. Literature reviews (López-Cózar et al., 2014; Laranjo et al., 2018; Montenegro, da Costa, and da Rosa Righi, 2019) show how applications are progressively widening their scope from telemedicine, patient follow-up and counselling, to virtual assistants for clinicians to get information from databases. Conversational agents are also used to interact with avatars in educational software for training healthcare professionals by means of virtual patients (VP) (Rombauts, 2014).

In a collaborative project with companies,[1] we collaborated in the creation of a virtual patient e-learning software for medical doctors. The software allows medical instructors to define a patient profile—with its health problems and medical history—so that medical students can practice their diagnosis skills. Our contribution was to create a dialogue system for the medical history taking step: medical doctors can train their anamnesis skills through a conversational agent, which was developed in French, English and Spanish. Figure 1 shows a sample of dialogue with the Spanish system.

Creating such a dialogue system brought up several challenges, namely: 1) To manage adequately the terminological variation in a domain of large vocabulary; 2) To provide reliable answers—correct according to the patient record—and realistic replies—to be expected from a patient perspective. In former work (Campillos-Llanos et al., 2019), we already developed on those aspects and reported the system architecture. Herein, we summarize its overall functioning (§2), explain the adaptation of the French version to

[1] `https://pvdial.limsi.fr/`

ISSN 1135-5948. DOI 10.26342/2019-63-31 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

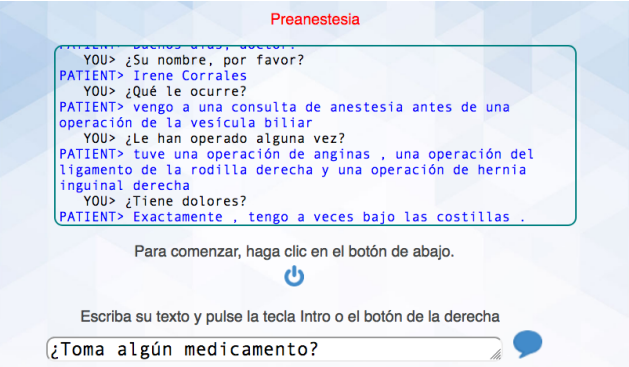


Figure 1: Sample of dialogue in Spanish

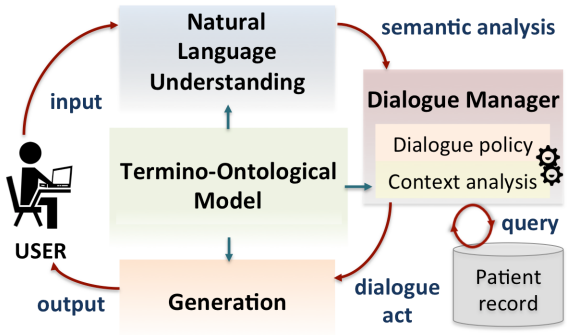


Figure 2: Schema of the dialogue system

the Spanish language (§3), and we describe the resources of the Spanish version (§4-§6). We conclude by reporting the current state of the system and future work (§7).

2 *Dialogue System Architecture*

The system is made up of different units (Figure 2), integrated in an infrastructure for dialogue systems (Rosset et al., 2005):

- A **Natural Language Understanding (NLU)** module: performs linguistic and semantic processing of user's input.
- A **Dialogue Manager**: determines the dialogue policy, processes the information state at each dialogue move (according to the previous dialogue history) and queries the virtual patient record.
- A **Linguistic and Termino-ontological model**: manages the linguistic and terminological variation of the dialogue task.
- A **Generation module**: replies to the user by means of templates to be filled with the contents of the VP record.

3 *Overall procedure of adaptation from French to Spanish*

The dialogue manager did not need any language-specific adaptation, but the other components needed language resources to cope with the processing steps in the Spanish language. The NLU module makes use of a general dictionary, and domain lists, which were collected following the same procedures as those applied for the French components. We extracted medical terms from the Unified Medical Language System® (hereafter, UMLS) according to semantic contents needed for the dialogue task; we summarize them in §4. We also translated lists from French to Spanish when no data could be extracted from domain terminologies; e.g. lay terms expressed in patient-language (e.g. *tener colesterol* instead of *hipercolesterolemia*); or general, out-of-domain lists (e.g. expressions of frequency or duration). Translation was enhanced with semi-automatic procedures and manually revised for quality check. Regular expressions and rules were translated manually.

As for the Termino-Ontological model, we collected equivalent resources to those needed in the French version (see §5). The Generation module needed both: 1) Collecting lexical resources from existing dictionaries and terminologies in Spanish; and 2) Translating manually the templates for generating the replies to be filled with the contents of the patient record. Finally, to build a small set of VP records for development, we translated from French 13 cases used for evaluating the French system. The translation of French to Spanish was rather feasible and fast (around 6 months). Both languages are close and have available terminological or language resources of similar type and size for the task.

4 *Resources for Natural Language Understanding*

This stage involves the following steps:

- **Linguistic preprocessing**: tokenization and Part-of-Speech (PoS) tagging with TreeTagger (Schmid, 1995). We use a dictionary of 627,099 entries for lemmatizing general and domain terms.
- **Spelling correction**: a spell-checker corrects the input or asks the user to confirm the correction of a misspelling.

- **Semantic annotation**: we use rules developed in Wmatch (Galibert, 2009), an engine for applying regular expressions using gazetteers. We created 149 NLU labels for annotating entity types (e.g. `symptom`), dialogue acts (e.g. `greeting`) and question types (e.g. `Qwhere`).

For each of the 149 labels, we formalized a grammar by means of regular expressions, rules of sentence structures and gazetteers for Named Entity Recognition. The NLU labels cover the range of topics of the medical history taking: patient's demographic data (e.g. name and age), medical history (e.g. diseases or surgeries), symptoms, medication intake, as well as lifestyle and social behavior. Some rules manage conversational acts (e.g. greetings), but out-of-the-task and out-of-domain acts are not covered (e.g. a question such as *What is your favourite film?*).

Lists of entities were extracted from the UMLS (Bodenreider, 2004), vs. 2017AA; or compiled manually. Lists of medical drugs come from the Spanish Drug Effect database (Segura-Bedmar et al., 2015) and the Agencia Española de Medicamentos y Productos Sanitarios.[2]

5 Resources for the Linguistic and Termino-Ontological Model

This component gathers structured thesauri for managing linguistic and terminological variation: 1) To match input terms against the content of the patient record (*concept normalization*); and 2) To favour replies using non-technical terms. To query the patient record, entities detected in the user's input are lemmatized and lowercased.

Linguistic variation requires managing inflectional and derivational variants, roots/affixes/stems, and general synonyms. **Inflectional variants** (e.g. *estornuda* ↔ *estornudo*) come from a lexicon gathering 45,763 word entries and 475,652 forms. **Derivational variants** (e.g. *fiebre* ↔ *febril*) and general **synonyms** (e.g. *andar* ↔ *caminar*) were obtained from the Spanish EuroWordNet in the Multilingual Central Repository 3.0 (Gonzalez, Laparra, and Rigau, 2012): for derivation, we used relations RELATED_TO, IS_DERIVED and PERTAINS_TO; and for synonymy, the relation SYNONYM. **Roots, stems and affixes** were translated from the lexicon used in the French version of the dialogue system, and also from English roots in the Specialist Lexicon (Browne, McCray, and Srinivasan, 2000).

To cater for **terminological variation** and map domain terms, we used the UMLS Metathesaurus, which makes it possible to map medical terms referring to the same concept, identified by a Concept Unique Identifier (CUI). Terms were collected from semantic type T121 (Pharmacological substance, e.g. *calmante* ↔ *analgésico*, C0002771, 2.5.1) and 3 semantic groups: ANAT, for anatomic entities (e.g. *vientre* ↔ *abdomen*, C0022658); DISO, for syndromes and disorders (e.g. *enfermedad de riñón* ↔ *nefropatía*, C0022658); and PROC, for procedures (e.g. *operación de hernia* ↔ *herniorrafía*, C0019328). We clustered in *ad hoc* lists domain synonym terms not found in the UMLS: e.g. symptom verbs and nouns (*adelgazar* ↔ *adelgazamiento*).

Lastly, we gathered **semantic relations** between CUIs from the UMLS Semantic Network (McCray, 1989). This helps the system to match input terms to concepts in the VP record beyond synonymy, when a hierarchical relation exists. For example, CHILD_OF relations may match a general entity to a specific entity (e.g. *enfermedad cardiovascular* ↔ *hipertensión*). We also use associative relations to link procedures and disorders (e.g. *herniorrafía* ↔ *hernia*), procedures and anatomic entities (e.g. *apendicectomía* ↔ *apéndice*), and pathologies and physiological entities (e.g. *vómito* ↔ *digestión*). These relations hold between concepts (not terms), hence they are language-independent.

6 Resources for Generation

This step involves three subtasks:

- **Gender/number agreement**: information extracted from the VP record is output through predefined templates, where nouns or adjectives need to agree. To generate the number and gender variant forms of the words to be output, we use linguistic information from DELAS-type dictionaries (Courtois, 1990).

- **Change of the content in the record to the patient's viewpoint**: in the generation step, data expressed in 3rd person is changed to the 1st person (e.g. *[el paciente] tose* → *[yo] toso*); we use a

[2] `www.aemps.gob.es/`

table with correspondences between 3rd and 1st person verb forms.

- **Reply with non-technical terms**: we favour system replies according to a patient's perspective (e.g. *anginas* instead of *amígdalas*); to do so, we use lists of equivalent terms.

7 Discussion and Conclusion

The tool is one of the few dialogue systems in Spanish for such a task in the medical domain. The system aims at filling the need for improving doctors interaction skills by means of an e-learning tool as a complement to real clinical practice. Developing the system required us to tackle challenges of the task and domain—e.g. managing terminological variation, generate realistic output, favouring correctness in system replies. All of those aspects revitalize an area of research with problems still not solved satisfactorily.

Our experience shows that adapting the system to a close language is feasible with the methods explained. This could spark interest in adapting the system to other Romance languages, provided that similar language and terminological resources exist for the task.

To date, only the French version of the system was evaluated by potential end-users (i.e. medical students and doctors, n=39). We conducted a quantitative evaluation—through the analysis of users' dialogue logs—and qualitative user evaluation—by means of 5-point Likert-scale questionnaires, collected with an on-line form which is also available for assessing the Spanish system. We aim at collecting more data to improve the Spanish version and evaluate it with medical students or residents. For that purpose, we prepared 13 different patient records by translating some cases used in the evaluation of the French system. The system can be tested and evaluated on-line.[3]

Acknowledgements

The SATT Paris Saclay funded this work.

References

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Browne, A. C., A. T. McCray, and S. Srinivasan. 2000. The Specialist Lexicon. *National Library of Medicine Technical Reports*, pages 18–21.

Campillos-Llanos, L., C. Thomas, E. Bilinski, P. Zweigenbaum, and S. Rosset. 2019. Designing a Virtual Patient Dialogue System Based on Terminology-rich Resources: Challenges and Evaluation. *Natural Language Engineering*. `https://doi.org/10.1017/S1351324919000329`.

Courtois, B. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Lang. française*, 87(1):11–22.

Galibert, O. 2009. *Approaches and methodologies for automatic Question-Answering in an open-domain, interactive setup*. Phd dissertation, Université Paris Sud, June.

Gonzalez, A., E. Laparra, and G. Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Laranjo, L. et al. 2018. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inf. Assoc.*, 25(9):1248–1258.

López-Cózar, R., Z. Callejas, D. Griol, and J. F. Quesada. 2014. Review of spoken dialogue systems. *Loquens*, 1(2):012.

McCray, A. T. 1989. The UMLS Semantic Network. In *Proc. Symp. on Comp. Applic. Medical Care*, pages 503–507. AMIA.

Montenegro, J. L. Z., C. A. da Costa, and R. da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*.

Rombauts, N. 2014. *Patients virtuels: pédagogie, état de l'art et développement du simulateur Alphadiag*. Phd dissertation, Faculty of Medicine, Lyon, France.

Rosset, S., O. Galibert, G. Illouz, and A. Max. 2005. Interaction et recherche d'information: le projet Ritel. *Traitement automatique des langues*, 46(3):155–179.

Schmid, H. 1995. Treetagger—a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Segura-Bedmar, I., P. Martínez, R. Revert, and J. Moreno. 2015. Exploring Spanish health social media for detecting drug effects. In *BMC Medical Informatics and Decision Making*, volume 15, page S6.

[3] `www.audiosurf.net/pg_2018/select_case.php`

Información General

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 8 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/la-revista/informacion-para-autores

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

L. Alfonso Ureña López (Director)
Universidad de Jaén
laurena@ujaen.es
Patricio Martínez Barco (Secretario)
Universidad de Alicante
patricio@dlsi.ua.es
Manuel Palomar Sanz
Universidad de Alicante
mpalomar@dlsi.ua.es
Felisa Verdejo Maillo
UNED
felisa@lsi.uned.es

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández Rodríguez	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Ramón López-Cózar Delgado	Universidad de Granada (España)

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Antònia Martí Antonín	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Eugenio Martínez Cámara	Universidad de Granada (España)
Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lluís Padró Cirera	Universidad Politécnica de Cataluña (España)
Manuel Palomar Sanz	Universidad de Alicante (España)
Ferrán Pla Santamaría	Universidad Politécnica de Valencia (España)
German Rigau Claramunt	Universidad del País Vasco (España)
Horacio Rodríguez Hontoria	Universidad Politécnica de Cataluña (España)
Paolo Rosso	Universidad Politécnica de Valencia (España)
Leonel Ruíz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Kepa Sarasola Gabiola	Universidad del País Vasco (España)
Encarna Segarra Soriano	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé Delor	Universidad de Barcelona
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignon/Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares Ferro	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
René Venegas Velásquez	Universidad Católica de Valparaiso (Chile)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Si desea inscribirse como socio de la Sociedad Española del Procesamiento del Lenguaje Natural puede realizarlo a través del formulario web que se encuentra en esta dirección http://www.sepln.org/sepln/la-sociedad
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/la-revista/consejo-de-redaccion
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/la-revista/consejo-asesor
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/sepln/inscripcion-para-nuevos-socios

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

Proyectos

MENHIR: mental health monitoring through interactive conversations
Zoraida Callejas, Kawtar Benghazi, Manuel Noguera, María Inés Torres, Raquel Justo 139
Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad
Alejandro Molina-Villegas, Oscar S. Siordia, Edwing Aldana-Bobadilla, César Aguilar, Olga Acosta . 143
LIHLITH: Learning to Interact with Humans by Lifelong Interaction with Humans
Eneko Agirre, Arantxa Otegi, Camille Pradel, Sophie Rosset, Anselmo Peñas y Mark Cieliebak..... 147
Unsupervised neural machine translation, a new paradigm solely based on monolingual text
Mikel Artetxe, Gorka Labaka, Eneko Agirre 151
Project CAVIAR CApturing VIewers' Affective Response
Fernando Fernández-Martínez, Zoraida Callejas, Ricardo Kleinlein, Cristina Luna Jiménez, Juan Manuel Montero, José Manuel Pardo 155
LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos
Itziar Aldabe, Josu Aztiria, Francho Beltrán, Myriam Bras, Klara Ceberio, Itziar Cortes, Jean-Baptiste Coyos, Benaset Dazeas, Luise Esher, Gorka Labaka, Igor Leturia, Kepa Sarasola, Aure Séguier, Jean Sibille..... 159
Prevención de enfermedades infecciosas basado en el análisis inteligente en RRSS y participación ciuidadana
Óscar Apolinario, José Medina-Moreira, Harry Luna-Aveiga, José Antonio García-Díaz, Rafael Valencia-Garcia, José Ignacio Estrade-Cabrera 163
Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del plan de Impulso de las tecnologías del lenguaje
Antonio Moreno, Doroteo Torre, Ana Valverde, Leonardo Campillos 167
EXTRAE: EXTRacción de Asociaciones entre Enfermedades y otros conceptos médicos
Lourdes Araujo, Juan Martínez-Romo, Andrés Duque, Fernando López-Ostenero, Ricardo Sánchez de Madariaga, Adolfo Muñoz Carrero, Mario Pascual Carrasco 171
Enriched communication across the lifespan
Martin Cooke, Simon King, Valerie Hazan, Yannis Stylianou, Esther Janse, Deniz Baskent, Volker Hohmann, Axel Winneke, Inma Hernaez 175
Social-Univ 2.0: tecnologías del lenguaje humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante
Isabel Moreno, Javi Fernández, Yoán Gutiérrez..... 179

Demostraciones

Asistente conversacional móvil para la rehabilitación de la tartamudez
David Griol, Javier Font, Zoraida Callejas 185
Plataforma web para la evaluación de la comprensión lectora aplicada a tareas de búsqueda
Javier Perea de la Casa, Fernando Javier Martínez Santiago, Miguel Ángel García Cumbreras, Manuel Carlos Díaz Galiano 189
Corpus viewer: NLP and ML-based platform for public policy making and implementation
David Pérez-Fernández, Jerónimo Arenas-García, Doaa Samy, Antonio Padilla-Soler, Vanesa Gómez-Verdejo 193
Atención automatizada a estudiantes en el proceso de matriculación en la Universidad de Alicante
Isabel Moreno, Yoan Gutierrez, Andrés Montoyo 197
Social-Univ 2.0: monitorización del entorno digital social universitario
Isabel Moreno, Javi Fernández, Yoán Gutiérrez..... 201
Terminological and language resources for developing a virtual patient dialogue system in Spanish
Leonardo Campillos-Llanos, Catherine Thomasa, Éric Bilinski, Piere Zwigenbaum, Sophie Rosset 205

Información General

Información para los autores 211
Información adicional..... 213