Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Grammatical error correction for Basque through a seq2seq neural architecture and synthetic examples
Zuhaitz Beloki, Xabier Saralegi, Klara Ceberio, Ander Corral 13
Un estudio de los métodos de reducción del frente de Pareto a una única solución aplicado al problema de resumen extractivo multi-documento
Jesús M. Sánchez-Gómez, Miguel A. Vega-Rodríguez, Carlos J. Pérez 21
Generación de frases literarias: un experimento preliminar
Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, Roseli S. Wedmann 29
Cross-lingual training for multiple-choice question answering
Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas 37
ContextMEL: Classifying contextual modifiers in clinical text
Paula Chocron, Álvaro Abella, Gabriel de Maeztu 45
Limitations of neural networks-based NER for resume data extraction
Juan F. Pinzón, Stanislav Krainikovsky, Roman S. Samarev 53
Minería de argumentación en el referéndum del 1 de octubre de 2017
Marcos Esteve Casademunt, Paolo Rosso, Francisco Casacuberta 59
Edición de un corpus digital de inventarios de bienes
Pilar Arrabal Rodríguez 67
Relevant content selection through positional language models: an exploratory analysis
Marta Vicente, Elena Lloret 75
Rantanplan, fast and accurate syllabification and scansion of Spanish poetry
Javier de la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, Elena González-Blanco 83

Proyectos

COST action “European network for web-centred linguistic data science”
Thierry Declerck, Jorge Gracia, John P. McCrae 93
MINTZAI: Sistemas de aprendizaje profundo E2E para traducción automática del habla
Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Álvarez, Inma Hernaez, Eva Navas, Ander González, Jaime Osácar, Edson Benites, Igor Ellakuria, Eusebi Calonge, Maite Martín 97
MISMIS: Misinformation and Miscommunication in social media: aggregating information and analysing language
Paolo Rosso, Francisco Casacuberta, Julio Gonzalo, Laura Plaza, J. Carrillo-Albornoz, E. Amigó, M. Felisa Verdejo, Mariona Taulé, Maria Salamó, M. Antònia Martí 101
AMALEU: a machine-Learned universal language representation
Marta R. Costa-jussà 105
Transcripción, indexación y análisis automático de declaraciones judiciales a partir de representaciones fonéticas y técnicas de lingüística forense
Antonio García-Díaz, Ángela Almela, Fernando Molina, Juan Salvador Castejón-Garrido, Rafael Valencia-García 109

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2020

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

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)
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)
René Venegas Velásquez	Pontificia Universidad Católica de Valparaíso (Chile)
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

Mario Almagro	Universidad Nacional de Educación a Distancia (España)
Laura Alonso Alemany	Universidad Nacional de Córdoba (Argentina)
Marco Casavantes	Universidad Autónoma de Chihuahua (México)
Víctor Manuel Darriba Bilbao	Universidad de Vigo (España)
Luis Espinosa-Anke	University of Cardiff (Reino Unido)
Juan Luis García Mendoza	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Horacio Jarquín	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Pilar López Úbeda	Universidad de Jaén (España)
Soto Montalvo	Universidad Rey Juan Carlos (España)
Arturo Montejo Ráez	Universidad de Jaén (España)
Flor Miriam Plaza del Arco	Universidad de Jaén (España)
Fracisco J. Ribadas-Pena	Universidad de Vigo (España)
Juan Javier Sánchez Junquera	Universidad Politécnica de Valencia (España)

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 65 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). 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 37 trabajos para este número, de los cuales 22 eran artículos científicos y 15 correspondían a resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas. De entre los 22 artículos recibidos, 10 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 45,4%.

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 2020
Los editores.

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 65th 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.

Thirty-seven papers were submitted for this issue, from which twenty-two were scientific papers and fifteen were either projects or tool description summaries. From these twenty-two papers, we selected ten (45,4%) 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, as long as it is equal or greater than 5 out of 7.

September 2020
Editorial board.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Grammatical error correction for Basque through a seq2seq neural architecture and synthetic examples
Zuhaitz Beloki, Xabier Saralegi, Klara Ceberio, Ander Corral 13
Un estudio de los métodos de reducción del frente de Pareto a una única solución aplicado al problema de resumen extractivo multi-documento
Jesús M. Sánchez-Gómez, Miguel A. Vega-Rodríguez, Carlos J. Pérez 21
Generación de frases literarias: un experimento preliminar
Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, Roseli S. Wedmann 29
Cross-lingual training for multiple-choice question answering
Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas 37
ContextMEL: Classifying contextual modifiers in clinical text
Paula Chocron, Álvaro Abella, Gabriel de Maeztu 45
Limitations of neural networks-based NER for resume data extraction
Juan F. Pinzón, Stanislav Krainikovsky, Roman S. Samarev 53
Minería de argumentación en el referéndum del 1 de octubre de 2017
Marcos Esteve Casademunt, Paolo Rosso, Francisco Casacuberta 59
Edición de un corpus digital de inventarios de bienes
Pilar Arrabal Rodríguez 67
Relevant content selection through positional language models: an exploratory analysis
Marta Vicente, Elena Lloret 75
Rantanplan, fast and accurate syllabification and scansion of Spanish poetry
Javier de la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, Elena González-Blanco 83

Proyectos

COST action “European network for web-centred linguistic data science”
Thierry Declerck, Jorge Gracia, John P. McCrae 93
MINTZAI: Sistemas de aprendizaje profundo E2E para traducción automática del habla
Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Álvarez, Inma Hernaez, Eva Navas, Ander González, Jaime Osácar, Edson Benites, Igor Ellakuria, Eusebi Calonge, Maite Martín 97
MISMIS: Misinformation and Miscommunication in social media: aggregating information and analysing language
Paolo Rosso, Francisco Casacuberta, Julio Gonzalo, Laura Plaza, J. Carrillo-Albornoz, E. Amigó, M. Felisa Verdejo, Mariona Taulé, Maria Salamó, M. Antònia Martí 101
AMALEU: a machine-Learned universal language representation
Marta R. Costa-jussà 105
Transcripción, indexación y análisis automático de declaraciones judiciales a partir de representaciones fonéticas y técnicas de lingüística forense
Antonio García-Díaz, Ángela Almela, Fernando Molina, Juan Salvador Castejón-Garrido, Rafael Valencia-García 109

Demostraciones

EmoCon: Analizador de emociones en el congreso de los diputados
Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, María Teresa Martín-Valdivia, Andrea López-Fernández 115
Nalytics: natural speech and text analytics
Ander González-Docasal, Naiara Pérez, Aitor Álvarez, Manex Serras, Laura García-Sardiña, Haritz Arzelus, Aitor García-Pablos, Montse Cuadros, Paz Delgado, Ane Lazpiur, Blanca Romero 119
RESIVOZ: dialogue system for voice-based information registration in eldercare
Laura García-Sardiña, Manex Serras, Arantza del Pozo, Mikel D. Fernández-Bhogal 123

The text complexity library
Rocío López Anguita, Jaime Collado Montañez, Arturo Montejo Ráez 127
The impact of coronavirus on our mental health
Jiawen Wu, Francisco Rangel, Juan Carlos Martínez 131
AREVA: augmented reality voice assistant for industrial maintenance
Manex Serras, Laura García-Sardiña, Bruno Simões, Hugo Álvarez, Jon Arambarri 135
UMUCorpusClassifier: compilation and evaluation of linguistic corpus for natural language processing tasks
José Antonio García-Díaz, Ángela Almela, Gema Alcaraz-Mármol, Rafael Valencia-García 139

Información General
Información para los autores 145
Información adicional 147

Artículos

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 13-20 recibido 02-04-2020 revisado 01-06-2020 aceptado 01-06-2020

Grammatical Error Correction for Basque through a seq2seq neural architecture and synthetic examples

Corrección gramatical para euskera mediante una arquitectura neuronal seq2seq y ejemplos sintéticos

Zuhaitz Beloki, Xabier Saralegi, Klara Ceberio, Ander Corral
Elhuyar Foundation
{z.beloki, x.saralegi, k.ceberio, a.corral}@elhuyar.eus

Abstract: Sequence-to-sequence neural architectures are the state of the art for addressing the task of correcting grammatical errors. However, large training datasets are required for this task. This paper studies the use of sequence-to-sequence neural models for the correction of grammatical errors in Basque. As there is no training data for this language, we have developed a rule-based method to generate grammatically incorrect sentences from a collection of correct sentences extracted from a corpus of 500,000 news in Basque. We have built different training datasets according to different strategies to combine the synthetic examples. From these datasets different models based on the Transformer architecture have been trained and evaluated according to accuracy, recall and F0.5 score. The results obtained with the best model reach 0.87 of F0.5 score.
Keywords: GEC, Seq2seq architectures, Basque, Less-resourced languages

Resumen: Las arquitecturas neuronales secuencia a secuencia constituyen el estado del arte para abordar la tarea de corrección de errores gramaticales. Sin embargo, su entrenamiento requiere de grandes conjuntos de datos. Este trabajo estudia el uso de modelos neuronales secuencia a secuencia para la corrección de errores gramaticales en euskera. Al no existir datos de entrenamiento para este idioma, hemos desarrollado un método basado en reglas para generar de forma sintética oraciones gramaticalmente incorrectas a partir de una colección de oraciones correctas extraídas de un corpus de 500.000 noticias en euskera. Hemos construido diferentes conjuntos de datos de entrenamiento de acuerdo a distintas estrategias para combinar los ejemplos sintéticos. A partir de estos conjuntos de datos hemos entrenado sendos modelos basados en la arquitectura Transformer que hemos evaluado y comparado de acuerdo a las métricas de precisión, cobertura y F0.5. Los resultados obtenidos con el mejor modelo alcanzan un F0.5 de 0.87.
Palabras clave: Corrección gramatical, Arquitecturas seq2seq, Euskera

1 Introduction

The task of *Grammatical Error Correction (GEC)* consists in detecting and correcting grammatical errors in a sentence, resulting in a grammatically correct sentence (e.g. *"I give her a book yesterday" → "I gave her a book yesterday"*). This is a task that arouses great interest within Natural Language Processing, and proof of this is the large number of works on GEC that we can find in the literature.

Initially, this task was addressed with relative success through symbolic approaches based on linguistic rules. Later, different authors introduced methods based on machine learning that can be divided into two groups: a) methods based on supervised classification, and b) methods based on machine translation. The methods based on supervised classification consist of building a classifier for each type of grammatical error from annotated corpora (Izumi et al., 2003; Gamon, 2010). Methods

ISSN 1135-5948. DOI 10.26342/2020-65-1

based on machine translation, on the other hand, address the task of grammatical correction as a problem of machine translation, or sequence-to-sequence learning (*seq2seq*), where the source sentences correspond to grammatically incorrect sentences and the target sentences to grammatically correct ones. These methods are more efficient than methods based on supervised classification, as their ability to deal with errors that follow complex patterns is greater (Rozovskaya and Roth, 2016).

The first methods based on machine translation used statistical machine translation models (Brockett et al., 2006). Lately, they have been replaced by superior neural models (Chollampatt and Ng, 2018; Grund-kiewicz and Junczys-Dowmunt, 2018) given the great capacity of the latter to generalize patterns.

However, neural translation models require even more training data than statistical translation models. This is a major problem, because even for widely used languages such as English, there are no large training corpora available for the GEC task. For this reason, several authors (Zhao et al., 2019; Lichtarge et al., 2019) propose to address the task of grammatical error correction as a *low-resource Neural Machine Translation task*, where the automatic generation of synthetic training data is essential. Different techniques have been proposed for generating synthetic data, such as techniques based on back-translation (Reiet al., 2017; Ge et al., 2018).

The problem of the lack of training data is even greater in languages with limited digital resources, where there are not even initial annotated corpora that can serve as a basis for generating synthetic training examples. This is the case of the Basque language, which is the case of the study that we propose in this paper.

Until now, the correction of grammatical errors in Basque texts has been tackled through rule-based strategies (Oronoz, 2009) with the limitations that this approach entails. In this paper we propose to address the problem by using *seq2seq* neural models. To this end, we have evaluated *seq2seq* neural models trained from different synthetic training corpora. Various training corpora were generated, by introducing grammatical errors derived from several rules into correct sentences. Four strategies were used to combine the various types of incorrect examples generated. Correct sentences were extracted from a large collection of news gathered from digital newspapers. The main contributions of this work are the following:

- To the best of our knowledge, this is the first work that studies the use of neural *seq2seq* models in the task of correcting grammatical errors in Basque.
- We propose a new rule-based method for generating synthetic training corpus oriented to the correction of grammatical errors in Basque.
- We provide a new benchmark for the GEC task in Basque, by making the synthetic training and evaluation datasets publicly available[1].

The article is structured as follows. In the following section we will explain the methodology we have followed to select the grammatical errors included in the study. In section 3 we describe in detail the methods proposed for the generation of synthetic training datasets from grammatically correct texts. In section 4 we will present the *seq2seq* neural architecture used to implement the grammatical correction system. Section 5 will discuss the results obtained with the systems trained on the basis of the different synthetic training datasets generated. We will conclude this article by presenting the main conclusions drawn from the work and also the future work planned.

2 Selection of grammatical errors

This work focuses on the most common grammatical errors in Basque texts. As we do not have a corpus of Basque texts with grammatical errors annotated, we have chosen to consult a professional translator/corrector and a professional lexicographer.

Each of them has been asked to select ten errors from the list of 33 grammatical errors proposed for Basque by Oronoz (2009).

[1] https://hizkuntzateknologiak.elhuyar.eus/assets/files/elh-gec-eu.tgz

Specifically, they had to select those errors which in their opinion are most common in Basque texts regardless of the register and domain.

In addition to their own judgment, these experts used as additional material the list of most common errors in the exams for the EGA title (*Euskararen Gaitasun Agiria*, or in English: Certificate of Proficiency in Basque).

From the intersection of the two sets of ten errors selected by both experts, six errors resulted (Table 1), of which two (erroneous use of suffixes in dates and times) were discarded because they are easily solved by rule-based techniques (Oronoz, 2009). Thus, the four errors selected for this paper are the following:

- E1: Wrong use of the verb tense or aspect. For example, the use of the verbal form of the present in a future context.
- E2: Misuse of the verbal paradigm. The verbal system in Basque consists of four paradigms: *nor* (monovalent intransitive), *nor-nork* (bivalent transitive), *nor-nori* (bivalent intransitive), *nor-nori-nork* (trivalent transitive). Each verb is conjugated according to its corresponding paradigm(s). It is a very common error to use the wrong paradigm.
- E3: Lack of concordance between the verb and the subject. Confusion of the declension suffix in the subject.
- E4: Misuse of the verbal suffix. Completive sentences in Basque are formed by adding the suffix (*-(e)la*) to the verb of the subordinate sentence. If the sentence is negative, the suffix should be *-(e)nik*.

Error-type	*Examples*
E1: Verb tense	*Ziur bihar **jakiten*** (→jakingo) *dugula. (I'm sure we'll find out tomorrow)* *Gustura egingo **nuen*** (→nuke) *orain. (I'd love to do it now)* *Gauza bat **faltatzen*** *(*→falta*) zait esateko. (There's one more thing I have to say)*
E2: Verbal paradigm	*Afaltzera gonbidatu **zidan*** (→ninduen). *(He/She invited me to dinner)* *Atzo kalean ikusi **nizun*** (→zintudan). *(I saw you yesterday on the street)* *Utzi behar **dugu*** (→diogu) *negar **egitea*** (→egiteari). *(We have to stop crying)*
E3: Concordance verb-subject	***Jon*** (→Jonek) *ez daki ezer. (Jon doesn't know anything)* ***Bidaiak*** (→Bidaiek) *atsedena hartzeko balio dute. (Travelling is good to rest)* *Jende askok uste **dute*** (→du). *(A lot of people think).*
E4: Completive sentences	*Ez dut uste hori egia **dela*** (→denik). *(I don't think that's true)* *Nire ustez, hori horrela **dela*** (→da). *(I think it's like that)* *Badago beste kutsadura bat **dela*** (→dena) *nuklearra. (There's another contamination which is nuclear contamination)*

Table 1: Selected errors and examples. In brackets, the English translation of the corrected example

3 Generation of synthetic datasets

3.1 Generation of synthetic errors

The size of the different training datasets available for GEC is insufficient for training *seq2seq* neural models, so different techniques for the generation of additional synthetic training data have been proposed in the literature. Some authors (Grundkiewicz y Junczys-Dowmunt, 2014; Lichtarge et al., 2019) have proposed to extract training data from Wikipedia revision histories, a source from which a large number of examples can be extracted, especially for English. Other authors (Reiet al., 2017; Ge et al., 2018) generate synthetic training data following the *back-translation* strategy proposed for machine translation systems. An intermediate model is trained from the initial training corpus to be applied to a corpus of correct sentences, and thus sentences with grammatical errors are automatically generated. Another alternative proposed in the literature to generate synthetic training corpora is to introduce "noise" in a corpus of correct texts. The "noise" or grammatical errors are introduced by means of

linguistic rules (Yuan and Felice, 2013), or more generic operations of token replacement, elimination, insertion or reordering (Zhao et al., 2019).

In our case, to generate the synthetic training corpus we will follow an approach based on linguistic rules, in line with the strategy adopted by Yuan and Felice (2013). However, unlike Yuan and Felice (2013), we will not use an initial annotated corpus as a reference.

The rules are designed to generate specific grammatical errors in grammatically correct sentences. That way, we can generate pairs of incorrect and correct sentences useful for compiling a training dataset.

The implemented rules (Table 2) generate errors of the types E1, E2, E3, E4 (Table 1) described in subsection 3.1. For each type of error a set of rules has been implemented so that most of the possible cases are covered. In some cases the application of the rule is bi-directional depending on whether the error generated in that way is also common.

The changes executed by the implemented rules consist of replacing specific words depending on the type of error (examples shown in Table 3). These replacements are made according to certain grammatical information and specific tokens that we obtain through the morphosyntactic analyzer for the Basque language Eustagger (Ezeiza et al., 1998):

- Rule R1.1 associated with the error type E1 is applied to sentences where the verb tense is future (suffixes *"ko"* and *"go"*) and the verb inflection is modified to transform it into present tense (suffixes *"ten"* and *"tzen"*).
- Rules R2.1, R2.2, R2.3, R2.4 associated with error type E2 modify the auxiliary verb to simulate the most common verbal paradigm confusions.
- Rule R3.1 associated with error type E3 modifies the subject's grammatical case (its declension) to transform ergative cases into absolutive ones.
- Rules R4.1, R4.2, R4.3 associated with error type E4 modify the auxiliary verb to simulate suffix errors in completive sentences.

Error	*Rules*
E1	R1.1: ko/go → ten/tzen
E2	R2.1: nor-nork → nor-nori-nork R2.2: nor-nori-nork ↔ nori-nor R2.3: nor-nork ↔ nor R2.4: nor-nork ↔ nori-nor
E3	R3.1: subj_erg → sub_abs
E4	R4.1: v_aux(-nik) → v_aux(-la) R4.2: v_aux(-na) →v_aux(-la) R4.3: v_aux(-laren) ↔v_aux(-lako)

Table 2: Rules for errors associated with selected grammatical errors

Rule	*Examples*
R1.1	Arratsaldean **ikusiko** (→*ikusten*) gara. (See you in the afternoon)
R2.1	Atzo hondartzan ikusi **zintudan** (→*nizun*). (I saw you on the beach yesterday)
R2.2	Aholku kontrajarriak ematen ari **zaigu** (→digu). (He/She is giving us contradictory advice)
R2.3	Azkenaldian asko argaldu **du** (→*da*). (He's lost a lot of weight lately)
R2.4	Paisaia asko gustatzen **zait** (→*nau*). (I really like the landscape)
R3.1	**Nik** (→*ni*) ez dut nahi. (*I don't want it).* **Langileek** (→*langileak*) lan handia egin dute. (The workers have worked very hard)
R4.1	Ez dut uste etorriko **denik** (→*dela*). (I don't think he/she's coming)
R4.2	Badago beste arazo bat zehaztasuna **dena** (→*dela*). (There's another problem that is precision)
R4.3	Laster zabaldu da denok gaixotuko **garelaren** (→*garelako*) albistea. (The news that we're all going to get sick has spread fast.)

Table 3: Examples of errors generated by the implemented rules. In brackets, the English translation of the correct example

3.2 Strategies for building datasets

Each example in the training -and evaluation- datasets we create is composed of a sentence pair that includes a sentence containing grammatical errors and its corresponding corrected version. In order to generate those pairs we apply the rules described in the previous subsection over grammatically correct sentences. We also add pairs composed of the

original unmodified sentences so that trained models also take those cases into account.

Those grammatically correct sentences are extracted from a news corpus compiled from several Basque news websites: Berria.eus, Argia.eus and the various proximity media of the Tokikom.eus network. The collected corpus consists of 500,015 news items from which we extract 4,927,748 correct sentences $O_c=\{oc_i\}$ including 66 million words.

To generate the training datasets (Tables 4 and 5) we have analyzed different strategies to apply the rules on a subset of O_c of 4,921,748 sentences:

- Baseline (D_{t0} dataset): We apply to each correct sentence oc_i a variation of the substitution rule proposed by Zhao et al. (2019), since *seq2seq* models trained on data generated using the original method performed very poorly. The original method consists in replacing, with a 10% probability, each word with another randomly selected word from the corpus. To avoid highly artificial sentences, our variation adds the constraint that the bigram (w_1, w_2) exists in the corpus, where w_2 is the randomly selected word and w_1 is the word preceding the original replaced word. For each correct sentence we generate the pair ($z(oc_i)$,oc_i) composed of the incorrect sentence $z(oc_i)$ and the correct oc_i, in addition to the unmodified pair (oc_i,oc_i).
- Strategy 1 (D_{t1} dataset): From the rules that can be applied to the correct sentence oc_i, a single R_i rule is randomly selected and the pair ($R_i(oc_i)$,oc_i) is generated in addition to the unmodified pair (oc_i,oc_i). If no rule can be applied, only the pair (oc_i,oc_i) is generated.
- Strategy 2 (D_{t2} dataset): For each R_i rule that can be applied to the correct oc_i sentence, the pair ($R_i(oc_i)$,oc_i) is generated, in addition to the unmodified pair (oc_i,oc_i). If no rule can be applied, only the pair (oc_i,oc_i) is generated.
- Strategy 3 (D_{t3} dataset): We apply a set of defined rules to each correct sentence oc_i. From the rules that can be applied, a set $\{R_j\}$ of n rules is selected at random, and applied sequentially to generate the pairs ($R_1(oc_i) \circ \ldots \circ R_n(oc_i)$,$oc_i$) and ($oc_i$,$oc_i$). If no rule can be applied, only the pair (oc_i,oc_i) is generated.

In the different training datasets created we differentiate three types of example pairs according to the number of rules applied for their generation: a) *None*: they are not the result of any rule, b) *Single*: they are the result of applying one rule, c) *Multi*: they are the result of applying several rules (Table 5).

	Pairs	*R1*	*R2*	*R3*	*R4*
D_{t0}	8.49M	-	-	-	-
D_{t1}	9.33M	0.37M	3.43M	0.54M	0.07M
D_{t2}	17.38M	1.04M	9.66M	1.60M	0.19M
D_{t3}	9.33M	0.59M	3.84M	0.89M	0.10M
D_{ea}	6000	662	2924	871	1291
D_{em}	672	49	307	92	143

Table 4: Number of total pairs (*None* included) and pairs generated by each rule included in the training (D_{t0}, D_{t1}, D_{t2}, D_{t3}) and evaluation (D_{ea}, D_{em}) datasets

	Pairs	*None*	*Single*	*Multi*
D_{t0}	8.49M	-	-	-
D_{t1}	9.33M	4.92M	4.41M	0
D_{t2}	17.38M	4.91M	12.47M	0
D_{t3}	9.33M	4.92M	3.17M	1.24M
D_{ea}	6000	2000	2000	2000
D_{em}	672	250	221	201

Table 5: Total number of pairs (*Pairs*) and pairs according to typology (*None*, *Single*, *Multi*) included in the training (D_{t0}, D_{t1}, D_{t2}, D_{t3}) and evaluation (D_{ea}, D_{em}) datasets

To create the evaluation datasets we use the same strategies as those used to create the training datasets, but on a different subset (6k correct sentences) of O_c. We guarantee a balance between the pair types *None*, *Single* and *Multi*. In this way, we generate a first evaluation dataset D_{ea} fully automatically. Taking into account that, in some cases, the application of the rules can generate grammatically correct sentences, we also built another D_{em} evaluation dataset consisting of a subset of 750 D_e pairs but reviewed manually. Pairs generated by the rules that do not really include grammatical errors are eliminated. 78 pairs were discarded in the manual review

process, leaving a subset of 672 pairs (see tables 4 and 5).

4 Seq2seq architecture for GEC

The *seq2seq* neural architectures are being used successfully to address the GEC task. Unlike statistical translation models, *seq2seq* neural architectures can model dependencies between words (or similar word sets) that are critical in correcting grammatical errors (Sakaguchi et al., 2016).

In the literature, we can distinguish three main sequence-to-sequence architectures proposed for the correction of grammatical errors: architectures based on recurrent neural networks (Ge et al., 2018), architectures based on convolutional networks (Chollampatt and Ng, 2018), and architectures based on self-attention (Junczys-Dowmunt et al., 2018). These last ones are the ones we are going to use in this work, since they are the ones that provide better results in this task according to Zhao et al. (2019).

For training the grammatical correction models we have chosen the Transformer architecture (Vaswani et al., 2017). Specifically, we have used the implementation of the OpenNMT-py library. The Transfomer architecture is based on an encoder-decoder system with an attention mechanism. Both the encoder and the decoder are composed of 6 layers composed in turn by a recurrent neural network and a mechanism of attention. We have used the default values of the architecture without any optimization of the parameters. The size of the recurrent neural network of each layer is 512. Thus, 512 size embeddings have been used for both incorrect and corrected sentences. The Adam optimizer has been used during the training, and a learning-rate of 2 with a warm-up phase of 8000 steps. The dropout ratio is 0.1, the batch size is 4096 sentences, and the models have been trained until the results on the development set have not shown any improvement. For the development set 5000 sentence pairs have been selected randomly from the training data.

To avoid the open vocabulary issue and for a better translation of unknown words, BPE tokenization (Sennrich et al., 2016) has been applied to source and target sequences. Rare or unseen words are represented as a sequence of subword units. In the case of Basque, this encoding is particularly useful as declensions generate a larger vocabulary.

5 Results

We present results for four GEC systems. All of them are based on the Transformer model introduced in the previous section and trained on the synthetic datasets presented in section 3. Those systems were evaluated according to the standard metrics used in GEC: precision, recall and $F_{0.5}$ with respect to the set of edits needed to correct the incorrect sentences. The upperbound would be an oracle system that makes only the necessary edits to correct the errors included in the incorrect sentences. The following systems were built and evaluated:

- D_{t0}+*tr system*: Training of the Transformer model from the training dataset D_{t0} (synthetic examples by random word replacement).
- D_{t1}+*tr system*: Training of the Transformer model from the training dataset D_{t1} (Synthetic examples by application of a rule by sentence).
- D_{t2}+*tr*: Training of the Transformer model from the training dataset D_{t2} (synthetic examples by application of n rules per sentence)
- D_{t3}+*tr*: Training of the Transformer model from the training dataset D_{t3} (synthetic examples by simultaneous application of *n* rules per sentence)

Tables 6, 7, 8 and 9 show the results obtained for each of the systems with respect to the evaluation datasets, both the automatic D_{ea} and the manually reviewed D_{em}.

The best results are obtained by the D_{t3}+*tr* system, on both evaluation datasets and also on the *Single* (sentences with one error) and *Multi* (sentences with more than one error) subsets, as well as on the four types of errors (E1, E2, E3 and E4). The Transformer model seems able to better learn the task from examples that can combine more than one error, which is the configuration of the D_{t3} training dataset. Error analysis revealed that D_{t3}+*tr* works well with sentences with more than one error, but tends to make incorrect fixes in sentences with no errors or containing a single error.

The D_{t0}+*tr* system trained from the baseline dataset has a very low performance, and points out that the generic replacement rules are not adequate to generate synthetic training datasets, at least in this case study. The model does not have enough information to perform the necessary fixes. It does not create new mistakes, but neither corrects them.

The results of the D_{t1}+*tr* and D_{t2}+*tr* systems differ slightly from each other, the former being better. But the results of both are notably lower than those of D_{t3}+*tr*, especially in terms of recall. This difference in performance with respect to D_{t3}+*tr* is especially accentuated (see table 7) when dealing with sentences with more than one error (*Multi*). These systems rarely solve more than one error in the same sentence.

With regard to the different types of errors, there are no major differences, and in general better results are obtained for types E1 and E2 (see tables 8 and 9).

	D_{ea}			D_{em}		
	P	*R*	$F_{0.5}$	*P*	*R*	$F_{0.5}$
D_{t0}+*tr*	0.26	0.02	0.07	0.26	0.02	0.07
D_{t1}+*tr*	0.86	0.57	0.78	0.88	0.58	0.80
D_{t2}+*tr*	0.83	0.56	0.75	0.80	0.60	0.75
D_{t3}+*tr*	**0.88**	**0.73**	**0.85**	**0.90**	**0.76**	**0.87**

Table 6: Precision (P), recall (R) and $F_{0.5}$ of the systems with respect to the automatic (D_{ea}) and manually reviewed (D_{em}) datasets

	D_{ea}			D_{em}		
	S	*M*	*S*+*M*	*S*	*M*	*S*+*M*
D_{t0}+*tr*	0.08	0.06	0.07	0.06	0.09	0.07
D_{t1}+*tr*	0.81	0.79	0.80	0.84	0.81	0.82
D_{t2}+*tr*	0.82	0.77	0.79	0.83	0.78	0.80
D_{t3}+*tr*	**0.83**	**0.88**	**0.86**	**0.86**	**0.90**	**0.89**

Table 7: $F_{0.5}$ results of the systems with respect to the *Single* (*S*) and *Multi* (*M*) subsets of the evaluation datasets

	E1	*E2*	*E3*	*E4*
D_{t0}+*tr*	0.07	0.07	0.04	0.06
D_{t1}+*tr*	0.81	0.81	0.74	0.77
D_{t2}+*tr*	0.79	0.79	0.72	0.78
D_{t3}+*tr*	**0.88**	**0.87**	**0.86**	**0.85**

Table 8: $F_{0.5}$ results of the systems with respect to the automatic evaluation dataset D_{ea} depending on the grammatical error to correct

	E1	*E2*	*E3*	*E4*
D_{t0}+*tr*	0.07	0.07	0.01	0.06
D_{t1}+*tr*	0.88	0.83	0.76	0.79
D_{t2}+*tr*	0.85	0.79	0.72	0.82
D_{t3}+*tr*	**0.94**	**0.90**	**0.90**	**0.87**

Table 9: $F_{0.5}$ results of the systems with respect to the manually revised dataset D_{em} depending on the grammatical error to be corrected

6 Conclusions

In this work we have been able to prove that it is possible to implement a grammar checker based on *seq2seq* neural models for a less-resourced language, represented by Basque in this case of study. For this type of language where no training data is available for the GEC task, we have found that a strategy based on building synthetic training datasets from monolingual corpora is feasible. The proposed method, based on combining different linguistic rules to generate grammatical errors, allows the creation of large valid datasets to train high performance *seq2seq* neuronal models. In the future, we plan to extend the repertoire of linguistic rules for generating synthetic errors, and also study other methods of synthetic data generation, in order to include more types of grammatical errors in the *seq2seq* model.

References

Brockett, C., W. B. Dolan and M. Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 249-256).

Chollampatt, S. and H.T. Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ezeiza, N., I. Alegria, J.M. Arriola, R. Urizar and I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 380-384).

Gamon, M. 2010. Using mostly native data to correct errors in learners' writing: a meta-classifier approach. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 163-171).

Ge, T., F. Wei and M. Zhou. 2018. Fluency boost learning and inference for neural grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1)* (pp. 1055-1065).

Grundkiewicz, R. and M. Junczys-Dowmunt. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing* (pp. 478-490).

Izumi, E., K. Uchimoto, T. Saiga, T. Supnithi and H. Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics* (pp. 145-148).

Junczys-Dowmunt, M., R. Grundkiewicz, S. Guha and K. Heafield. 2018. Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 595-606).

Lichtarge, J., C. Alberti, S. Kumar, N. Shazeer, N. Parmar and S. Tong. 2019. Corpora Generation for Grammatical Error Correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (pp. 3291-3301).

Oronoz, M. 2009. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura* (Doctoral dissertation, Universidad del País Vasco-Euskal Herriko Unibertsitatea).

Rei, M. and H. Yannakoudakis. 2017. Auxiliary Objectives for Neural Error Detection Models. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 33-43).

Rozovskaya, A. and D. Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2205-2215).

Sakaguchi, K., C. Napoles, M. Post and J. Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, *4*, 169-182.

Sennrich, R., B. Haddow, and A. Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task* (pp. 371-376).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Yuan, Z. and M. Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 52-61).

Zhao, W., L. Wang, K. Shen, R. Jia and J. Liu. 2019. Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 156-165).

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 21-28 recibido 06-03-2020 revisado 14-06-2020 aceptado 14-06-2020

Un estudio de los métodos de reducción del frente de Pareto a una única solución aplicado al problema de resumen extractivo multi-documento

A study of methods for reducing the Pareto front to a single solution applied to the extractive multi-document summarization problem

Jesus M. Sanchez-Gomez[1], **Miguel A. Vega-Rodríguez**[1], **Carlos J. Pérez**[2]
[1]Dpto. de Tecnología de Computadores y Comunicaciones, Universidad de Extremadura
[2]Dpto. de Matemáticas, Universidad de Extremadura
jmsanchezgomez@unex.es, mavega@unex.es, carper@unex.es

Resumen: Los métodos de resumen automático son actualmente necesarios en muchos contextos diferentes. El problema de resumen extractivo multi-documento intenta cubrir el contenido principal de una colección de documentos y reducir la información redundante. La mejor manera de abordar esta tarea es mediante un enfoque de optimización multi-objetivo. El resultado de este enfoque es un conjunto de soluciones no dominadas o conjunto de Pareto. Sin embargo, dado que solo se necesita un resumen, se debe reducir el frente de Pareto a una única solución. Para ello, se han considerado varios métodos, como el mayor hipervolumen, la solución consenso, la distancia más corta al punto ideal y la distancia más corta a todos los puntos. Los métodos han sido probados utilizando conjuntos de datos de DUC, y han sido evaluados con las métricas ROUGE. Los resultados revelan que la solución consenso obtiene los mejores valores promedio
Palabras clave: Resumen extractivo multi-documento, Optimización multi-objetivo, Frente de Pareto, Solución única

Abstract: Automatic summarization methods are currently needed in many different contexts. The extractive multi-document summarization problem tries to cover the main content of a document collection and to reduce the redundant information. The best way to address this task is through a multi-objective optimization approach. The result of this approach is a set of non-dominated solutions or Pareto set. However, since only one summary is needed, the Pareto front must be reduced to a single solution. For this, several methods have been considered, such as the largest hypervolume, the consensus solution, the shortest distance to the ideal point, and the shortest distance to all points. The methods have been tested using datasets from DUC, and they have been evaluated with ROUGE metrics. The results show that consensus solution achieves the best average values
Keywords: Extractive multi-document summarization, Multi-objective optimization, Pareto front, Single solution

1 Introducción

Hoy en día, la información en Internet crece exponencialmente, y obtener la más importante sobre un tema concreto es de interés en muchas áreas. Extraer la información más relevante es posible mediante las herramientas de minería de texto, las cuales son capaces de generar un resumen automático a partir de información textual (Hashimi, Hafez, y Mathkour, 2015).

Existen varios tipos de resumen. Dependiendo de dónde se obtiene la información, un resumen puede ser mono-documento o multi-documento (Zajic, Dorr, y Lin, 2008): los mono-documento reducen la información de un único documento, y los multi-documento seleccionan la información de una colección. Un resumen también puede ser abstractivo, donde las palabras y oraciones que lo forman pueden no existir en el texto original, o extractivo, donde sus oraciones sí existen en la fuente original (Wan, 2008).

El problema del resumen extractivo multi-documento puede ser formulado como un pro-

ISSN 1135-5948. DOI 10.26342/2020-65-2

blema de optimización tanto mono-objetivo como multi-objetivo. En la optimización mono-objetivo solo se optimiza una única función objetivo, la cual incluye todos los criterios de forma ponderada (Alguliev, Aliguliyev, y Mehdiyev, 2011). Por otro lado, en la optimización multi-objetivo todas las funciones objetivo se optimizan de manera simultánea. Además, la optimización multi-objetivo ha logrado mejores resultados que la mono-objetivo (Saleh, Kadhim, y Attea, 2015). Las funciones objetivo utilizadas en estos trabajos fueron la cobertura del contenido y la reducción de la redundancia.

En la optimización multi-objetivo la solución generada no es única, sino que es un conjunto de soluciones no dominadas denominado conjunto de Pareto (Sudeng y Wattanapongsakorn, 2015). Los métodos de reducción del frente de Pareto pueden clasificarse en tres grupos. Primero, los métodos basados en preferencias, que necesitan que el usuario las asigne de forma previa (Antipova et al., 2015). Segundo, los métodos de *clustering*, que seleccionan uno o varios subconjuntos de soluciones del frente mediante el uso de técnicas basadas en similitud, como en Taboada y Coit (2007). Y tercero, los métodos basados en distancias seleccionan una única solución del frente basándose en la distancia al punto ideal (Padhye y Deb, 2011).

En este trabajo se han estudiado y comparado varios métodos automáticos de reducción del frente de Pareto a una única solución aplicados al problema de resumen extractivo multi-documento. Estos métodos se han basado en los conceptos del hipervolumen, de la solución consenso y de varios tipos de distancias. La experimentación ha sido realizada con los conjuntos de datos de DUC (*Document Understanding Conferences*), y los resultados se han evaluado con las métricas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). Además, también se ha realizado un análisis estadístico de los resultados obtenidos.

2 Trabajo relacionado

A continuación se analizan los métodos más usados para reducir el frente de Pareto.

En primer lugar, los métodos basados en preferencias de usuario necesitan que los usuarios asignen sus preferencias a priori para obtener un conjunto reducido de soluciones. Estas preferencias suelen estar relacionadas con los pesos de las funciones objetivo. El método de función de estrés ponderado propuesto por Ferreira, Fonseca, y Gaspar-Cunha (2007) integra las preferencias del usuario para encontrar la mejor región del frente de Pareto de acuerdo con estas preferencias. De la misma manera, el método de utilidades aditivas discriminatorias (Soylu y Ulusoy, 2011) requiere que el usuario asigne algunas referencias que reflejen sus preferencias. Otro método es el basado en el filtro de Pareto, desarrollado por Antipova et al. (2015), usado para reducir y facilitar el análisis post-óptimo con la idea de clasificar las soluciones de acuerdo a una eficiencia global y seleccionar aquellas que muestren un mejor equilibrio entre los objetivos.

En segundo lugar, los métodos de *clustering* dividen el frente de Pareto en varios grupos. Estos métodos comienzan con un número concreto de grupos, calculando el centroide para cada uno de ellos de forma iterativa y agrupando los puntos del frente de Pareto con el centroide más similar. El algoritmo de partición de grupos k-means (Taboada y Coit, 2007) proporciona conjuntos más pequeños de soluciones intermedias óptimas, calculando el centroide de cada grupo y asignando cada solución al grupo con el centroide más cercano. Aguirre y Taboada (2011) usaron el enfoque de árbol de crecimiento dinámico auto-organizado para reducir de forma inteligente el tamaño del conjunto, obteniendo así soluciones representativas. Además, la técnica del agrupamiento jerárquico consiste en la formación de subconjuntos de acuerdo a las decisiones de diseño (Veerappa y Letier, 2011).

Por último, los métodos basados en distancias seleccionan el punto del frente de Pareto que tiene la distancia más corta al punto ideal, que representa una solución en la que los valores de las funciones objetivo son óptimos. Padhye y Deb (2011) utilizaron el método de la métrica L_2 para seleccionar la solución con la distancia Euclídea más corta al punto ideal. Del mismo modo, Siwale (2013) presentó la solución de compromiso Chebyshev, que utiliza el criterio de la cercanía a un punto ideal basándose en esta distancia.

Los métodos basados en preferencias de usuario y los basados en *clustering* no seleccionan una única solución. Además, los métodos basados en distancias que se han revisado solo tienen en cuenta dos tipos de distancias.

En este trabajo se han evaluado otros tipos de distancias, como la Manhattan, la Mahalanobis y la Levenshtein, además de otros métodos, como el del mayor hipervolumen y la solución consenso.

3 Definición del problema

El problema de resumen extractivo multi-documento se formula a continuación como un problema de optimización multi-objetivo. Los métodos más usados en este campo son los basados en vectores de palabras. En ellos, una oración se representa como un vector de palabras, y para medir la similitud entre oraciones se utiliza un criterio particular, siendo la similitud coseno el más utilizado.

3.1 La similitud coseno

Dado el conjunto $T = \{t_1, t_2, ..., t_m\}$ que contiene los m términos distintos existentes en la colección de documentos D. Cada oración $s_i \in D$ se representa como $s_i = (w_{i1}, w_{i2}, ..., w_{im}), i = 1, 2, ..., n$, siendo n el número de oraciones en D. Cada componente w_{ik} representa el peso del término t_k en la oración s_i, que se calcula mediante el esquema tf_isf (*term-frequency inverse-sentence-frequency*) de la siguiente forma:

$$w_{ik} = tf_{ik} \cdot log(n/n_k) \tag{1}$$

donde tf_{ik} cuenta cuántas veces está el término t_k en la oración s_i, y $log(n/n_k)$ es el factor isf, siendo n_k el número de oraciones que contienen el término t_k.

El contenido principal de D puede representarse como la media de los pesos de los m términos en T mediante un vector de medias $O = (o_1, o_2, ..., o_m)$, cuyas componentes se calculan como:

$$o_k = \frac{1}{n} \sum_{i=1}^{n} w_{ik} \tag{2}$$

La similitud coseno se basa en los pesos definidos previamente. Concretamente, mide la semejanza entre el par de oraciones $s_i = (w_{i1}, w_{i2}, ..., w_{im})$ y $s_j = (w_{j1}, w_{j2}, ..., w_{jm})$ de la siguiente forma:

$$sim(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}} \tag{3}$$

3.2 La optimización multi-objetivo

La colección de documentos $D = \{d_1, d_2, ..., d_N\}$, que contiene N documentos, también puede representarse como un conjunto de n oraciones: $D = \{s_1, s_2, ..., s_n\}$. El fin es generar un resumen $R \subset D$ que tenga en cuenta los siguientes tres aspectos:

- Cobertura del contenido. El resumen debe cubrir el contenido principal de la colección de documentos.
- Reducción de la redundancia. El resumen no debe contener oraciones de la colección de documentos similares entre sí.
- Longitud. El resumen debe tener una longitud prefijada L.

El problema de resumen extractivo multi-documento implica la optimización simultánea de la cobertura del contenido y de la reducción de la redundancia. Sin embargo, estos criterios son contradictorios entre sí. Por lo tanto, la mejor forma de resolver este problema es mediante un enfoque de optimización multi-objetivo.

Antes es necesario definir la representación de una solución, $X = (x_1, x_2, ..., x_n)$, donde x_i es una variable binaria que tiene en cuenta la presencia o ausencia ($x_i = 1$ o $x_i = 0$) de la oración s_i en el resumen R.

El primer objetivo a optimizar, $\Phi_{CC}(X)$, es el relativo al criterio de la cobertura del contenido. Dadas las oraciones $s_i \in R$, este objetivo se representa como la similitud entre las oraciones s_i y todas las oraciones en la colección D, representada por el vector medio O. Por lo tanto, la siguiente función debe ser maximizada:

$$\Phi_{CC}(X) = \sum_{i=1}^{n} sim(s_i, O) \cdot x_i \tag{4}$$

El segundo objetivo a optimizar, $\Phi_{RR}(X)$, se refiere a la reducción de la redundancia. En este caso, se necesita otra variable binaria y_{ij} que relacione la presencia o ausencia simultánea ($y_{ij} = 1$ o $y_{ij} = 0$) del par de oraciones s_i y s_j en el resumen R. Para cada par de oraciones $s_i, s_j \in R$, su similitud $sim(s_i, s_j)$ debe ser minimizada, lo que es equivalente a maximizar la siguiente función:

$$\Phi_{RR}(X) = \frac{1}{\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(s_i, s_j) \cdot y_{ij}\right) \cdot \sum_{i=1}^{n} x_i} \tag{5}$$

Tras la definición de los objetivos, se puede formular el problema de optimización multi-objetivo:

$$\text{máx } \Phi(X) = \{\Phi_{CC}(X), \Phi_{RR}(X)\} \tag{6}$$

$$\text{sujeto a } L - \varepsilon \leq \sum_{i=1}^{n} l_i \cdot x_i \leq L + \varepsilon \qquad (7)$$

donde l_i es la longitud de la oración s_i y ε es la tolerancia de la longitud del resumen:

$$\varepsilon = \max_{i=1,2,\ldots,n} l_i - \min_{i=1,2,\ldots,n} l_i \qquad (8)$$

4 Métodos de reducción del frente de Pareto a una única solución

En esta sección se presentan los diferentes métodos estudiados para reducir el frente de Pareto a una única solución.

4.1 Mayor hipervolumen

El hipervolumen mide el espacio cubierto por los valores de las funciones objetivo correspondientes a cada solución del frente de Pareto (Beume et al., 2009). Al tratarse de un problema bidimensional, el hipervolumen es el área cubierta por cada punto (solución). Este método selecciona la solución asociada al punto con mayor hipervolumen (MH) (ver Figura 1).

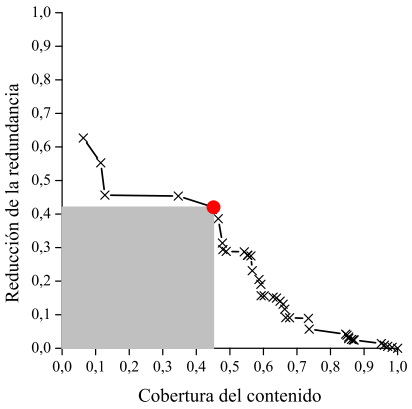

Figura 1: Mayor hipervolumen

4.2 Solución consenso

La solución consenso (SC) es aquella que se genera a partir de todas las soluciones del frente de Pareto (Pérez et al., 2017). En este problema se forma con las oraciones de los resúmenes asociados, seleccionando las oraciones más utilizadas hasta alcanzar la restricción de longitud $L \pm \epsilon$ indicada en la Ecuación (7). Al tratarse de un resumen generado a partir de otros, su solución asociada puede no existir en el frente (ver Figura 2).

4.3 Distancia más corta al punto ideal

El punto ideal es aquel en el que los valores de los objetivos son los mejores posibles.

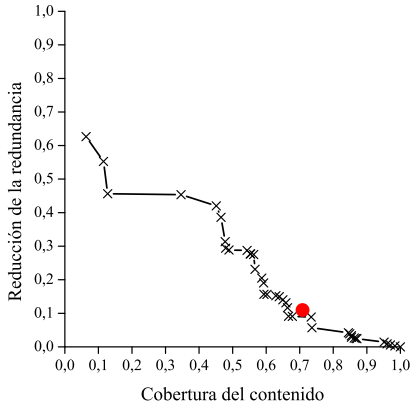

Figura 2: Solución consenso

En un espacio objetivo normalizado, con rango [0,1], donde las funciones objetivo deben ser maximizadas, el punto ideal está situado en (1,1). Por lo tanto, este método (DCPI) mide la distancia entre cada punto del frente y el punto ideal, y selecciona el que tiene la distancia más corta (ver la Figura 3, que se basa en la distancia Euclídea). Dados dos puntos del espacio $P_1 = (p_{11}, p_{12})$ y $P_2 = (p_{21}, p_{22})$, la distancia d entre P_1 y P_2 se define como $d(P_1, P_2)$. Además de las distancias Euclídea (DCPI_E) y Chebyshev (DCPI_C), también se han estudiado las distancias Manhattan (DCPI_M) y Mahalanobis (DCPI_B).

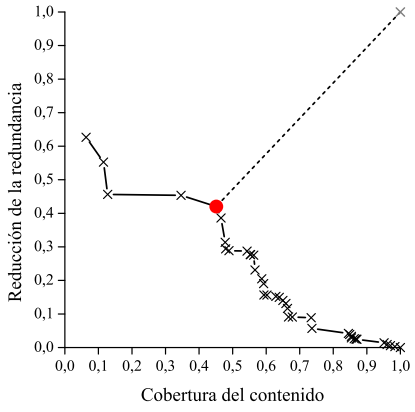

Figura 3: Distancia más corta al punto ideal

En primer lugar, la distancia Euclídea es la distancia ordinaria entre dos puntos en un espacio Euclídeo (Padhye y Deb, 2011):

$$d_E(P_1, P_2) = \sqrt{\sum_{i=1}^{2} (p_{1i} - p_{2i})^2} \qquad (9)$$

En segundo lugar, la distancia Manhattan está basada en la geografía callejera cuadri-

culada del distrito de Manhattan (Wu et al., 2015). Se define como la suma de las longitudes de las proyecciones en los ejes de coordenadas del segmento de línea entre dos puntos:

$$d_M(P_1, P_2) = \|P_1, P_2\| = \sum_{i=1}^{2} |p_{1i} - p_{2i}| \quad (10)$$

En tercer lugar, la distancia Chebyshev es la mayor de las longitudes de las proyecciones en los ejes de coordenadas del segmento de línea entre dos puntos (Siwale, 2013):

$$d_C(P_1, P_2) = \max_{i=1,2} |p_{1i} - p_{2i}| \quad (11)$$

Y en cuarto lugar, la distancia Mahalanobis determina la similitud entre dos variables aleatorias multi-dimensionales, teniendo en cuenta sus correlaciones (Zhao et al., 2017). En este contexto se define como:

$$d_B(P_1, P_2) = \sqrt{\sum_{i=1}^{2} \left(\frac{p_{1i} - p_{2i}}{\sigma_i} \right)^2} \quad (12)$$

donde σ_i es la desviación estándar no corregida, que se calcula midiendo la dispersión de los valores de la función objetivo i a través de los G puntos del frente de Pareto:

$$\sigma_i = \frac{1}{G} \sum_{g=1}^{G} (p_{gi} - \bar{p_i})^2 \quad (13)$$

siendo $\bar{p_i} = \frac{1}{G} \sum_{g=1}^{G} p_{gi}$.

4.4 Distancia más corta a todos los puntos

Este método (DCTP) evalúa la suma de las distancias entre un punto y todos los restantes, y selecciona el punto del frente con la menor distancia (ver Figura 4). Para este método, además de las cuatro distancias definidas anteriormente (Euclídea DCTP_E, Chebyshev DCTP_C, Manhattan DCTP_M y Mahalanobis DCTP_B), también se ha considerado la distancia Levenshtein (DCTP_L) (Ristad y Yianilos, 1998). En este contexto se define como el número de inserciones y eliminaciones de oraciones necesario para transformar un resumen en otro. Dados dos resúmenes (o conjuntos de oraciones) $\mathcal{R}$ y $\mathcal{S}$, la distancia de Levenshtein entre ellos se calcula como:

$$d_L(\mathcal{R}, \mathcal{S}) = |\mathcal{R}| + |\mathcal{S}| - 2 \cdot |\mathcal{R} \cap \mathcal{S}| \quad (14)$$

donde $|\cdot|$ es la cardinalidad o número de elementos en un conjunto.

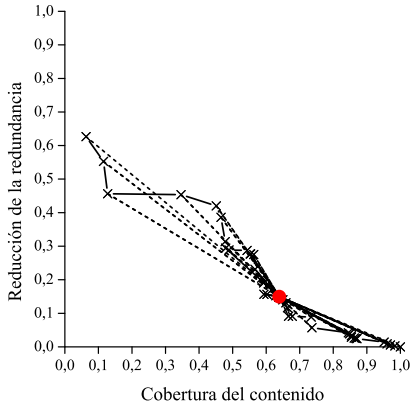

Figura 4: Distancia más corta a todos los puntos

5 Resultados

5.1 Conjuntos de datos

Para la ejecución de los experimentos se han utilizado los conjuntos de datos DUC, que es un banco de pruebas para la evaluación de resúmenes automáticos. En este estudio se han utilizado 10 temas (del d061j al d070f) del conjunto de datos DUC2002 (NIST, 2014). La longitud de resumen establecida en estos conjuntos es de 200 palabras.

5.2 Métricas de evaluación

El rendimiento de los métodos ha sido evaluado con las métricas ROUGE. Estas métricas son las que se utilizan en DUC (Lin, 2004). ROUGE mide la similitud entre un resumen generado automáticamente y un resumen generado por un experto. Las puntuaciones ROUGE empleadas han sido el ROUGE-2 (R-2) y el ROUGE-L (R-L), y se ha usado la media aritmética para medir la tendencia central en cada tema.

Además, se ha realizado un análisis estadístico de las puntuaciones ROUGE mediante pruebas de hipótesis adecuadas. Los p-valores menores que 0,05 han sido considerados como estadísticamente significativos.

5.3 Configuración

Sanchez-Gomez, Vega-Rodríguez, y Pérez (2018) propusieron y aplicaron un enfoque de optimización multi-objetivo basado en el algoritmo de colonia de abejas artificiales aplicado al problema de resumen extractivo multi-documento. Los siguientes parámetros produjeron buenos resultados: tamaño de población = 50; número de ciclos = 1000; y probabilidad de mutación = 0,1.

Método	d061j	d062j	d063j	d064j	d065j	d066j	d067f	d068f	d069f	d070f	Media
MH	**0,297**	0,186	0,212	0,201	0,112	0,189	0,273	0,242	0,131	0,099	0,194
SC	0,268	**0,400**	0,199	**0,203**	0,167	0,195	0,273	0,277	**0,254**	0,172	**0,241**
DCPI$_{\mathbf{E}}$	**0,297**	0,191	0,205	0,168	0,115	0,059	0,231	0,242	0,233	0,099	0,184
DCPI$_{\mathbf{M}}$	0,248	0,295	**0,256**	0,187	0,168	0,063	0,172	0,208	0,195	**0,258**	0,205
DCPI$_{\mathbf{C}}$	0,222	0,140	0,183	0,197	0,124	0,189	**0,353**	0,260	0,145	0,101	0,191
DCPI$_{\mathbf{B}}$	**0,297**	0,190	0,153	0,170	0,143	0,059	0,231	**0,318**	0,111	0,098	0,177
DCTP$_{\mathbf{E}}$	0,219	0,245	0,192	0,175	0,163	0,203	0,272	0,291	0,119	0,111	0,199
DCTP$_{\mathbf{M}}$	0,232	0,270	0,185	0,169	0,164	**0,221**	0,268	0,286	0,122	0,112	0,203
DCTP$_{\mathbf{C}}$	0,222	0,242	0,196	0,169	0,164	0,202	0,284	0,282	0,118	0,117	0,200
DCTP$_{\mathbf{B}}$	0,219	0,245	0,191	0,175	0,166	0,203	0,274	0,292	0,117	0,111	0,199
DCTP$_{\mathbf{L}}$	0,272	0,232	0,175	0,167	**0,177**	0,196	0,268	0,261	0,148	0,130	0,203

Tabla 1: Resultados para ROUGE-2 (en negrita los mejores valores)

Método	d061j	d062j	d063j	d064j	d065j	d066j	d067f	d068f	d069f	d070f	Media
MH	**0,541**	0,342	0,469	**0,404**	0,366	0,428	0,508	0,450	0,321	0,402	0,423
SC	0,509	**0,571**	0,458	0,390	**0,407**	0,461	0,517	0,506	0,517	0,480	**0,482**
DCPI$_{\mathbf{E}}$	0,540	0,347	0,459	0,347	0,375	0,278	0,417	0,450	**0,536**	0,402	0,414
DCPI$_{\mathbf{M}}$	0,508	0,483	**0,489**	0,373	0,393	0,281	0,413	0,450	0,461	**0,481**	0,433
DCPI$_{\mathbf{C}}$	0,494	0,291	0,414	0,400	0,375	0,425	**0,558**	0,465	0,380	0,406	0,421
DCPI$_{\mathbf{B}}$	0,540	0,345	0,357	0,312	0,341	0,279	0,417	**0,531**	0,256	0,402	0,378
DCTP$_{\mathbf{E}}$	0,481	0,463	0,440	0,368	0,390	0,458	0,525	0,518	0,441	0,435	0,452
DCTP$_{\mathbf{M}}$	0,492	0,477	0,430	0,365	0,390	**0,468**	0,518	0,512	0,449	0,440	0,454
DCTP$_{\mathbf{C}}$	0,438	0,462	0,445	0,364	0,391	0,460	0,532	0,507	0,443	0,440	0,453
DCTP$_{\mathbf{B}}$	0,481	0,463	0,438	0,368	0,392	0,458	0,525	0,518	0,443	0,435	0,452
DCTP$_{\mathbf{L}}$	0,518	0,406	0,384	0,349	0,395	0,456	0,488	0,485	0,322	0,431	0,423

Tabla 2: Resultados para ROUGE-L (en negrita los mejores valores)

5.4 Resultados

Los métodos han sido evaluados con los frentes de Pareto resultantes de las 20 repeticiones ejecutadas para cada tema. Las Tablas 1 y 2 contienen los resultados para R-2 y R-L respectivamente, mostrando que la solución consenso (SC) ha conseguido el mejor valor medio en ambos ROUGE. También es el método que ha obtenido el mejor valor en mayor número de temas (3 en R-2 y 2 en R-L). Los métodos de la distancia más corta al punto ideal y a todos los puntos con la distancia Manhattan (DCPI$_{\mathrm{M}}$ y DCTP$_{\mathrm{M}}$) han logrado el segundo mejor valor medio (DCPI$_{\mathrm{M}}$ en R-2 y DCTP$_{\mathrm{M}}$ en R-L). Además, DCPI$_{\mathrm{M}}$ y DCTP$_{\mathrm{M}}$ han obtenido el mejor valor en 2 temas y en 1 tema en ambos ROUGE, respectivamente. El método del mayor hipervolumen (MH) ha obtenido el mejor resultado en 1 tema en R-2 y en 2 temas en R-L. En cuanto a los métodos basados en distancias, y agrupando todas ellas, DCPI ha obtenido el mejor resultado en 5 temas en ambos ROUGE, mientras que DCTP los ha obtenido en 2 temas para R-2 y en 1 tema para R-L.

Una vez analizados los resultados de las puntuaciones ROUGE, se muestra el análisis estadístico. Se han aplicado pruebas de hipótesis estadísticas para analizar si existen diferencias estadísticamente significativas entre los diferentes métodos. Las condiciones de normalidad y homokedasticidad para el ANOVA uni-factorial no se pueden asumir para ningún tema, por lo que se ha utilizado el test de Kruskal-Wallis. Los p-valores son inferiores a 0,001 para todos los temas, por lo que al menos un par de métodos son significativamente diferentes. Las comparaciones por pares se han realizado entre el método de la solución consenso y el resto de los métodos mediante el test de Conover con la corrección de Bonferroni. Las Tablas 3 y 4 muestran las diferencias estadísticamente significativas, y de ellas se concluye que el método de la solución consenso (SC) aporta diferencias estadísticamente significativas con respecto al resto. Concretamente, SC ha obtenido diferencias estadísticamente significativas con el 80 % de los métodos en R-2 en al menos la mitad de los temas. En cuanto a R-L, SC ha obtenido diferencias estadísticamente significativas con todos los métodos en al menos la mitad de los temas.

Para terminar, la Tabla 5 muestra los por-

Método	d061j	d062j	d063j	d064j	d065j	d066j	d067f	d068f	d069f	d070f	Total
MH	**<0,001**	**<0,001**	0,056	0,077	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	**<0,001**	7
DCPI$_E$	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	1,000	**<0,001**	8
DCPI$_M$	1,000	**<0,001**	**0,002**	1,000	1,000	**<0,001**	**<0,001**	**<0,001**	**0,002**	1,000	6
DCPI$_C$	0,057	**<0,001**	**0,005**	0,429	**<0,001**	**0,006**	**<0,001**	**0,008**	**0,025**	**<0,001**	8
DCPI$_B$	**<0,001**	**<0,001**	**<0,001**	**<0,001**	0,683	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	8
DCTP$_E$	**<0,001**	**<0,001**	1,000	0,202	1,000	0,415	1,000	1,000	**<0,001**	**<0,001**	4
DCTP$_M$	**0,007**	**<0,001**	**0,017**	**<0,001**	1,000	**<0,001**	1,000	1,000	**<0,001**	**<0,001**	7
DCTP$_C$	**0,002**	**<0,001**	1,000	**<0,001**	1,000	0,197	1,000	1,000	**<0,001**	**<0,001**	5
DCTP$_B$	**<0,001**	**<0,001**	1,000	0,202	1,000	0,619	1,000	1,000	**<0,001**	**<0,001**	4
DCTP$_L$	1,000	**<0,001**	**<0,001**	**<0,001**	1,000	1,000	1,000	0,442	**<0,001**	**0,003**	5

Tabla 3: p-valores obtenidos de la comparación por pares entre el método de la solución consenso y el resto de métodos para ROUGE-2 (en negrita las diferencias estadísticamente significativas)

Método	d061j	d062j	d063j	d064j	d065j	d066j	d067f	d068f	d069f	d070f	Total
MH	**<0,001**	**<0,001**	1,000	0,112	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	**<0,001**	7
DCPI$_E$	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	1,000	**<0,001**	8
DCPI$_M$	0,457	**<0,001**	**0,024**	1,000	**0,027**	**<0,001**	**<0,001**	**<0,001**	1,000	1,000	6
DCPI$_C$	1,000	**<0,001**	**<0,001**	0,900	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	8
DCPI$_B$	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	**<0,001**	1,000	**<0,001**	**<0,001**	9
DCTP$_E$	**<0,001**	**<0,001**	0,419	**0,011**	0,275	1,000	1,000	1,000	**<0,001**	**<0,001**	5
DCTP$_M$	0,134	**<0,001**	**0,003**	**0,004**	0,205	1,000	1,000	1,000	**<0,001**	**<0,001**	5
DCTP$_C$	**0,005**	**<0,001**	1,000	**<0,001**	0,303	1,000	0,635	1,000	**<0,001**	**<0,001**	5
DCTP$_B$	**<0,001**	**<0,001**	0,206	**0,027**	0,599	1,000	1,000	1,000	**<0,001**	**<0,001**	5
DCTP$_L$	1,000	**<0,001**	**<0,001**	**<0,001**	1,000	1,000	1,000	**0,048**	**<0,001**	**<0,001**	6

Tabla 4: p-valores obtenidos de la comparación por pares entre el método de la solución consenso y el resto de métodos para ROUGE-L (en negrita las diferencias estadísticamente significativas)

Método	MH	DCPI$_E$	DCPI$_M$	DCPI$_C$	DCPI$_B$	DCTP$_E$	DCTP$_M$	DCTP$_C$	DCTP$_B$	DCTP$_L$	Media
ROUGE-2	24,23	30,98	17,56	26,18	36,16	21,11	18,72	20,50	21,11	18,72	23,53
ROUGE-L	13,95	16,43	11,32	14,49	27,51	6,64	6,17	6,40	6,64	13,95	12,35

Tabla 5: Porcentajes de mejora obtenidos por el método de la solución consenso

centajes de mejora obtenidos por la solución consenso con respecto al resto de métodos. Los porcentajes de mejora obtenidos por la solución consenso varían del 17,56 % al 36,16 % para R-2 y del 6,17 % al 27,51 % para R-L. Además, los porcentajes de mejora global obtenidos son del 23,53 % y del 12,35 % en R-2 y R-L, respectivamente.

6 Conclusiones

El problema del resumen extractivo multi-documento se ha resuelto aplicando optimización multi-objetivo. Como el resultado es un conjunto de soluciones no dominadas o frente de Pareto, el objetivo de este trabajo ha sido llevar a cabo un análisis para seleccionar un único resumen del frente. Para ello, se han implementado, evaluado y comparado diferentes métodos basados en el mayor hipervolumen, en la solución consenso, en la distancia más corta al punto ideal y en la distancia más corta a todos los puntos, evaluando para estos dos últimos varios tipos de distancias. Los resultados indican que la solución consenso es el método que mejores resultados ha obtenido. Los porcentajes de mejora global alcanzados son del 23,53 % para ROUGE-2 y del 12,35 % para ROUGE-L.

Agradecimientos

Esta investigación ha sido apoyada por la Agencia Estatal de Investigación (proyectos PID2019-107299GB-I00 y MTM2017-86875-C3-2-R), por la Junta de Extremadura (proyectos GR18090 y GR18108) y por la Unión Europea (Fondo Europeo de Desarrollo Regional). Jesus M. Sanchez-Gomez está apoyado por un Contrato Predoctoral financiado por la Junta de Extremadura (contrato PD18057) y la Unión Europea (Fondo Social Europeo).

Bibliografía

Aguirre, O. y H. Taboada. 2011. A Clustering Method Based on Dynamic Self Organizing Trees for Post-Pareto Optima-

lity Analysis. *Procedia Computer Science*, 6:195–200.

Alguliev, R. M., R. M. Aliguliyev, y C. A. Mehdiyev. 2011. Sentence selection for generic document summarization using an adaptive differential evolution algorithm. *Swarm Evol. Comput.*, 1(4):213–222.

Antipova, E., C. Pozo, G. Guillén-Gosálbez, D. Boer, L. F. Cabeza, y L. Jiménez. 2015. On the use of filters to facilitate the post-optimal analysis of the Pareto solutions in multi-objective optimization. *Comput. Chem. Eng.*, 74:48–58.

Beume, N., C. M. Fonseca, M. López-Ibáñez, L. Paquete, y J. Vahrenhold. 2009. On the Complexity of Computing the Hypervolume Indicator. *IEEE Trans. Evol. Comput.*, 13(5):1075–1082.

Ferreira, J. C., C. M. Fonseca, y A. Gaspar-Cunha. 2007. Methodology to Select Solutions from the Pareto-Optimal Set: A Comparative Study. En *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, páginas 789–796. ACM.

Hashimi, H., A. Hafez, y H. Mathkour. 2015. Selection criteria for text mining approaches. *Comput. Hum. Behav.*, 51:729–733.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. En *Proceedings of the ACL-04 Workshop*, volumen 8, páginas 74–81. ACL.

NIST. 2014. Document Understanding Conferences. http://duc.nist.gov. Último acceso: 11 de agosto de 2020.

Padhye, N. y K. Deb. 2011. Multi-objective optimisation and multi-criteria decision making in SLS using evolutionary approaches. *Rapid Prototyping Journal*, 17(6):458–478.

Pérez, C. J., M. A. Vega-Rodríguez, K. Reder, y M. Flörke. 2017. A Multi-Objective Artificial Bee Colony-based optimization approach to design water quality monitoring networks in river basins. *Journal of Cleaner Production*, 166:579–589.

Ristad, E. S. y P. N. Yianilos. 1998. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532.

Saleh, H. H., N. J. Kadhim, y B. A. Attea. 2015. A genetic based optimization model for extractive multi-document text summarization. *Iraqi Journal of Science*, 56(2):1489–1498.

Sanchez-Gomez, J. M., M. A. Vega-Rodríguez, y C. J. Pérez. 2018. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Syst.*, 159:1–8.

Siwale, I. 2013. Practical Multi-Objective Programming. Informe técnico, Technical Report RD-14-2013. APEX Research.

Soylu, B. y S. K. Ulusoy. 2011. A preference ordered classification for a multi-objective max–min redundancy allocation problem. *Comput. Oper. Res.*, 38(12):1855–1866.

Sudeng, S. y N. Wattanapongsakorn. 2015. Post Pareto-optimal pruning algorithm for multiple objective optimization using specific extended angle dominance. *Eng. Appl. Artif. Intell.*, 38:221–236.

Taboada, H. A. y D. W. Coit. 2007. Data Clustering of Solutions for Multiple Objective System Reliability Optimization Problems. *Qual. Technol. Quant. Manag.*, 4(2):191–210.

Veerappa, V. y E. Letier. 2011. Understanding Clusters of Optimal Solutions in Multi-Objective Decision Problems. En *19th Requirements Engineering Conference*, páginas 89–98. IEEE.

Wan, X. 2008. An exploration of document impact on graph-based multi-document summarization. En *Proceedings of the Conference on Empirical Methods in NLP*, páginas 755–762. ACL.

Wu, L., X. Xu, X. Ye, y X. Zhu. 2015. Repeat and near-repeat burglaries and offender involvement in a large Chinese city. *Cartogr. Geogr. Inf. Sci.*, 42(2):178–189.

Zajic, D. M., B. J. Dorr, y J. Lin. 2008. Single-document and multi-document summarization techniques for email threads using sentence compression. *Inf. Process. Manage.*, 44(4):1600–1610.

Zhao, L., Z. Lu, W. Yun, y W. Wang. 2017. Validation metric based on Mahalanobis distance for models with multiple correlated responses. *Reliab. Eng. Syst. Saf.*, 159:80–89.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 29-36 recibido 30-04-2020 revisado 26-05-2020 aceptado 26-05-2020

Generación de frases literarias: un experimento preliminar

Generation of Literary Sentences: a Preliminary Approach

Luis-Gil Moreno-Jiménez[1,4], **Juan-Manuel Torres-Moreno**[1,3], **Roseli S. Wedemann**[2]
[1]Université d'Avignon/LIA
[2]Universidade do Estado do Rio de Janeiro
[3]Polytechnique Montréal
[4]Universidad Tecnológica de la Selva
luis-gil.moreno-jimenez@alumni.univ-avignon.fr,
juan-manuel.torres@univ-avignon.fr, roseli@ime.uerj.br

Resumen: En este trabajo abordamos la generación automática de frases literarias en español. Proponemos un modelo de generación textual basado en algoritmos estadísticos y análisis sintáctico superficial. Presentamos resultados preliminares que son bastante alentadores.
Palabras clave: Generación de texto, Modelos de lenguaje, Word embedding

Abstract: In this paper we address the automatic generation of literary phrases in Spanish. We propose a model for text generation based on statistical algorithms and Shallow Parsing. We present preliminary results that are quite encouraging.
Keywords: Natural Language Generation, Language Models, Word Embedding

1 Introducción

La Generación Automática de Texto (GAT), es un área del Procesamiento de Lenguaje Natural (PLN), que en los últimos años ha logrado avances importantes, bajo la convicción de desarrollar modelos computacionales capaces de simular cómo las personas manipulan el lenguaje. La mayoría de estos trabajos persiguen la automatización de ciertos procesos para aumentar la productividad en el ámbito industrial, académico, tecnológico, etc. (Szymanski y Ciota, 2002; Sridhara et al., 2010; Fu et al., 2014).

Sin embargo, existe un enfoque dentro de GAT que ha sido poco abordado: la generación automática de texto literario. Desde hace algún tiempo, diversos investigadores han trabajado en modelos generativos de este tipo, la mayor parte se avocan directamente a poemas, poesías o narrativas (Lebret, Grangier, y Auli, 2016; Brantley et al., 2019). Consideramos que la literatura, como "Proceso Creativo" Boden (2004), es el resultado que las personas desarrollan como parte de un proceso cognitivo complejo, y que la mejor manera de abordarla, es partir de conceptos más amplios como: emociones, asociación de conceptos (semántica) y estilos de redacción. Pensamos que, combinados adecuadamente, estos conceptos podrían usarse para modelar computacionalmente una parte del proceso creativo que una persona sigue para la generación de literatura.

La complejidad para automatizar la generación literaria reside en el análisis de los textos literarios, que sistemáticamente han sido dejados a un lado por varias razones. En primer lugar, el nivel de discurso literario es más complejo que el de los otros géneros. En segundo lugar, a menudo, los documentos literarios hacen referencia a mundos o situaciones imaginarias o alegóricas, a diferencia de géneros como el peridístico o enciclopédico que describen mayoritariamente situaciones o hechos factuales. Estas y otras características presentes en los textos literarios, vuelven sumamente compleja la tarea de análisis automático de este tipo de textos. En este trabajo nos proponemos utilizar corpora literarios, a fin de generar realizaciones literarias (frases nuevas) no presentes en dichos corpora.

Este proceso automatizado da lugar a un campo de investigación denominado Creatividad Computacional (CC) (Pérez y Pérez, 2015), donde el objetivo es modelar el "proce-

ISSN 1135-5948. DOI 10.26342/2020-65-3

so creativo" en un lenguaje que sea interpretable y reproducible en el dominio de lo calculable. Una gran variedad de modelos de IA han sido adaptados e incluso mejorados para lograr simular el proceso creativo a través de modelos computacionales (Colton, 2012). Debe considerarse que el objetivo principal de la CC no es solucionar problemas específicos en el ámbito industrial o académico, sino proponer nuevos paradigmas para la creación de obras artísticas (Colton, Wiggins, y others, 2012).

Otro problema es la falta de una definición universal de literatura; por lo que diversas definiciones son encontradas. Esto complica la tarea de evaluación, ya que, para lograr una percepción literaria homogénea, se debería partir de la misma definición de literatura. En este trabajo optaremos por introducir una definición pragmática de frase literaria, que servirá para nuestros modelos y experimentos.

Definición. *Una frase literaria se caracteriza por poseer elementos (sustantivos, verbos, adjetivos y adverbios) que son percibidos como elegantes y menos coloquiales que sus equivalentes en lengua general.*

Por ejemplo, la frase en lengua general:

- *"Me detuve a descansar luego de haber caminado mucho."*

puede ser ligeramente re-escrita para generar una frase literaria según nuestra definición:

- *"Tomé unos instantes para respirar y oxigenar mis pulmones... pues mi andar se había prolongado largo tiempo."*

En particular, proponemos crear artificialmente frases literarias utilizando modelos generativos y aproximaciones semánticas basados en corpora de lengua literaria. Buscamos, através de la combinación de estos elementos, una homosintaxis, es decir, la producción de texto nuevo a partir de formas de discurso de diversos autores. La homosintaxis no tiene el mismo contenido semántico, tampoco las mismas palabras, pero guarda la misma estructura sintáctica. En este artículo estudiamos el problema de la generación de texto literario en forma de frases aisladas, no a nivel de párrafos. La generación de párrafos puede ser objeto de trabajos futuros. Un protocolo de evaluación de la calidad de las frases generadas será presentado.

Este artículo está estructurado como sigue. La Sección 2 presenta un estado del arte de la creatividad computacional. La Sección 3 describe los corpora utilizados en nuestros experimentos. Los modelos usados son descritos en la Sección 4. Los resultados se encuentran en la Sección 5, antes de concluir en Sección 6 con algunas ideas de trabajos futuros.

2 Estado del arte

En la GAT se encuentran algunos trabajos con diversos objetivos. Szymanski y Ciota (2002) han desarrollado un modelo basado en cadenas de Markov para la generación de texto en polaco. El proceso inicia con un término dado por el usuario (estado inicial). Un proceso probabilístico calcula los estados siguientes. Cada estado es representado por n-gramas de letras o de palabras. El método demostró un mejor comportamiento, generando palabras de hasta 4 o 5 letras, considerando que en polaco esta es la longitud media de palabras.

Sridhara et al. (2010) utilizan un enfoque distinto. Presentan un algoritmo generativo de comentarios descriptivos aplicados a bloques de código en lenguaje Java. Se consideran algunas variables lingüísticas como los nombres de métodos, funciones e instancias. Estos elementos son procesados heurísticamente para generar texto descriptivo.

También existen trabajos menos extensos pero más precisos. Huang et al. (2012) proponen un modelo basado en redes neuronales para la generación de subconjuntos multipalabras. Este mismo objetivo se considera en (Fu et al., 2014), en donde se busca establecer y detectar la relación hiperónimo-hipónimo usando un modelo Word2vec (también basado en redes neuronales (Mikolov, Yih, y Zweig, 2013)). Los autores reportan una precisión de 0.70, al ser evaluado sobre un corpus manualmente etiquetado.

En cuanto a los trabajos relacionados a la GAT, se percibe una diferencia entre aquellos orientados a la generación literaria y aquellos que buscan la generación de texto en lengua general. Por ejemplo, en (Zhang y Lapata, 2014) se propone un modelo para la generación de poemas basado en dos premisas básicas: *¿qué decir?* y *¿cómo decirlo?* El modelo considera una lista de palabras clave para seleccionar un conjunto de frases. Estas frases son procesadas con una red neuronal (Mikolov y Zweig, 2012) para construir nuevas

combinaciones y formular nuevos contextos. El modelo fue evaluado manualmente por 30 expertos en una escala de 1 a 5, analizando legibilidad, coherencia y significatividad en frases de 5 palabras, obteniendo una precisión de 0,75. Sin embargo, la coherencia entre frases resultó ser muy pobre.

Otros trabajos, como los de Oliveira (2012; Oliveira y Cardoso (2015) proponen modelos para la generación de poemas, basados en plantillas lingüísticas. Se utilizan listas de palabras clave para controlar el contexto bajo el cual los poemas son generados. Estos trabajos utilizan la herramienta PEN[1] para obtener la información gramatical de las palabras y crear nuevas plantillas. La ventaja de utilizar métodos basados en plantillas es que ayudan a mantener la coherencia y la gramaticalidad de los textos generados.

Modelos enfocados a la generación de poesía pueden ser analizados en los trabajos de Oliveira (2017) y Agirrezabal et al. (2013). Este último presenta un modelo estocástico, donde se calcula la probabilidad de aparición de etiquetas POS (*Part-of-Speech*), considerando las ocurrencias de estas etiquetas extraídas de diversos corpora. Después se generan nuevas secuencias y posteriormente se procede a la sustitución de las etiquetas POS de sustantivos y adjetivos.

3 Corpora utilizados

En esta sección, describimos los corpora utilizados en nuestros modelos para los experimentos. Se trata del corpus 5KL y del corpus 8KF, ambos creados en idioma español.

3.1 Corpus 5KL

Este corpus fue constituido con aproximadamente 5 000 documentos (en su mayor parte libros) en español. Los textos, en su mayoría, corresponden a géneros literarios: narrativa, poesía, teatro, ensayos, etc[2]. Los documentos originales, en formatos muy heterogéneos[3], fueron procesados para crear un único documento codificado en *utf8*. Dada su heterogeneidad, este corpus presenta una gran cantidad de errores: palabras cortadas o pegadas, símbolos, números, disposición no convencional de párrafos, etc. Lo que complica la tarea de análisis. Estas son, sin embargo, las condiciones que presenta un corpus literario real.

Las herramientas clásicas como FreeLing tienen mucha dificultad en tratar este tipo de documentos. Por ello, decidimos construir un segmentador de frases ad hoc para este tipo de corpus ruidoso. Las frases fueron segmentadas automáticamente, usando un programa en PERL 5.0 y expresiones regulares, para obtener una frase por línea.

Las características del corpus 5KL se encuentran en la Tabla 1[4]. Este corpus es empleado para entrenar el modelo Word2vec (Sección 4).

Corpus	**Frases**	**Tokens**
5KL	9 M	149 M
Media por doc.	2.4 K	37.3 K

Tabla 1: Corpus 5KL: obras literarias

El corpus literario 5KL posee la ventaja de ser muy extenso y adecuado para el aprendizaje automático. Tiene sin embargo, la desventaja de que no todas las frases son *necesariamente* frases literarias. Muchas de ellas son frases de lengua general, que a menudo otorgan una fluidez a la lectura y proporcionan los enlaces necesarios a las ideas.

3.2 Corpus 8KF

Decidimos crear un pequeño corpus controlado, exclusivamente compuesto de "frases literarias", que será utilizado para generar las plantillas o estructuras gramaticales (Sección 4.1). Este corpus de casi 8 000 frases literarias fue constituido manualmente, a partir de poemas, discursos, citas, cuentos y otras obras.

Se evitaron cuidadosamente las frases de lengua general, y también aquellas demasiado cortas ($N \leq 3$ palabras) y demasiado largas ($N \geq 30$ palabras). Algunos elementos que sirvieron para seleccionar manualmente las frases "literarias" fueron: un vocabulario complejo y estético, el cual rara vez es empleado en el lenguaje común, además de la identificación de ciertas figuras literarias como la rima, la anáfora, la metáfora y otras. Las características del corpus 8KF se muestran en la Tabla 2.

4 Modelo de Generación Textual

A continuación, se describen las dos fases que conforman nuestro modelo. La primera

[1] http://code.google.com/p/pen

[2] Dada la dimensión de este corpus, no nos fue posible cuantificar los géneros manualmente. Una aproximación automática podrá realizarse a futuro.

[3] pdf, txt, html, doc, docx, odt, etc.

[4] M = 10^6 y K = 10^3.

Corpus	Frases	Tokens
8KF	7 679	114 K
Media por frase	–	15

Tabla 2: Corpus 8KF: frases literarias

consiste en la generación de una estructura gramatical que hemos denominado *estructura gramatical parcialmente vacía* (EGP), la cual se compone de palabras funcionales (conectores, verbos auxiliares, artículos, etc.) y etiquetas POS (*Part-of-Speech*)[5]. Estas últimas son obtenidas a través de un análisis morfosintáctico realizado con FreeLing [6] (Padró y Stanilovsky, 2012).

La segunda, consiste en la sustitución de las etiquetas POS con palabras semánticamente relacionadas a un contexto (*query*, Q), que es solicitado al usuario. Para ello, se emplea un modelo Word2vec (Mikolov, Yih, y Zweig, 2013) con interpretaciones geométricas.

4.1 Generación de EGP basado en texto enlatado

Esta técnica conocida como *texto enlatado* (*Canned Text*) (Molins y Lapalme, 2015) cuenta con la ventaja de agilizar el análisis sintáctico y permitir centrarnos directamente en el vocabulario a emplear (McRoy, Channarukul, y Ali, 2003; van Deemter, Theune, y Krahmer, 2005). La EGP es generada usando una frase del corpus 8KF (Sección 3), donde se reemplazan únicamente verbos, sustantivos o adjetivos $\{V, S, A\}$, por sus respectivas etiquetas POS. Las otras entidades lingüísticas, en particular las palabras funcionales, son conservadas.

El proceso inicia con la selección aleatoria de una frase original $f_o \in$ corpus 8KF de longitud $|f_o| = N$. Se selecciona una f_o para cada frase que se desee generar. f_o será analizada con FreeLing para identificar los sintagmas. Los elementos $\{V, S, A\}$ de los sintagmas de f_o serán reemplazados por sus respectivas etiquetas POS. Estos elementos lingüísticos son los que mayor información aportan en un texto, independientemente de su género (Bracewell, Ren, y Kuriowa, 2005).

[5]Etiquetas que aportan información gramatical de cada palabra http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html

[6]Desarrollado en el centro TALP (Universidad Politécnica de Cataluña) y puede ser obtenido en la dirección: http://nlp.lsi.upc.edu/freeling

Según nuestra hipótesis, al sustituirlos, lograremos la generación de frases nuevas por homosintaxis: semántica diferente, misma estructura[7].

La arquitectura del modelo se ilustra en la Figura 1. Los cuadros llenos representan palabras funcionales y los cuadros vacíos etiquetas POS a ser reemplazadas.

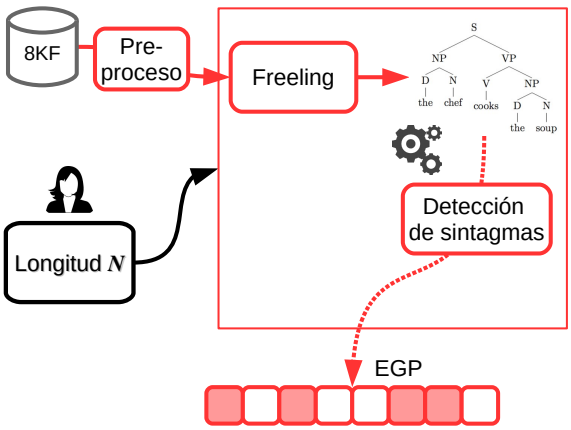

Figura 1: Modelo generativo *canned text*

4.2 Reemplazo de etiquetas POS

En esta fase, empleamos el modelo de aproximación semántica *Word2vec*, que utiliza un algoritmo basado en redes neuronales. El objetivo es obtener la representación vectorial de las palabras (*embeddings*) del corpus 5KL en un espacio n-dimensional[8], y poder calcular las distancias entre estas. El entrenamiento del modelo Word2vec se describe a continuación.

4.2.1 Modelo Word2vec

Para el entrenamiento y la implementación de Word2vec se utiliza la biblioteca Gensim[9], una implementación en Python de Word2vec[10]. El corpus 5KL es pre-procesado para uniformizar el formato del texto, eliminando caracteres que no son importantes en los análisis de PLN (como puntuación, números, etc.) (Torres-Moreno, 2014).

Se consideraron palabras con más de 5 ocurrencias en el corpus. Se definió una longitud de 10 palabras para la ventana contextual. Para las representaciones vectoriales se

[7]Al contrario de la paráfrasis que busca conservar la semántica, alterando la estructura sintáctica.

[8]Word2vec pertenece a un amplio campo de investigación dentro de PLN conocido como *Representation Learning* (Bengio, Courville, y Vincent, 2013).

[9]Disponible en: https://pypi.org/project/gensim/

[10]https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

consideraron vectores de 60 dimensiones. El modelo Word2vec empleado en este trabajo es *continuous skip-gram model (Skip-gram)* (Mikolov et al., 2013).

Con el modelo entrenado, es posible obtener un conjunto de palabras, o *embeddings*, asociadas a una entrada definida por un *query* Q. Word2vec recibe un término Q y devuelve un léxico $L(Q) = (w_1, w_2, ..., w_m)$, que representa un conjunto de $m = 10$ palabras semánticamente próximas a Q. El valor de m fue definido de esta manera ya que se percibió que, mientras más se extiende el número de palabras proximas a Q, estas pierden más su relación con respecto a Q. Formalmente, representamos Word2vec: $Q \rightarrow L(Q)$.

4.2.2 Interpretación geométrica

En esta parte, determinamos las palabras más adecuadas que sustituirán las etiquetas POS de una EGP y así generar una nueva frase. Para cada etiqueta POS_k, $k = 1, 2, ...$ $\in$ EGP, que se desea sustituir, usamos el algoritmo descrito a continuación.

Se efectúa un análisis morfosintáctico del corpus 5KL usando FreeLing y se usan las etiquetas POS para crear conjuntos de palabras que posean la misma información gramatical (etiquetas POS idénticas). Una Tabla Asociativa (TA) es generada como resultado de este proceso. La TA consiste en entradas de pares POS_k y una lista de palabras asociadas. Formalmente, se reprenta $\text{POS}_k \rightarrow V_k = \{v_{k,1}, v_{k,2}, ..., v_{k,i}\}$.

Luego se construye un vector para cada una de las tres palabras siguientes.

- o: es la palabra k de la frase f_o correspondiente a la etiqueta POS_k. Esta palabra permite recrear un contexto del cual la nueva frase debe alejarse, evitando producir una paráfrasis.
- Q: es la palabra que define al *query* proporcionado por el usuario.
- w: es la palabra candidata que podría reemplazar POS_k. Esta palabra pertenece a un vocabulario V_k de tamaño $|V_k| = m$ palabras, $w \in V_k$, que es recuperado de TA.

Las 10 palabras o_i más próximas a o, las 10 palabras Q_i más próximas a Q y las 10 palabras w_i más próximas a w (en este orden y obtenidas con Word2vec), son concatenadas y representadas en un vector simbólico $\vec{U}$ de 30 dimensiones. El número de dimensiones fue fijado a 30 de manera empírica, como un compromiso razonable entre diversidad léxica y tiempo de procesamiento.

El vector $\vec{U}$ puede ser escrito como

$$\vec{U} = (u_1, ..., u_{10}, u_{11}, ..., u_{20}, u_{21}, ..., u_{30})$$

donde cada elemento $u_j, j = 1, ..., 10$, representa una palabra próxima a o; $u_j, j = 11, ..., 20$, representa una palabra próxima a Q; y $u_j, j = 21, ..., 30$, es una palabra próxima a w. $\vec{U}$ puede ser re-escrito de la siguiente manera,

$$\vec{U} = (o_1, ..., o_{10}, Q_{11}, ..., Q_{20}, w_{21}, ..., w_{30})$$

o, Q y w generan respectivamente tres vectores numéricos de 30 dimensiones:

$$\begin{aligned} o : \vec{X} &= (x_1, ..., x_{30}) \\ Q : \vec{Q} &= (q_1, .., q_{30}) \\ w : \vec{W} &= (w_1, ..., w_{30}) \end{aligned}$$

donde los valores de $\vec{X}$ son obtenidos tomando la distancia entre la palabra o y cada palabra $u_j \in \vec{U}, j = 1, ..., 30$. La distancia, $x_j = dist(o, u_j)$ es recuperada de Word2vec, donde $x_j \in [0, 1]$. Evidentemente la palabra o estará más próxima a las 10 primeras palabras u_j que a las restantes.

Un proceso similar permite obtener los valores de $\vec{Q}$ y $\vec{W}$ a partir de Q y w, respectivamente. En estos casos, el *query* Q estará más próximo a las palabras u_j en las posiciones $j = 11, ..., 20$ y la palabra candidata w estará más próxima a las palabras u_j en las posiciones $j = 21, ..., 30$.

Enseguida, se calculan las similitudes coseno entre $\vec{Q}$ y $\vec{W}$ (1) y entre $\vec{X}$ y $\vec{W}$ (2),

$$\theta = \cos(\vec{Q}, \vec{W}) = \frac{\vec{Q} \cdot \vec{W}}{|\vec{Q}||\vec{W}|} \quad (1)$$

$$\beta = \cos(\vec{X}, \vec{W}) = \frac{\vec{X} \cdot \vec{W}}{|\vec{X}||\vec{W}|} \quad (2)$$

Estos valores de θ y β están normalizados en [0,1]. Se itera el proceso para todas las palabras del léxico $w \in V_k$. Esto genera otro conjunto de vectores $\vec{X}, \vec{Q}$ y $\vec{W}$ para los cuales se deberán calcular nuevamente las similitudes. Al final se obtienen m valores de similitudes θ_i y β_i, $i = 1, ..., m$, y se calculan los promedios $\langle\theta\rangle$ y $\langle\beta\rangle$.

El cociente normalizado $\left(\frac{\langle\theta\rangle}{\theta_i}\right)$ indica qué tan grande es la similitud de θ_i con respecto al promedio $\langle\theta\rangle$ (interpretación de tipo maximización); es decir, que tan próxima se encuentra la palabra candidata w al *query* Q. El cociente normalizado $\left(\frac{\beta_i}{\langle\beta\rangle}\right)$ indica qué tan reducida es la similitud de β_i con respecto a $\langle\beta\rangle$ (interpretación de tipo minimización); es decir, qué tan lejos se encuentra la palabra candidata w de la palabra o de f_o.

Estas fracciones se obtienen en cada par (θ_i, β_i) y se combinan (minimización-maximización) para calcular un score S_i, según la ecuación

$$S_i = \left(\frac{\langle\theta\rangle}{\theta_i}\right) \cdot \left(\frac{\beta_i}{\langle\beta\rangle}\right) \tag{3}$$

Mientras más elevado sea el valor S_i, mejor obedece a nuestros objetivos: acercarse al *query* y alejarse de la semántica original.

Finalmente, ordenamos en forma decreciente la lista de valores de S_i y se escoge, de manera aleatoria, entre los 3 primeros, la palabra candidata w que reemplazará la etiqueta POS_k en cuestión. El resultado es una nueva frase $f_3(Q, N)$ que no existe en los corpora utilizados para construir el modelo (ver Figura 2).

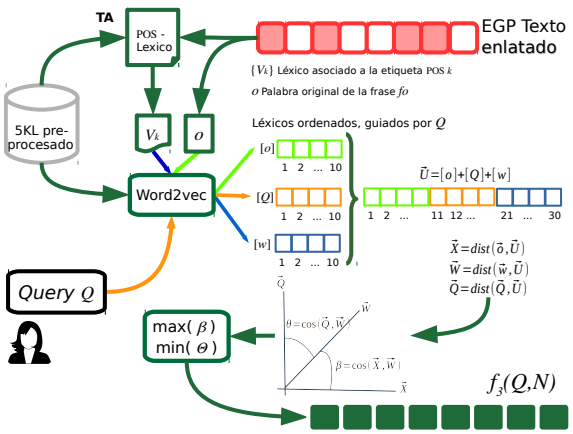

Figura 2: Aproximación semántica basada en interpretación geométrica min-max

5 *Resultados y evaluación*

Dados las características de nuestro modelo (idioma, corpora empleados, generación inspirada en homosintaxis), no es posible compararse con otros, ya que estos se centran en áreas específicas de la literatura y no parten de un enfoque general como es el caso en este trabajo. Sin embargo, previo a este modelo, realizamos dos experimentos basados en métodos estocásticos, empleando Cadenas de Markov, y un integración de *Canned Text* y Word2vec, con un enfoque simplificado. Los resultados de esos experimentos se detallan en (Moreno-Jiménez, Torres-Moreno, y Wedemann, 2020), pero a grosso modo los criterios evaluados así como la escala empleada fueron los mismos que se aplicaron en este trabajo. Para el modelo basado en Markov, se obtuvieron los siguientes resultados: gramaticalidad=0,55, coherencia=0,25 y contexto=0,67. El experimento integrado por *Canned Text* y Word2vec obtuvo: gramaticalidad=0,74, coherencia=0,56 y contexto=0,35.

Los resultados de esos experimentos nos permitieron detectar algunas deficiencias en los modelos iniciales, y proponer el modelo actual, cuyos resultados se muestran a continuación.

5.1 Resultados

Presentamos algunos ejemplos de frases generadas por nuestro modelo. Para el *query* Q, y una longitud en número de palabras N, los resultados se muestran en el formato $f(Q, N) =$ frase generada.

1. $f($`AMOR,10`$) =$ `En el aprecio está el cariño forzoso de una simpatía.`
2. $f($`AMOR,10`$) =$ `Los cariños no conocen de nada a un respeto loco.`
3. $f($`AMOR,10`$) =$ `No está la simpatía en las bondades de la envidia.`
4. $f($`GUERRA,9`$) =$ `Existe demasiada innovacion en torno a muy pocos sucesos.`
5. $f($`GUERRA,9`$) =$ `En la pelea todo debe motivo, menos la retirada.`
6. $f($`GUERRA,10`$) =$ `La codicia, siempre adversa, es terrible engendrada contra un desgraciado.`
7. $f($`SOL,9`$) =$ `Si tus dulces fueran amanecer, mis ojos marchitas fueran.`
8. $f($`SOL,11`$) =$ `Con rapidez, los monógamos impedimentos buscan para iluminar nos la luz.`
9. $f($`SOL,10`$) =$ `Incluso los luceros ingratos son comilones, y por tanto antiguos.`

5.2 Evaluación

Presentamos en esta sección un protocolo de evaluación manual. El experimento completo consistió en la generación de 45 frases. Se consideraron tres *queries*, $Q = \{$`AMOR, GUERRA, SOL`$\}$ y se generaron 15 frases de cada uno. Las frases fueron mezcladas antes de presentarlas a los 7 evaluadores seleccionados. Los evaluadores se eligieron considerando que poseen estudios universitarios y son hispanohablantes nativos, además de cierto hábito a la lectura que les permita la buena compresión de este tipo de textos. Se les pidió evaluar, usando la escala 0=mal, 1=aceptable y 2=correcto, los criterios siguientes:

- **Gramaticalidad:** ortografía, conjugaciones correctas, concordancia en género y número.
- **Coherencia:** legibilidad, percepción de una idea general.
- **Contexto:** relación de la frase con respecto al *query*.

Se realizó una adaptación del Turing. A los evaluadores se les hizo creer que había algunas frases escritas por personas y otras escritas por los algoritmos. Se les pidió indicar cuáles frases pensaban que habían sido generadas por personas (0) y cuáles por algoritmos (1). La Tabla 3 presenta la media aritmética y la desviación estándar de los resultados obtenidos normalizados a una escala entre 0-1.

Criterio	Resultados
Gramaticalidad	**0.77 ± 0.13**
Coherencia	**0.60 ± 0.14**
Contexto	**0.53 ± 0.19**
Turing	**0.44 ± 0.15**

Tabla 3: Evaluación del sistema

Se observa que las frases son percibidas como gramaticales y coherentes, con una evaluación de 0,77 y 0,60 respectivamente. El contexto obtuvo un resultado más bajo. Esto puede deberse a que las EGP contienen elementos fijos (palabras funcionales) que en ocasiones pueden alterar el contexto o semántica de las palabras insertadas. A pesar de ello, los resultados desde una perspectiva general son bastante alentadores.

En el test de Turing, los evaluadores perciben un 44 % como frases generadas por una persona. Esto, aunque parece ser un score bajo, comparado con los trabajos relacionados con este tema, resulta ser una buena evaluación, considerando que el objetivo no es generar texto aleatorio, sino un texto con características literarias.

6 Conclusiones

En este trabajo presentamos una primera aproximación de un modelo generativo de texto literario usando modelos neuronales de tipo Word embedding. El modelo produce frases aisladas en español con un cierto contenido literario. No se requiere prácticamente intervención del usuario (excepto por el query y la longitud de la frase requerida). La generación de párrafos y el uso de rimas sera el objeto de trabajos futuros (Medina-Urrea y Torres-Moreno, 2019). Dada la estructura del modelo propuesto, la extensión a otros idiomas (francés y portugués) también será contemplada (Charton y Torres-Moreno, 2011).

Bibliografía

Agirrezabal, M., B. Arrieta, A. Astigarraga, y M. Hulden. 2013. Pos-tag based poetry generation with wordnet. En *14th European Workshop on Natural Language Generation*, páginas 162–166. ACL.

Bengio, Y., A. Courville, y P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Boden, M. A. 2004. *The creative mind: Myths and Mechanisms*. Routledge.

Bracewell, D., F. Ren, y S. Kuriowa. 2005. Multilingual single document keyword extraction for information retrieval. En *2005 International Conference on Natural Language Processing and Knowledge Engineering*, páginas 517–522, Wuhan, China. IEEE.

Brantley, K., K. Cho, H. Daumé, y S. Welleck. 2019. Non-monotonic sequential text generation. En *2019 Workshop on Widening NLP*, páginas 57–59, Florence, Italy. ACL.

Charton, E. y J.-M. Torres-Moreno. 2011. Automatic modeling of logical connectors by statistical analysis of context. *Canadian Journal of Information and Library Science*, 35(3):287–306.

Colton, S. 2012. *Automated theory formation in pure mathematics*. Springer Science & Business Media.

Colton, S., G. A. Wiggins, y others. 2012. Computational creativity: The final frontier? En *20th European Conference on Artificial Intelligence*, páginas 21–26. ACL.

Fu, R., J. Guo, B. Qin, W. Che, H. Wang, y T. Liu. 2014. Learning semantic hierarchies via word embeddings. En *52nd Annual Meeting of the ACL*, volumen 1, páginas 1199–1209, Baltimore, Maryland, USA. ACL.

Huang, E. H., R. Socher, C. D. Manning, y A. Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. En *50th Annual Meeting of the ACL*, volumen 1, página 873–882, USA. ACL.

Lebret, R., D. Grangier, y M. Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv*, página arXiv 1603.07771.

McRoy, S., S. Channarukul, y S. Ali. 2003. An augmented template-based approach to text realization. *Natural Language Engineering*, 9:381 – 420.

Medina-Urrea, A. y J.-M. Torres-Moreno. 2019. Rimax: Ranking semantic rhymes by calculating definition similarity.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., páginas 3111–3119.

Mikolov, T., W.-t. Yih, y G. Zweig. 2013. Linguistic regularities in continuous space word representations. En *NACACL: Human Language Technologies*, páginas 746–751, Atlanta, Georgia, USA. ACL.

Mikolov, T. y G. Zweig. 2012. Context dependent recurrent neural network language model. En *2012 IEEE Spoken Language Technology Workshop (SLT)*, páginas 234–239, Miami, FL, USA. IEEE.

Molins, P. y G. Lapalme. 2015. JSrealB: A bilingual text realizer for web programming. En *15th ENLG*, páginas 109–111, Brighton, UK. ACL.

Moreno-Jiménez, L.-G., J.-M. Torres-Moreno, y R. S. Wedemann. 2020. Generación automática de frases literarias. *Linguamatica*. Accepted.

Oliveira, H. G. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. En *10th ICNLG*, páginas 11–20.

Oliveira, H. G. 2012. Poetryme: a versatile platform for poetry generation. En *Computational Creativity, Concept Invention and General Intelligence*, volumen 1, Osnabrück, Germany. Institute of Cognitive Science.

Oliveira, H. G. y A. Cardoso. 2015. Poetry generation with poetryme. En *Computational Creativity Research: Towards Creative Machines*, volumen 7, Paris, France. Atlantis Thinking Machines.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *8th LREC)*, Istanbul, Turkey.

Pérez y Pérez, R. 2015. *Creatividad Computacional*. Larousse - Grupo Editorial Patria, México.

Sridhara, G., E. Hill, D. Muppaneni, L. Pollock, y K. Vijay-Shanker. 2010. Towards automatically generating summary comments for java methods. En *IEEE/ACM International Conference on Automated Software Engineering*, página 43–52, Antwerp, Belgium. ACM.

Szymanski, G. y Z. Ciota. 2002. Hidden markov models suitable for text generation. En N. Mastorakis V. Kluev, y D. Koruga, editores, *WSEAS*, páginas 3081–3084, Athens, Greece. WSEAS - Press.

Torres-Moreno, J.-M. 2014. *Automatic Text Summarization*. ISTE, Wiley, London, UK, Hoboken, USA.

van Deemter, K., M. Theune, y E. Krahmer. 2005. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24.

Zhang, X. y M. Lapata. 2014. Chinese poetry generation with recurrent neural networks. En *2014 EMNLP*, páginas 670–680, Doha, Qatar. ACL.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 37-44 recibido 30-04-2020 revisado 04-06-2020 aceptado 04-06-2020

Cross-lingual Training for Multiple-Choice Question Answering

Entrenamiento Croslingüe para Búsqueda de Respuestas de Opción Múltiple

Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas
Universidad Nacional de Educación a Distancia (UNED)
{gblanco, alvarory, anselmo}@lsi.uned.es

Abstract: In this work we explore to what extent multilingual models can be trained for one language and applied to a different one for the task of Multiple Choice Question Answering. We employ the RACE dataset to fine-tune both a monolingual and a multilingual models and apply these models to another different collections in different languages. The results show that both monolingual and multilingual models can be zero-shot transferred to a different dataset in the same language maintaining its performance. Besides, the multilingual model still performs good when it is applied to a different target language. Additionally, we find that exams that are more difficult to humans are harder for machines too. Finally, we advance the state-of-the-art for the QA4MRE Entrance Exams dataset in several languages.
Keywords: Question Answering; Multiple-Choice Reading Comprehension; Multilinguality

Resumen: En este trabajo exploramos en qué medida los modelos multilingües pueden ser entrenados para un solo idioma y aplicados a otro diferente para la tarea de respuesta a preguntas de opción múltiple. Empleamos el conjunto de datos RACE para ajustar tanto un modelo monolingüe como multilingüe y aplicamos estos modelos a otras colecciones en idiomas diferentes. Los resultados muestran que tanto los modelos monolingües como los multilingües pueden transferirse a un conjunto de datos diferente en el mismo idioma manteniendo su rendimiento. Además, el modelo multilingüe todavía funciona bien cuando se aplica a un idioma de destino diferente. Asimismo, hemos comprobado que los exámenes que son más difíciles para los humanos también son más difíciles para las máquinas. Finalmente, avanzamos el estado del arte para el conjunto de datos QA4MRE Entrance Exams en varios idiomas.
Palabras clave: Búsqueda de Respuestas; Opción múltiple; Multilingüismo

1 Introduction

Question Answering (QA) has manifold dimensions, depending on the source of the information (i.e. free text vs. knowledge bases), and how the question is to be responded. In this work, we focus on Multiple-Choice QA, where systems has to select the correct answer from a set of candidates according to a given text. This format is usually applied to evaluate language understanding with humans.

To solve this task, the tendency has steered towards deep, attention-based language models, like BERT, or XLNET ((Devlin et al., 2019; Yang et al., 2019)), based on the idea of transformers (Vaswani et al., 2017). These models have boosted all available NLP tasks, becoming the go-to technique in many cases, and QA is not an exception.

These models could be a handicap for underrepresented languages in terms of available resources to train them. In addition, the training of these models requires a huge computation power. Hence, there is a great interest from the research community towards developing multilingual models trained once for many different languages. So far, mono-

ISSN 1135-5948. DOI 10.26342/2020-65-4

lingual models still seem to perform better than multilingual ones for some tasks (Martin et al., 2019; Agerri et al., 2020; Cañete et al., 2020).

In order to use these pre-trained models, systems must be fine-tuned to the target task. Unfortunately, the majority of languages don't have large enough datasets for several tasks, this is the case of Multiple-Choice QA. Given the size of the RACE dataset (Lai et al., 2017), it is possible to fine-tune an English model with it, but you cannot do the same for other languages as the datasets are too small. Such is the case of Spanish or Italian. The creation of these datasets is usually very costly both in time and money. Thus, the majority of collections are either too small to train a deep model or are only available in English (Hsu, Liu, and Lee, 2019).

In this work, we study how to apply the models trained with a dataset in a language to a different collection in another language. For this purpose, we use the RACE collection to train a model and we test the model using the Entrance Exams (EE) datasets, which are available in several languages (Rodrigo et al., 2018).

According to these observations, the objectives of this work are:

i) Compare the results obtained in EE and RACE. In principle, they target similar human language skills: middle and high school English level in the case of RACE and university admission in the case of EE.

ii) Determine if there is a correlation between the difficulty of the exercises for both humans and computers.

iii) Determine if the knowledge learnt by fine-tuning with a collection (RACE in English) can be transferred to perform with another collection (EE English and other languages).

iv) Test the performance of both monolingual and multilingual BERT models once they are trained in one language and evaluated in different ones.

v) Advance the state-of-the-art for the EE task in various languages.

These objectives are motivated by the following research questions:

RQ 1 How monolingual and multilingual models perform for Multiple-Choice QA when they are trained for a specific language to work in a different one?

RQ 2 When monolingual and multilingual models are trained and tested for the same language, is their performance comparable?

RQ 3 Can multilingual models advance the current state-of-the-art for some languages where there is not enough training data?

2 Related Work

Question Answering (QA) is the task of returning a precise and short answer given a Natural Language question. QA can be approached from two main perspectives (Rogers et al., 2020): 1) Open QA, where systems collect evidences and answers across several sources such as Web pages and knowledge bases (Fader, Zettlemoyer, and Etzioni, 2013) and, 2) Reading Comprehension (RC), where the answer is gathered from a single document.

RC systems can be oriented to: (1) extract spans of text with the answer (extractive QA), (2) select an answer from a set of candidates (multiple-choice QA) or (3) generate an answer (generative QA). Extractive QA has received a lot of attention fostered by the availability of popular benchmarks such as SQuAD (Rajpurkar, Jia, and Liang, 2018). On the other hand, generative QA has received less attention given that it is difficult to perform an exact evaluation and there are few datasets (Kočiský et al., 2018).

In this work we focus on Multiple-Choice (MC) QA. Since MC is a common way to measure reading comprehension in humans, the task is very realistic. In fact, the datasets employed in this research (presented ahead) are based on real world exams. Besides, some researches have pointed MC format as a better format to test language understanding of automatic systems (Rogers et al., 2020).

There exists several MC collections, mostly in English. In some cases it involves paying crowd-workers to gather documents and/or pose questions regarding those documents. MCTest (Richardson, Burges, and Renshaw, 2013), for example, proposed for the workers to invent short, children friendly, fictional stories and four questions

with four answers each, including deliberately wrong answers. As a way to encourage a deeper understanding of texts, the QuAIL dataset includes unanswerable questions (Rogers, Kovaleva, and Rumshisky, 2020). Other datasets were created from real world exams. This is the case of the well known MC dataset RACE (Lai et al., 2017), or the multilingual Entrance Exams (Rodrigo et al., 2018), described in more detail in the next Section.

When doing QA, there usually exists the constraint on the language and size of the datasets available. In this sense, many times there is not enough training data to fine-tune a model in a specific language. (Asai et al., 2018) tried to solve this issue by translating the target collection to a language with enough training data and using a QA system trained in the second language. However, this approach relies too much on the quality of the translation.

A common practice to fill this gap is zero-shot learning, which aims to solve a task without receiving any example of that task at the training phase. That is, we fine-tune for task A and evaluate in task B, possibly in another language. Thus, we expect for the knowledge to be transferred from one task to the other (in another language) with minimum overhead.

We have found similar efforts in the literature. Hsu, Liu, and Lee (2019), for example, studies how to train Multi BERT for extractive QA in a language to test it in another language, they obtain promising results. We differ from them in: (1) the languages employed: they test English, Korean and Chinese; (2) the task: they work on extractive QA and; (3) the type of collections: we use collections crafted for human evaluation, which allows us to study how difficulty for humans correlates with difficulty for automatic systems. More specifically, we zero-shot transfer a model from RACE (fine-tune) to Entrance Exams (no training data available) in multiple, unseen languages.

3 Datasets

In our experiments we use RACE and Entrance Exams. Both collections are derived from real human evaluations. The following subsection gives further details of each collection.

3.1 RACE

RACE (Lai et al., 2017) was collected from the English exams for middle (subset named RACE-M) and high (subset named RACE-H) school Chinese students. There are two subsets depending on the level of the exams: *RACE-M* for middle school and *RACE-H* for high school.

The authors proposed it to evaluate the reading comprehension task using MC format. RACE consists of more that 100K questions generated from human experts (English instructors). Table 1 shows the details of RACE. We can see in the table that RACE-H contains more data than RACE-M.

In order to evaluate the difficulty of the collection, the authors employed Amazon Mechanical Turk[1] to annotate question types of a subset. The authors found a higher ratio of reasoning questions with respect to CNN (Hermann et al., 2015), SQUAD and NEWSQA (Trischler et al., 2017), which justifies that RACE is more difficult than those datasets.

3.2 Entrance Exams

The Entrance Exams (EE) data was collected from standardized English examinations for university admission in Japan and used in the Entrance Exams task at CLEF in 2013, 2014 and 2015 (Rodrigo et al., 2018). Exams were created by the Japanese National Center for University Admission Tests. Only the exams with MC format were included in the dataset. We show in Table 2 the number of documents and questions released in each edition, as well as the number for the overall set.

The organizers of the task proposed also the same task in other languages different from English by collecting parallel translations from volunteers at the translation for progress website[2]. EE data is also available in French, Italian, Spanish and Russian. Translations for German are only available for the 2015 dataset. Thus, EE allows to test QA systems in other languages besides English, which is the language where almost all the QA datasets are available. However, EE received only participants for the English and French tasks. Therefore, this paper represents the first attempt to solve EE in the other languages.

[1] https://www.mturk.com/mturk/welcome

[2] `http://www.translationsforprogress.org/`

Dataset	RACE-M			RACE-H			RACE			
Subset	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	All
# documents	6,409	368	362	18,728	1,021	1,045	25,137	1,389	1,407	27,933
# questions	25,421	1,436	1,436	62,445	3,451	3,498	87,866	4,887	4,934	97,687

Table 1: Details of RACE-M, RACE-H and RACE collections

	2013	2014	2015	All
# documents	9	12	19	40
# questions	46	56	89	191

Table 2: Details of Entrance Exams collection

We want to remark also that the EE organizers proposed two kinds of evaluation. The first one is based in traditional QA evaluation, measuring the overall performance of a system over the whole set of questions. The second approach proposed to measure the number of tests passed by a system. According to this approach, each test is made of a document and the questions about it. Then, a system passes a test if it manages to answer correctly to 50% or more of the questions, similar to human evaluations.

4 BERT and Multilingual BERT

BERT and Multilingual BERT (M-BERT from now on) are transformer-based language representation models. They have been pre-trained from unlabeled text to do Masked Language Modeling and Next Sentence Prediction (Devlin et al., 2019). Afterwards, each model can be fine-tuned in specific tasks such as those at Glue (Wang et al., 2018) or QA. Albeit both models share the same architecture (a twelve layer transformer), they were trained with different corpus. BERT was trained with BooksCorpus (800M words) and M-BERT with the Wikipedia in 104 different languages. Even so, both use the same word piece vocabulary (and tokenizer) and have no information about the language in training.

The idea behind M-BERT is to learn all languages at once, delivering a single model capable of operating in multiple languages. There are several caveats with this approach:

1. Languages compete against one another for a fraction of hyper parameters, affecting underrepresented languages. Although oversampling is applied, it is not completely solved (Artetxe, Ruder, and Yogatama, 2019).
2. No language specific knowledge is used to improve in any part of the model. Intuitively, one should be able to pre-train a BERT model in any language, applying language-specific knowledge and improve over M-BERT (e.g.: In (Agerri et al., 2020), the authors employ a basque-specific word piece vocabulary to improve the basque model).

In this paper, we compare both models and their ability to perform zero-shot cross-lingual transfer in multiple-choice QA. We describe the experiments in the next Section.

5 Experiments

To the best of our knowledge, there exists no Spanish MC collections big enough to fine tune a model. So, we have fine-tuned a model on English RACE and test it on other languages.

In our experiment, we use a simple BERT model and a M-BERT. Each model has been fine-tuned over RACE for three epochs, as recommended by the developers. We have employed the well known transformers library from huggingface[3], following the hyper-parameters stated by the first BERT[4] base model result on RACE's leaderboard[5]. Additionally, all the source code is available in a Github repository[6]. Every model was trained on Google Cloud Platform with a Tesla T4 for three epochs.

The experiments followed with each model are:

1. Fine-tune model on RACE train collection, with both high and middle splits.
2. Measure[7] performance on RACE test collection.

[3] https://huggingface.co/transformers/

[4] https://github.com/NoviScl/BERT-RACE

[5] http://www.qizhexie.com/data/RACE_leaderboard.html

[6] https://github.com/m0n0l0c0/race-experiments

[7] We use accuracy

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	**0.6114**	0.2500	0.3078
RACE High	0.4774	**0.5031**	0.2500	0.3059
RACE All	0.4917	**0.5347**	0.2500	0.3059
EE English	0.4921	**0.4974**	0.2500	0.2304
EE Spanish	0.3665	**0.4503**	0.2500	0.2932
EE Italian	0.2880	**0.4293**	0.2500	0.2775
EE French	0.3037	**0.4346**	0.2500	0.2565
EE Russian	0.2618	**0.3403**	0.2500	0.2723
EE German**	0.3708	**0.4494**	0.2500	0.2584

Table 3: Accuracy of each model, including baselines: BERT, M-BERT, Random and Longest over each dataset RACE (every split) and Entrance Exams (over every year and language) **German data is a single result, there is data only for 2015

3. Measure performance on EE in: English, Spanish, Italian, French, Russian and German

We have also ran two baselines to establish a lower bound every model should surpass. The first baseline choose every answer at random. Since we have four candidates per question, this baseline achieves an accuracy score of 0.25. The second, following the work of (Rogers et al., 2020), just yields the longest answer.

6 Results

Table 3 shows the results, according to accuracy, obtained with the conducted experiments. We list the results obtained in each dataset (RACE and Entrance Exams) by each employed model (BERT, M-BERT, Random and Longest baselines). In all cases, we report the results for the test split. For Entrance Exams we show the results for all years averaged together.

BERT scores similar in RACE and Entrance Exams, though it obtains its best scores in RACE middle. Even so, it is outperformed by M-BERT in all cases. The latter performs better in RACE than Entrance Exams, which are harder. Both models are above the baselines excluding BERT with Russian EE.

Both models' scores visibly decrease when raising the education grade. RACE middle is the highest scored, which corresponds to middle school exams, the easiest of the three collections for humans. Following, RACE high and, in the last place, Entrance Exams, which are tests for university entrance. This tendency matches that of the humans, when increasing in difficulty, students score lower.

Both models are above the baselines, excluding BERT's Entrance Exams - Russian result which is the worst by far.

Among the Entrance Exams collection, the English version obtains the highest score for both models. Taking into account that both models were fine-tuned with a single task in English, this was expected. On the other hand, EE Russian was the lowest scored collection. Our intuition is that it is due to using a different alphabet.

No model had previous clue of Entrance Exams task, it was zero-shot transferred from RACE. That is only available in English.

This means that in the case of BERT, there are languages that it has never seen before, because it is monolingual. Furthermore, it's performance worsens hardly when exposed to unseen languages, specially with Russian and German, this is the expected behavior since languages with very different semantics are not tokenized nor understood correctly by the model (Artetxe, Ruder, and Yogatama, 2019; Agerri et al., 2020). This is the case of Russian.

M-BERT scores above 0.4 in all cases but Russian. Additionally, it is very close to passing the exam (to answer correctly at least 50% of the questions) for English.

Tables 4, 5 and 6 show the results of EE divided by years (from 2013 to 2015). We give results according to accuracy and the proportion of tests (a document with their corresponding questions) passed (an accuracy of at least 0.5). These results correspond to the two evaluation perspectives applied in EE and described in Section 3.2. We include results in each language for each model, the two baselines and the best performing system at EE (in each edition). There are results from

Entrance Exams 2013

	BERT		MultiBERT		Random		Longest		NIIJ-3*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	**0.43**	0.22	0.41	0.44	0.25	0	0.22	0.00	0.35	0.33
Spanish	0.37	0.33	**0.43**	0.22	0.25	0	0.22	0.11	-	-
Italian	0.28	0.11	**0.33**	0.22	0.25	0	0.28	0.22	-	-
French	0.35	0.11	**0.43**	0.33	0.25	0	0.17	0.00	-	-
Russian	0.22	0.11	**0.26**	0.00	0.25	0	0.17	0.00	-	-

Table 4: Accuracy of each model and proportion of passed tests, 2013 edition had 9 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2013. The best previous work (*) on (Rodrigo et al., 2018) from the National Institute of Informatics of Japan (Li et al., 2013). They originally presented results only for English

Entrance Exams 2014

	BERT		MultiBERT		Random		Longest		Synapse*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	0.50	0.58	**0.52**	0.67	0.25	0	0.30	0.33	0.45	0.58
Spanish	0.38	0.33	**0.45**	0.50	0.25	0	0.34	0.25	-	-
Italian	0.32	0.33	**0.43**	0.33	0.25	0	0.29	0.17	-	-
French	0.30	0.17	0.48	0.50	0.25	0	0.30	0.25	**0.59**	0.75
Russian	0.30	0.17	0.32	0.17	0.25	0	**0.34**	0.17	-	-

Table 5: Accuracy of each model and proportion of passed tests, 2014 edition had 12 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2014. The best previous work (*) on (Rodrigo et al., 2018) from Synapse (Laurent et al., 2014). They originally presented results only for French and English

Entrance Exams 2015

	BERT		MultiBERT		Random		Longest		Synapse*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	0.52	0.63	0.53	0.63	0.25	0	0.19	0.21	**0.58**	0.84
Spanish	0.36	0.32	**0.46**	0.53	0.25	0	0.30	0.26	-	-
Italian	0.27	0.26	**0.48**	0.53	0.25	0	0.27	0.16	-	-
French	0.28	0.11	0.40	0.47	0.25	0	0.27	0.32	**0.56**	0.84
Russian	0.26	0.11	**0.39**	0.47	0.25	0	0.28	0.32	-	-
German**	0.37	0.42	**0.45**	0.58	0.25	0	0.26	0.16	-	-

Table 6: Accuracy of each model and proportion of passed tests, 2015 edition had 19 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2015. The best previous work (*) on (Rodrigo et al., 2018) from Synapse (Laurent et al., 2015). They originally presented results only for French and English
**This is the only year with German data

previous systems for English in all editions and French in 2014 and 2015. Thus, our work is setting the results for these collections in several languages.

BERT model shows its best result in English, 2013, but M-BERT passes a higher proportion of tests (almost twice). This means that BERT finds the correct answer for more questions than M-BERT but distributed across just a few documents, obtaining higher grades. However, M-BERT scores better in general, passing more than 44% of the tests.

The rest of results for 2013 are dominated by M-BERT model, which states a new best result in English, outperforming the previous systems. In the case of the second campaign, M-BERT lands second on the results table for French (where the best result is obtained by (Laurent et al., 2014)), and surpasses previ-

ous results in English. M-BERT model also sets the new results in Spanish and Italian (the best result for Russian is achieved by the longest baseline). From table 6, the best results in English and French go for Synapse (Laurent et al., 2015), who developed a complex system including background knowledge and trained over the previous dataset of EE. Thus, it seems that simple pre-trained models, transferred over a different dataset, are close to those results but not enough. For the rest of languages, the best results come from M-BERT model, that sets the state-of-the-art without requiring a dataset in those languages.

Overall, we can observe a best performance of M-BERT with respect to BERT, which is focused on the English language. In fact, according to our experiments, we can fine-tune M-BERT in English for MC using one dataset and transfer that knowledge to datasets in other languages.

7 Conclusions and Future Work

The results obtained show that both monolingual and multilingual models can be fine-tuned for task and transferred to another task in the same language. Furthermore, multilingual models are transferable also to different languages.

Also, we obtain evidence that systems performance is hampered by exams difficulty in the same way human grades do.

In this work we established the state-of-the-art results over the Entrance Exams task in four more languages.

We would like to continue pursuing methods to cope with low resource languages. To do so, we will continue exploring how fine-tuned transformer bodies can be transferred to reuse knowledge about specific tasks, following the lead of (Artetxe, Ruder, and Yogatama, 2019; Otegi et al., 2020).

Acknowledgments

This work has been funded by the Spanish Research Agency under CHIST-ERA LIHLITH project (PCIN-2017-085/AEI) and deepReading (RTI2018-096846-B-C21 / MCIU/AEI/FEDER,UE).

References

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your Text Representation Models some Love: the Case for Basque. mar.

Artetxe, M., S. Ruder, and D. Yogatama. 2019. On the Cross-lingual Transferability of Monolingual Representations. oct.

Asai, A., A. Eriguchi, K. Hashimoto, and Y. Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Cañete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *to appear in PML4DC at ICLR 2020*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, oct. Association for Computational Linguistics.

Fader, A., L. Zettlemoyer, and O. Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hsu, T.-Y., C.-L. Liu, and H.-y. Lee. 2019. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Kočiský, T., J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 785–794, apr.

Laurent, D., B. Chardon, S. Nègre, C. Pradel, and P. Séguéla. 2015. Reading comprehension at entrance exams 2015. In L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, editors, *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Laurent, D., B. Chardon, S. Nègre, and P. Séguéla. 2014. French Run of Synapse Développement at Entrance Exams 2014. In *CLEF (Working Notes)*, pages 1415–1426.

Li, X., R. Tian, N. L. T. Nguyen, Y. Miyao, and A. Aizawa. 2013. Question Answering System for Entrance Exams in QA4MRE. In *CLEF (Working Notes)*. Citeseer.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. 2019. CamemBERT: a Tasty French Language Model. nov.

Otegi, A., A. Agirre, J. A. Campos, A. Soroa, and E. Agirre. 2020. Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. In *12th International Conference on Language Resources and Evaluation*.

Rajpurkar, P., R. Jia, and P. Liang. 2018. Know What You Don{'}t Know: Unanswerable Questions for {SQ}u{AD}. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Richardson, M., C. J. C. Burges, and E. Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. Technical report, nov.

Rodrigo, A., A. Peñas, Y. Miyao, and N. Kando. 2018. Do systems pass university entrance exams? *Information Processing & Management*, 54(4):564–575, jul.

Rogers, A., O. Kovaleva, M. Downey, and A. Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks.

Rogers, A., O. Kovaleva, and A. Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. feb.

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 5998–6008.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 {EMNLP} Workshop {B}lackbox{NLP}: Analyzing and Interpreting Neural Networks for {NLP}*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., jun, pages 5754–5764.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 45-52 recibido 30-04-2020 revisado 05-06-2020 aceptado 05-06-2020

ContextMEL: Classifying Contextual Modifiers in Clinical Text

ContextMEL: un Clasificador de Modificadores Contextuales en Texto Clínico

Paula Chocrón, Álvaro Abella, Gabriel de Maeztu
IOMED
{paula.chocron,alvaro.abella,gabriel.maeztu}@iomed.es

Abstract: Taking advantage of electronic health records in clinical research requires the development of natural language processing tools to extract data from unstructured text in different languages. A key task is the detection of contextual modifiers, such as understanding whether a concept is negated or if it belongs to the past. We present ContextMEL, a method to build classifiers for contextual modifiers that is independent of the specific task and the language, allowing for a fast model development cycle. ContextMEL uses annotation by experts to build a curated dataset, and state-of-the-art deep learning architectures to train models with it. We discuss the application of ContextMEL for three modifiers, namely Negation, Temporality and Certainty, on Spanish and Catalan medical text. The metrics we obtain show our models are suitable for industrial use, outperforming commonly used rule-based approaches such as the NegEx algorithm.
Keywords: Clinical text, Temporality, Negation, Certainty, deep learning, annotation

Resumen: Las historias clínicas electrónicas pueden traer grandes avances en la investigación médica, pero requieren el desarrollo de herramientas para procesar texto no estructurado en diferentes idiomas. Una tarea clave es la detección de distintos modificadores contextuales, como el aspecto temporal de un concepto, o si está negado. En este trabajo presentamos ContextMEL, un método para construir clasificadores para modificadores contextuales que es independiente tanto de la tarea específica como del lenguaje, permitiendo un ciclo de desarrollo dinámico. ContextMEL usa anotaciones de expertos para crear un dataset curado, y las últimas tecnologías en aprendizaje profundo. En este artículo discutimos la aplicación de ContextMEL para tres modificadores contextuales (temporalidad, negación, y certeza) en texto médico en castellano y catalán. Los resultados obtenidos muestran que nuestros modelos pueden utilizarse en un entorno industrial, y que son más precisos que conocidos métodos basados en reglas, como el algoritmo NegEx.
Palabras clave: Texto clínico, clasificación, aprendizaje profundo, anotación

1 Introduction

The growing adoption of electronic health records (EHR) in hospitals provides a unique opportunity to analyze clinical text data automatically. Firstly, being able to efficiently and accurately extract data from these unstructured text corpora would greatly accelerate the analysis process. Secondly, it would make it easier to comply with the privacy policies that such sensitive text requires. Driven by these considerations, Natural Language Processing (NLP) applications for the clinical domain has been a very active field of research in the past decades.

The main goal of our NLP system is to find the units of observation (i.e data points) required by clinical research from the EHR of the patients. Therefore extracting data points from clinical text usually means, at its core, understanding whether a given written observation is true or false for each case. For example, a study may require the identification of all the cases of anemia, to later analyze pathologies that are commonly shared with that diagnosis. This requires as a first step a tool to identify concepts in a text, commonly known as a *concept extractor*. These tools standardly take advantage

ISSN 1135-5948. DOI 10.26342/2020-65-5

of some of the large and curated knowledge bases that exist for medical terminology such as SNOMED CT[1], performing matches between their entities and the ones that are found in the text. These matches can either use syntactic distances or rules (Abacha and Zweigenbaum, 2011), or machine learning (Torii, Wagholikar, and Liu, 2011).

To construct an appropriate dataset, however, it is also essential to be able to identify contextual modifiers that modify the concepts in question. For example, consider the following text: *Previous episodes of Anemia, the patient's Hgb levels are now normal.* Although a good concept extractor would find a reference to *anemia*, it is as important to determine that it refers to the past, and that it is therefore not an active pathology anymore. Another typical case in which contextual modification is pivotal is in the identification of whether a concept is negated or not. In the next section we discuss different approaches which have been proposed to address these kind of problems. These tools are, in general, language-specific, and designed to identify one specific contextual modifier (such as Negation). However, the field of clinical research is constantly changing, and it is very likely to have studies that will need to handle modifiers that have never been taken into account before.

In this paper we present ContextMEL, a system that allows to train models to perform contextual concept analysis, irrespective of their source language and the particular task. At its core, ContextMEL proceeds by annotating data and then training models with these annotations. Here, we propose to use established state-of-the-art deep learning models that are particularly well-suited to solve the problem of assigning labels to a concept in the context of a longer text. These models are trained and evaluated on three different context-related tasks, all applied to clinical data, using both Spanish and Catalan sources. Since our method and the deep learning models we use are domain and language independent, our method can be straightforwardly applied to other domains or languages. Our technique shows promising results for these tasks, and it outperforms one of the best-known methods for Negation detection.

[1] http://www.snomed.org/

2 Related Work

The problem of identifying contextual concept modifiers is very broad, and it has been researched extensively. We will restrict to discuss approaches developed for medical text. Most of the existent approaches tackle the detection of a particular contextual modifier, and for a specific language. We will discuss solutions that identify Negation, Certainty and Temporality.

The development of tools to determine contextual concept modifiers for medical text has evolved side by side with general progress in NLP techniques. Early approaches relied on syntactic and rule-based methods. For example, the NegEx algorithm (Chapman et al., 2001) applies a rule-based system to perform Negation identification in English. NegEx has been extended to include other modifiers in systems such as ConText (Harkema et al., 2009), which can also identify the experiencer (i.e. the patient or a member of her family) and a concept's temporal modifier. NegEx has also been extended to other languages. Relevant examples for Spanish are (Costumero et al., 2014; Stricker, Iacobacci, and Cotik, 2015) and `NegEx-MES`, which we will discuss later.

More recent techniques incorporate the use of machine learning methods to the problem of contextual modifier identification, such as Conditional Random Fields (Agarwal and Yu, 2010). The work in (Cruz Diaz et al., 2012) uses machine learning techniques to identify Negation and Speculation (Certainty) for Spanish medical text. Finally, the latest approaches rely on deep learning techniques. The work in (Dalloux, Claveau, and Grabar, 2019) uses Long-Short Term Memory (LSTM) networks to solve Negation and Certainty in French corpora, while (Fancellu, Lopez, and Webber, 2016) applies them for English. The work on (Tourille et al., 2017) uses them to detect temporal modifiers in English. Finally, the latest work on modifier detection for clinical text uses modern language models based on Transformers such as BERT. Examples are (Khandelwal and Sawant, 2019) for Negation in English, (García-Pablos, Perez, and Cuadros, 2020) for Negation in Spanish, and (Sergeeva et al., 2019), which also tackles Speculation.

3 Context-MEL: a General System for Contextual Classification

ContextMEL is a system to tackle the problem of *contextual modifier analysis*. This problem can be defined formally as follows: Let t be a text, and c be a concept formed by a subset of one or more words in t. Given a set of labels L, assign a label from L to c in the context of t. For example, in the problem of identifying negations, the set of labels is composed of $L = \{\texttt{Positive, Negative}\}$. Given $t =$ *"The patient does not have anemia."* and the concept $c =$ *anemia*, the objective of the task is to assign one of the two labels to c in t. In theory, concepts could be any n-gram. In practice, we will use (and therefore develop methods for) only those that are recognized as medical concepts by a knowledge base.

ContextMEL provides a general framework to build classification models that can be used to solve the problem of contextual modifier analysis for different tasks, independently of the language of the text or its domain. Concretely, the system consists of a first phase, during which it uses manual expert annotators to build a labeled dataset for training. In a second phase we use this dataset to train a classification model that can assign labels to concept-text pairs. We use well-known model architectures that were originally developed for the problem of Aspect-Based Sentiment Analysis, and we show that we can apply them without changes to a variety of other tasks.

The models that are obtained with ContextMEL can be applied directly on a concept extraction pipeline. First, texts are preprocessed and analyzed by a concept extractor which links expressions to entities in a knowledge base. Then, these concepts together with the original text are fed to the contextual models built with ContextMEL, which assigns them a label. The methods that conform our concept extractor are outside of the scope of the paper, and therefore are not discussed in depth here.

ContextMEL assumes the following resources are available: 1. A large corpora of unlabeled text. 2. A relatively basic system that finds *concepts*, or the entities that are of interest and would require contextual labels. 3. A team of expert annotators.

The first and the second items are normally available when working with clinical data. Unlabeled text abounds, and there are reliable, well-curated knowledge bases that can be useful to identify concepts. In our case, we used a simple NER system built on ULMS vocabulary, with most of the terms coming from the Spanish SNOMED CT (see (Torii, Wagholikar, and Liu, 2011) and (Soldaini and Goharian, 2016) for examples of other medical concept extraction systems). Annotators, on the other hand, are normally expensive. However, we will show that a relatively small amount of annotated data is necessary.

Our system aims to be as generalizable as possible, providing a procedure that can be adapted to different domains and contextual modifiers that need to be identified. In our particular case, we used the system to build models to classify concepts in medical text written in Spanish and Catalan. In many cases one same text contains bits of both languages. We used the framework to build models for three different contextual modifiers:

Temporality. The goal of the Temporality task is to understand whether a concept belongs to the *current episode*, to the *clinical history* or to the *future plan* of a patient. The set of labels for this modifier is {`Antecedent, Plan, Current Episode`}. For example, in *"Patient with kidney transplant in 2010. Has mild back pain. I prescribe Ibuprofen every 8 hours."*, the concept *kidney transplant* is `Antecedent`, *back pain* is `Current Episode`, and *Ibuprofen* is `Plan`.

Certainty. This task is about identifying whether a concept is certain, or if it refers to a hypothesis or conditional that could be true or not. The set of labels for this modifier is {`Certain, Uncertain`}. For example, in *"If there is high fever, take ibuprofen"*, the concept *fever* is a hypothesis, while *ibuprofen* is conditional (depends on the fever). They both are, therefore, uncertain. It could be argued that no concept is totally certain, since the doctor could be wrong. It is important to keep in mind that we are not trying to classify whether a concept is true or not, but rather if the doctor considers it as a fact. Certainty is sometimes referred to as *speculation* in the literature.

Negation. Negation is perhaps the simplest of the tasks. It aims to identify whether

a concept is negated or not in a text. The set of labels for this modifier is {`Positive, Negative`}. For example, in the sentence *"Patient with anemia had no liver abnormality."*, *anemia* is positive and *liver abnormality* is negative. We have already discussed related work for this task in the introduction.

4 *Annotation: Building a training dataset for contextual modifier analysis*

The first step of Context-MEL consists on producing a high-quality dataset that can be used for training a machine learning model. While the resulting datasets are specific to a particular contextual modifier, language, and domain, we provide a general framework to obtain annotations.

Suppose we are trying to solve a *contextual modifier analysis* problem with set of labels L. The annotation phase requires to employ a team of domain expert annotators, who will assign labels from L to pairs of text-concept. To do so, annotators have access to an interface where they can read a text with a highlighted concept. The highlighted concepts are obtained automatically from the text, using the automatic *concept matching* system we mentioned in the previous section. Next to the text there is a table where they can select labels from L (whether it is possible to select more than one label depends on the specific semantics of the contextual modifier being analyzed). Finally, annotators can choose to finish the annotation of the concept by pressing "OK" (meaning that they submit their labels), "NO" (meaning that there is something wrong with the concept) or "PASS" (meaning that they are unsure of the annotation and they prefer to jump to another concept). Once a concept is annotated, another one from the same text is highlighted, until the annotator labels them all and jumps to the next text.

In our particular case, we employed a team of five medical practitioners. To make annotation more efficient, and since we were developing models for the three modifiers (*Temporality, Certainty, Negation*) in parallel, we merged their labels to create one unique annotation task. We only included labels that were not the default one for each of the modifiers: `History, Plan, Uncertain` and `Negative`. Since we were mixing independent modifiers, we made labels not exclusive, and annotators could choose more than one. For example, they could mark the word *ibuprofen* in *"take ibuprofen if there is pain"* as `Plan, Uncertain`. Figure 1 shows an example of a concept annotation.

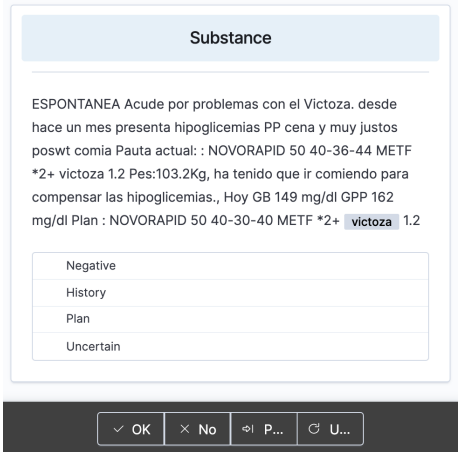

Figure 1: Labels for the annotator's task

The full annotation cycle lasted 3 months. We organized a meeting in person at the beginning and one midterm, to explain the task and clarify possible doubts.

We gave each pair concept-text to three different annotators, varying the composition of the groups. We considered an annotation as valid if at least two of the three annotators agreed on the label.

4.1 A more efficient annotation strategy: Confirming Entities First

We noticed rapidly of a problem in our approach: it relied too much on the automatic concept matcher, which was not perfect. In particular, our matcher had many false positives that were considered concepts although they should not be. This meant that annotators had to spend valuable time deciding how to label something that was not really a concept worth labeling. To tackle this problem, we set up a preliminary task that consisted on confirming whether a sequence of words was a medical concept or not. Annotators where very fast on this task, since they only had to choose between two labels. Later, we used these results on the main task, by assigning them only those concepts that had been verified as correct.

5 *Training: building classifiers for contextual modifier analysis*

Once we obtained the annotated dataset, we proceeded to train a model that would perform the labeling automatically. We propose to consider the problem of contextual modifier analysis as a generalization of *aspect-based sentiment analysis (ABSA)*. ABSA is a type of sentiment analysis that takes into account the fact that a text can express different sentiments about different things. For example, the text *"The movie has a very interesting plot, but I found the main actor bland"* is not positive or negative per-se, but rather positive about the plot and negative about the actor. ABSA can be defined as classifying the polarity of a concept c (which they call *aspect*) in a context t. This is equivalent to our problem when the set of labels L express polarity (for example `Positive, Negative`). The community working on solving ABSA has proposed different architectures to perform this kind of targeted classification; a survey can be found in (Schouten and Frasincar, 2015). For ContextMEL we compared the performance of an approach based on LSTMs and one based on BERT.

	Temp	Certainty	Negation
Train	6286/1000	1743/282	4931/689
Val	1459/255	384/84	1162/ 218
Test	261/79	246/90	173/75

Table 1: Size of the datasets, with the number of Spanish/Catalan examples on each one

Before discussing the models in the following subsections, let us comment on the datasets we used for training. Our data was naturally highly imbalanced (for example the *Temporality* task was skewed to the "Current Episode" label). To minimize the effect of this bias, we used balanced datasets that contained 65% of the prevalent class and 35% of the other one for the datasets with two labels (*Certainty* and *Negation*) and a $40\% - 20\% - 20\%$ split for the *Temporality* dataset. Table 1 shows the dimensions of the train and validation datasets that we used for training, together with a testset that we will discuss in the next section. As we can see, the datasets are mixed in Spanish and Catalan, although Catalan is notably underrepresented. Note that the datasets are not very large. For Certainty, for example, training with only 2025 examples (of which just 709 are uncertain) is enough to achieve the results we report next.

5.1 Contextual analysis via LSTM models

The first alternative we explored was to use a method proposed in (Tang et al., 2016). Their approach consists essentially in solving ABSA with two unidirectional LSTM networks, one modeling the part of the context to the left of the concept, and the other one the part to the right. The LSTM that models the right part receives the text reversed, so that the input to both networks ends with the concept. More formally, if the context c has n tokens and the concept to analyze is in positions i to $i + t$, the left LSTM receives input $c[0 : i + t]$ and the right LSTM receives $reverse(c[i : n])$. Using a concrete example, if the we want to classify the concept *anemia* in the text *"60 years old woman. History of anemia. Normal values now."*, the inputs for the left and right LSTMs would be:

Left: `60 years old woman . History of anemia`
Right: `. now values Normal . anemia`

These models are sometimes referred to as TD-LSTM (*Target-Dependent LSTM*).

We trained an LSTM with 1 layer for each of our context dimensions. To vectorize the input, we used a specialized embedding that we trained on our large corpora of unlabeled Spanish and Catalan texts. This embedding had 200 dimensions and was trained using word2vec (Mikolov et al., 2013). We trained the LSTMs for 15 epochs on each dataset. We were able to train these models on a machine with only one CPU and 8 GB of RAM memory.

5.2 Contextual analysis via BERT

The second approach we investigated was a BERT-based methodology. Language models such as BERT (Devlin et al., 2018) are neural networks based on the Transformer architecture (Vaswani et al., 2017) which have been trained on a huge dataset for a general task such as guessing a masked word. This setup makes them very flexible, which has proved very beneficial to perform transfer learning. In other words, the knowledge in the language model can be reused to solve

different specific tasks, needing only a small task-specific dataset.

BERT architectures have been used in different ways to solve ABSA (Zeng et al., 2019). In our case we used a very simple approach that took advantage of the type of input expected by BERT. In addition to the masked word task that we already mentioned, BERT is also trained for the *next sentence prediction* task. Shortly, the task consists on deciding whether two sentences are consecutive or not. To implement this task, the model uses a special token `[SEP]`, which indicates the separation between two sentences. If the input sentences are a and b, the model input is then [`[CLS]` a `[SEP]` b], where `[CLS]` is a special token used at the beginning of text. We used this mechanism to give, as input to the model, both the text and the concept, using `[SEP]` to indicate their boundaries. That is, for a pair of text t and concept c, our input was [`[CLS]` t `[SEP]` c]. Using the same example as before, to classify *anemia* in *"60 years old woman. History of anemia. Normal values now."* the input would be [`[CLS] 60 years old woman . History of anemia . Normal values now . [SEP] anemia [SEP]`].

Since our text is in Spanish and Catalan, we used the Multilingual (*bert-base-multilingual-cased*) version of BERT as a base, which we fine-tuned with the same datasets we used for the LSTM-version[2]. The models required only 3 epochs to achieve their best performance. To train the BERT models, we used a machine with 1 GPU and 30 GB RAM.

6 Results

In this section we present and discuss the results we obtained using both techniques (LSTM and BERT-based) for the three tasks we described before. To evaluate our models as accurately as possible, we built a golden testset for each task which was manually revised by us, using the annotator's data as a basis. For each label l (including a label `Blank` as a common one for `Positive, Current Episode, Certain`), we selected 100 examples that at least 2 annotators had marked with l, and 100 examples where only one one annotator had marked it as such (which often consist on harder or edge cases). Then we labeled the examples ourselves, and used the results as a golden test dataset.

To ensure a fair evaluation framework, we used in the testset only notes that had not been seen at train time. That is, we not only made sure the pair concept-text was previously unseen, but we also used completely new texts. Table 2 shows the results for the three different tasks. As we can see, we obtain good accuracy levels, which could be used industrially. Results are particularly good for the Negation task, which is unsurprising since it is notably simpler than the other two. Using BERT models improves the accuracy significantly for Tense and Certainty, which is expected since it is a very large model. When evaluating the results, it is important to keep in mind that training BERT requires a GPU and more memory, while the LSTMs can be trained on a much simpler computer; the size of the model would also impact the inference time. Notably, BERT models have better results even when they use the very generic Multilingual model as a base, while the LSTM ones use a vector space specifically designed for the task. Investigating whether further training the base model leads to better results is part of our planned future research.

In Table 3 we also show the precision ans recall values for each class. As we can see, the *Temporality* case is particularly good identifying future plans, which are normally very clear in the text, while the difference between Current Episode and Antecedent is sometimes more subtle. The Negation model is, as expected, better at identifying positives (the prevalent class). In Certainty, we observe the certain case is slightly better.

	Temp	Certainty	Negation
LSTM	81.82%	85.71%	95.70%
BERT	84.41%	88.10%	95.16%

Table 2: Accuracies for the three tasks for each model

We consider the possibility of building models specifically designed for a particular language, task, and domain to be one of the main advantages of our system. However, this makes comparing our techniques

[2]While there are Spanish-only versions of BERT, we are not aware of any Catalan-Spanish one, or even any Catalan-only version. Catalan is, however, included in the multilingual version of BERT

	Precision	Recall
Certain	85.06%	87.06%
Uncertain	86.42%	84.34%
Positive	96.67%	95.47%
Negative	92.86%	94.71%
Past	77.93%	78.03%
Present	77.89%	77.93%
Future	89%	88.80%

Table 3: Precision and recall per class for each task (LSTM)

NegEx	Context-MEL
83.14%	94.22%

Table 4: Results for the Spanish dataset

against other tools challenging. To the best of our knowledge, there are no available tools to perform temporal analysis of clinical text in Spanish and Catalan. The task of identifying negations has been the most prolific of the three we worked on. In what follows we show how our model compares with the `NegEx-MES`[3] system, which applies the `NegEx` technique to Spanish medical texts. `NegEx` is, at this moment, the best-known method to identify negated concepts, and its implementations are commonly applied industrially.

Since NegEx-MES was developed for Spanish texts, we compared the systems only for that portion of our testset (a 73.59%, according to Table 1). These results can be seen in table 4. As we can see, our system improves significantly the accuracy over NegEx-MES.

7 Conclusions and Future Work

We described ContextMEL, a system to build, in a general way, models to detect contextual modifiers. The framework includes an efficient way of annotating data by experts and state-of-the-art model architectures to harness the knowledge in the resulting dataset. Importantly, the development of the system is not bounded to a particular language or a particular task, allowing for a fast, simple, and organized model development cycle.

[3]https://github.com/PlanTL-SANIDAD/NegEx-MES

We applied ContextMEL for the identification of three context modifiers: Temporality, Negation, and Certainty. We obtained promising results that are already, after only one development iteration, suitable for industrial use. Moreover, the results for Negation outperformed the ones obtained with a commonly used rule-based tool. In summary, we think ContextMEL harnesses the power of the latest deep learning architectures to make the best possible use of a dataset that is carefully built, but not huge. For example, for Certainty we obtained an accuracy of 88.1% with only 2025 (and 709 uncertain) examples.

We have two clear directions of development and research for this project. First, we plan to study whether using a BERT model specially trained for Spanish and Catalan clinical text as a base for our fine-tuning would improve the results we already have. This model does not exist yet, so we plan to build it ourselves with our corpora of unlabeled data. If results are better, it would mean another step towards building a general framework to make fine-tuning for specific tasks even more lightweight.

As another line for future work we plan to work on a Quality Assessment-based development cycle to refine models using annotations. Our objective is to define a workflow that allows us to build the training dataset iteratively and on demand. This will be achieved by evaluating systematically the models that are obtained with each iteration until they reach an industry-suitable accuracy threshold. This avoids annotating unnecessarily large amounts of data, which is yet another way of making model development faster.

References

Abacha, A. B. and P. Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(S5):S4.

Agarwal, S. and H. Yu. 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American medical informatics association*, 17(6):696–701.

Chapman, W., W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. 2001. A simple algorithm for identifying negated

findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310, 11.

Costumero, R., F. Lopez, C. Gonzalo-Martín, M. Millan, and E. Menasalvas. 2014. An approach to detect negation on medical documents in spanish. In *Brain Informatics and Health*. Springer International Publishing, pages 366–375.

Cruz Diaz, N., M. López, J. Mata Vázquez, and V. Pachón. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology*, 63:1398–1410, 07.

Dalloux, C., V. Claveau, and N. Grabar. 2019. Speculation and negation detection in French biomedical corpora. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 223–232, Varna, Bulgaria, September. INCOMA Ltd.

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

Fancellu, F., A. Lopez, and B. Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 495–504.

García-Pablos, A., N. Perez, and M. Cuadros. 2020. Sensitive data detection and classification in spanish clinical text: Experiments with bert. *arXiv preprint arXiv:2003.03106*.

Harkema, H., J. N. Dowling, T. Thornblade, and W. W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42 5:839–51.

Khandelwal, A. and S. Sawant. 2019. Negbert: A transfer learning approach for negation detection and scope resolution.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space.

Schouten, K. and F. Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28:1–1, 01.

Sergeeva, E., H. Zhu, A. Tahmasebi, and P. Szolovits. 2019. Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187.

Soldaini, L. and N. Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Stricker, V., I. Iacobacci, and V. Cotik. 2015. Negated findings detection in radiology reports in spanish: an adaptation of negex to spanish. 07.

Tang, D., B. Qin, X. Feng, and T. Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Torii, M., K. Wagholikar, and H. Liu. 2011. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*, 18(5):580–587.

Tourille, J., O. Ferret, A. Neveol, and X. Tannier. 2017. Neural architecture for temporal relation extraction: a bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need.

Zeng, B., H. Yang, R. Xu, W. Zhou, and X. Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389, 08.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 53-58 recibido 19-03-2020 revisado 09-06-2020 aceptado 09-06-2020

Limitations of Neural Networks-based NER for Resume Data Extraction

Limitaciones en Reconocimiento de Entidades Nombradas (REN) basado en Redes Neuronales para la extracción de datos de Curriculum Vitae

Juan F. Pinzon, Stanislav Krainikovsky and Roman Samarev
dotin Inc. California, USA
info@dotin.us

Abstract: We provided research about the abilities of neural network-based NER models, quality, and their limitations in resolving entities of different types of complexity (emails, names, skills, etc.). It has been shown that the quality depends on the entity type and complexity, and estimate "ceilings", which model quality can achieve in the case of proper realization and well-labelled dataset.
Keywords: NER, Neural-Networks, Curriculum Vitae, Resumes, NLP, Data Extraction

Resumen: Proporcionamos investigación sobre las capacidades de los modelos REN basados en redes neuronales, la calidad y sus limitaciones para resolver entidades de diferentes tipos de dificultad (correos electrónicos, nombres, habilidades, etc.). Se ha demostrado que la calidad depende del tipo y la complejidad de la entidad, y estima los límites, que la calidad del modelo puede lograr en el caso de una realización adecuada y un corpus bien etiquetado.
Palabras clave: REN, Redes Neuronales, Curriculum Vitae, PLN, Extracción de Datos

1 Introduction

The task of parsing information from Resumes is of utmost importance and relevance in the NLP and Computer Linguistics communities. The Human Resources industry and recruiting companies will greatly benefit from its successful outcomes, not only reducing time and costs in the processing of CVs and their information but also making better quality matches between job postings and candidates. The impact and positive outcomes that these solutions can provide has been extensively discussed on research papers, such as *"The Impact of Semantic Web Technologies on Job Recruitment Processes"* (Bizer et al., 2005).

Resume Parsing solutions usually rely on complex rules statistical algorithms to correctly capture desired information from resumes. There are many variations of writing styles, words, syntax and to make things worse, the ever-increasing globalized world of nowadays means it is also important to take into account for cultural differences that affect the style and word choices among others. We intend to analyze if the Machine Learning and AI recent improvements can obtain better results for this task over rule-based approaches which use a predefined set of rules to extract the content.

Our main goal is to produce a machine learning model that will be able to extract main entities, such as name, email, education/university, companies worked at, skills, etc., from semi-structured and unstructured resumes while obtaining good quality and performance. We believe that the recent advancements in NLP can help achieve this task with smaller, but properly annotated, datasets and at the same time relying as little as possible on pre- or post-processing contrasting what was presented in the models of the "Resume Parser: Semi-structured Chinese Document Analysis" (Zhang et al., 2009) and "Study of Information Extraction in Resume" (Nguyen, Pham, and Vu, 2018) studies.

2 Methods

For this task, we used the named entity recognition (NER) approach, using the FlairNLP

ISSN 1135-5948. DOI 10.26342/2020-65-6

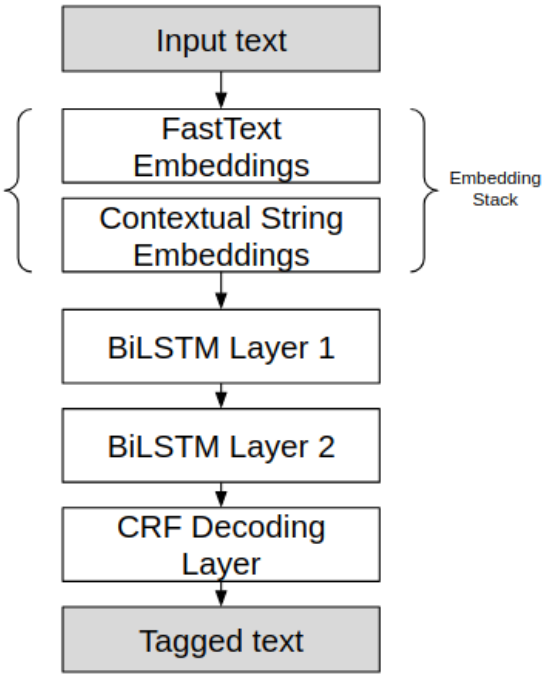

Figure 1: Neural Network NER Model Architecture

framework (Akbik et al., 2019). This framework allows us to easily implement the current state-of-the-art NER model that uses the LSTM variant of bidirectional recurrent neural networks (BiLSTMs) and a conditional random field (CRF) decoding layer, architecture depicted in Figure 1. The biggest advantage of using this framework was being able to leverage the power of Contextual String Embeddings - these embeddings can capture the hidden syntactic-semantic information, exceeding far beyond the standard word embeddings. Their distinct properties are: "(a) they are trained without any explicit notion of words and thus fundamentally model words as sequences of characters, and (b) they are contextualized by their surrounding text, meaning that the same word will have different embeddings depending on its contextual use" (Akbik, Blythe, and Vollgraf, 2018). Additionally, the stacking embeddings technique was used, which consists of combining different types of embedding models by concatenating each embedding vector to generate the final word vectors. It is proven to be beneficial to add classic word embeddings to enhance latent word-level semantics; for our task, we stacked FastText embeddings (Mikolov et al., 2018), pre-trained over web crawls, with the Flair contextual string embeddings, which are pre-trained with 1 billion word corpus.

2.1 Data

For the training of the before mentioned NER model it was required to obtain considerable amounts of annotated resumes. Two main sources of datasets with resumes were found, first one contained 220 annotated resumes from Indeed (employment website and search engine). These CVs are in a semi-structured form and have annotations for extracting 10 entities categories (Name, College Name, Degree, Graduation Year, Years of Experience, Companies worked at, Designation, Skills, Location Email Address), they were annotated collaboratively, by DataTurks community with the DataTurks online annotation tool (Trilldata-Technologies, 2018), the dataset is available in their repository (Narayanan and DataTurks, 2018). The second source was obtained from a Kaggle dataset (Palan, 2019). This dataset contains 1,200 non-annotated and unstructured resumes in a CSV file.

After analysing both sources of resumes it was discovered that the 220 annotated CVs from DataTurks were very poorly annotated, generating excessive noise on our initial NER training. For that reason, the annotations were removed and both sources were merged and re-annotated by us, using the Doccano open-source text annotation tool (Nakayama et al., 2018), initially using the same 10 entities used by DataTurks (Trilldata-Technologies, 2018).

Posterior to training several NER models, it was evident that the entity *Skills* was too broad and general and caused difficulties with the training of the NER models. The *Skills* entity was separated into *Job-Specific Skills*, *Soft Skills* and *Tech Tools*, the models were trained and compared against the 10 entities annotations, using the same resumes, improving the results.

A total of 507 resumes were annotated for each of these annotations types (10 and 12 entities) and a set-aside test set of 50 CVs (also with 10 and 12 entities annotations) were used for testing the best model. It is important to note that due to computational limitations, the resumes that had a count of tokens bigger than 2,000, were discarded before training. The corpus used to obtain the

	TRAIN	DEV	TEST
# of Documents	456	51	50
# of Tokens per Tag:			
No-tag	254,522	28,923	26,189
Name	1,479	124	170
Email_Address	2,010	227	222
Designation	6,866	670	613
Job_Specific_Skills	10,434	1,232	1,108
Tech_Tools	7,393	992	712
Companies_worked_at	5,412	524	494
Location	5,872	698	705
Years_of_Experience	6,203	696	641
College_Name	3,440	345	311
Degree	4,400	410	437
Graduation_Year	1,007	93	105
Soft_Skills	2,216	268	308

Table 1: Annotated Corpus description

highest score had the following composition of tokens and tags, see Table 1.

2.2 Cross-validated Data Quantity Analysis

A 5 fold cross-validation analysis was performed with different resumes amount, to understand the behaviour and performance of the models, as a whole (micro average) and per entity.

The second purpose was to evaluate if obtaining new annotated data will produce big enough improvements to the results. The amounts of CVs were **50**, **100**, **200**, **300**, **400** and **500**, they were taken from the Train and Development sets of annotated CVs presented in Table 1.

2.3 Best Model

The best model performance was achieved using the corpus mentioned in in Table 1 (using 12 annotated entities), the following hyperparameters and embeddings were provided to the Flair sequence tagger (NER) model:

- Embedding stack:
  - FastText word embeddings (Mikolov et al., 2018).
  - Flair Contextual String Embeddings ("news-forward") & ("news-backward"), (Akbik, Blythe, and Vollgraf, 2018).
- Initial learning rate: 0.5
- Dropout: p=0.12
- RNN layers: 2 BiLSTM layers
- Hidden size: 128

	10 Annotated Ent. F1(%)	12 Annotated Ent. F1(%)
Micro Avg.	72.1%	**78%**
Graduation Year	**92.2%**	87.2%
Name	**89.1%**	88.9%
Email	97.3%	**98.4%**
Location	87.9%	**89%**
Degree	81.5%	**85.5%**
Years Exp.	88%	**90.9%**
Designation	77.4%	**81%**
College Name	85.3%	**87.2%**
Company	67.6%	**76.1%**
Skills	56.1%	-
Job Specific Skills	-	**58.4%**
Soft Skills	-	**63.8%**
Tech Tools	-	**73.8%**

Table 2: 10 & 12 Annotated entities comparison

- Anneal factor: 0.5
- Patience: 3
- Use CRF: True

2.4 Evaluation Metrics

For the evaluation of the NER models, the metrics presented in the *"CoNLL-2003 shared task: language-independent Named Entity Recognition"* (Tjong Kim Sang and De Meulder, 2003) were used. These are exact match precision, recall and F1 score with with true-, false-positives (TP, FP) and false-negatives (FN):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

Mainly the F1 (3) metrics will be presented since this measure represents a harmonic mean between Precision and Recall.

3 Results

3.1 10 & 12 Entities Comparison

The results presented in Table 2 demonstrate the success of the experiment of dividing the *Skills* entity into 3 separate entities: *Job-Specific Skills, Soft Skills* & *Tech Tools*, and therefore, helped the model in overall. Mainly, due to the difficulty and implicit nature of the *Job-Specific Skills* is localized and contained allowing for the "easier" entities like *Soft Skills* and *Tech Tools*, to be trained more precisely, thus, improving the model as a whole.

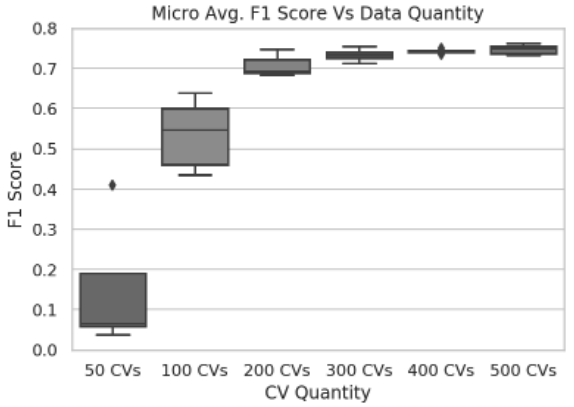

Figure 2: Micro average F1 vs. data quantity

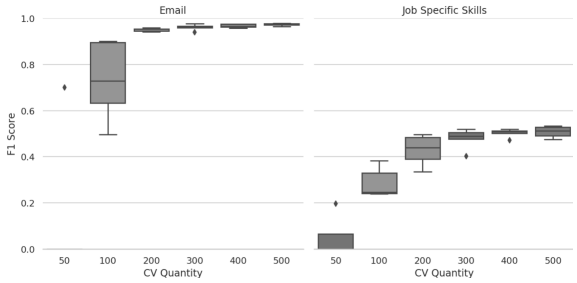

Figure 3: F1 Score vs. CV quantity for Email, Job-Specific Skills

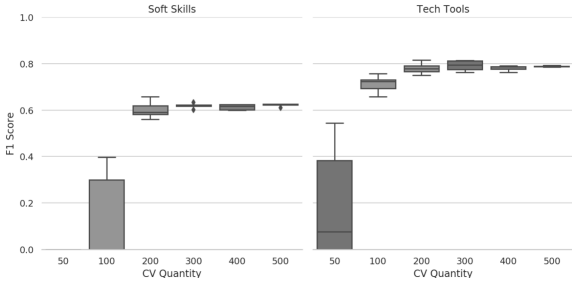

Figure 4: F1 Score vs. CV quantity for Soft Skills & Tech Tools entities

	F1 (%)	Recall (%)	Precision (%)
Micro Avg.	78.71 ±0.48	77.69	79.75
College Name	85.72 ±1.51	87.58	83.93
Company	74.21 ±2.38	78.8	70.13
Degree	85.65 ±1.53	81.86	89.8
Designation	84.08 ±1.74	91.21	77.99
Email	98.5 ±1.64	97.04	100
Graduation Year	92.73 ±2.98	92.73	92.73
Location	89.76 ±0.84	91.05	88.51
Name	89.43 ±2.28	90.16	88.71
Years Exp.	90.88 ±1.17	93.13	88.74
Job Spec. Skills	61.16 ±4.73	58.62	63.93
Soft Skills	65.27 ±2.03	60.56	70.78
Tech Tools	74.34 ±3.30	79.48	69.83

Table 3: Best model results, F1 Score, Precision and Recall

3.2 Cross-validated Data Quantity Results

Figures 2, 3 and 4 present a clear trend when more data is added for training, both for Micro average (Figure 2) and for individual entities (Figures 3 and 4). As data quantity increases the variability of the obtained F1 scores decreases. The performance of the model rapidly increases until it reaches a "ceiling", plateauing after 300 CVs. Furthermore, Figure 2 illustrates the inverse correlation between the variability of the models and training data quantity.

3.3 Best Model Results

The best model scores obtained are presented in Table 3.

4 Discussion

Figures 2, 3 and 4 demonstrate that the model reached a plateau in the scores, regarding the data quantity. Even if the labelled data is doubled, the increase in scores is going to be very limited. It might help reduce statistical variability across experiments but without much gain in the F1 score. Given this, we consider it is not worth investing the time and resources trying to increase the amount of well-labelled data.

It is important to note that the entity "nature" or complexity is not only present for the *Job-Specific Skills* entity only, Figure 3 and Table 3 illustrate how varied the scores among entities are, having different "ceilings" amongst them, there is a 37.3% difference between the highest and the lowest entities F1 scores (in best model scores). 25% (or 3 entities) present a standard deviation greater than 2.5%, *Job-Specific Skills* having the highest value (4.73%), 58.3% (or 7 entities) with a standard deviation between 2.5% and 1.5%. 25% or 3 entities present a standard deviation lower than 1.5%.

The uneven results obtained across the different entities indicate that a Neural Network NER model alone is not enough to accomplish this task with satisfactory results for production deployment. A big amount of pre- and post-processing will be required in order to identify missed entities and resolve disambiguation among the predictions. As it has been presented on previous studies by Zhang et al. (2009) and Nguyen, Pham, and Vu (2018). It can be observed that entities that are easily identifiable and are pre-

sented with the same patterns across different CV's, such as *Email Address*, *Graduation Year*, *Years of Experience* and *Location* get higher scores. Email addresses will always have a "@" sign in the middle and will finish with ".com" or ".edu" or similar. Graduation Year and Years of Experience will always be a set of start and finish date or a duration period (ex. 2004 - 2008, 4 Years, 9 Months). These easily identifiable and repeated patterns are understood by the Neural Network-based NER. Therefore, good results are obtained for these entities. On the other hand, entities which do not show any type of repeated patterns or are unique to a very small type of jobs and/or CV's, such as *Job Specific Skills*, *Soft Skills* and *Company* exhibit how this type of model fails to understand and generalize due to high amount of variability that these entities pose.

Table 3 presents how the three Skills entities present a high variation in their scores, this is also due to each of these Skills "Nature", *Job Spec Skills* being clearly the most difficult for this type of model. This is caused by the big amount of Jobs found in a diverse corpus of CV's, each different Job will have their own set of skills, making it extremely difficult for the model to understand. A solution to this would be to train different models for specific domains, using specific domain training corpus. Contrarily, results show how the entity *Tech Tools* presents a very different "Nature". Tech Tools tend to be repeated more often across CV's, even if the job category or position of the Resumes are very different. Common Tech Tools like Microsoft Word, Excel, PowerPoint, Microsoft Windows... will be found more frequently, helping the model obtain a higher percentage of total relevant results correctly classified (Recall) compared to the other skill entities.

33% of the entities obtain F1 scores below 75%, these particular entities are critical for accurate data extraction from resumes, they are *Company*, *Job-Specific Skills*, *Soft skills* and *Tech tools*. Without obtaining at least 90% F1 score for these entities it is our opinion that it can not be considered as successful results for a standalone solution. We consider 90% as a success criterion based on scores obtained by the best current commercial CV parser, Rapidparser (RapidParser, 2020). This CV parser among other commercial and open-source parsers were compared by Neumer (2018) in his Master Thesis. In this work, Rapidparser obtained F1 scores above 94% for start-date, end-date, work description and skills entities (Neumer, 2018).

NLP models have advanced greatly in recent years but still fall short when trying to implement them by themselves, especially for this task, given that the goal is to build a system that can process successfully **unstructured** and **semi-structured** CVs.

5 Conclusion

Our results demonstrate that the advancements in the NLP field, such as the "Contextual String Embeddings" (Akbik, Blythe, and Vollgraf, 2018) and state-of-the-art NER architectures have failed to obtain satisfactory outcomes to be considered useful, for HR and recruiting industry applications, on their own. The results obtained contain high variance among the different entities intended to be extracted. Obtaining good quality when entities have low complexities and are explicitly presented, such as *Email*, *Name* and *Location*, but for the more critical and complex entities such as *Job-Specific Skills* and *Soft skills* this type of model alone can not cope with the ambiguity and implicitness of this information in **unstructured** resumes.

In order to be able to meet our sponsor needs for this task, this model will be improved and enhanced with different approaches. More classical text segmentation approaches, as the ones discussed in the *"Applying named entity recognition and coreference resolution for segmenting English texts"* article (Fragkou, 2017) combined with other rule-based techniques will be tested, to resolve disambiguation of the more complex entities while at the same time trying to catch the missed ones by the model.

As a positive outcome from this project, a diverse data-set has been produced. It contains 550 CVs, 25% **semi-structured** and 75% **unstructured**, ranging from a wide variety of industries paired with well-performed annotations. This provides a unique and good quality corpus for solving this task. The corpus can be found in the following GitHub repository: https://github.com/dotin-inc/resume-dataset-NER-annotations.

References

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Akbik, A., D. Blythe, and R. Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Bizer, C., R. Heese, M. Mochol, R. Oldakowski, R. Tolksdorf, and R. Eckstein. 2005. The impact of semantic web technologies on job recruitment processes. In O. K. Ferstl, E. J. Sinz, S. Eckert, and T. Isselhorst, editors, *Wirtschaftsinformatik 2005*, pages 1367–1381, Heidelberg. Physica-Verlag HD.

Fragkou, P. 2017. Applying named entity recognition and co-reference resolution for segmenting english texts. *Progress in Artificial Intelligence*, 6, 05.

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nakayama, H., T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Narayanan, A. and DataTurks. 2018. traindata.json, Jun. https://github.com/DataTurks/Entity-Recognition-In-Resumes-SpaCy/blob/master/traindata.json.

Neumer, T. 2018. Efficient natural language processing for automated recruiting on the example of a software engineering talent-pool. Master's thesis, TECHNISCHE UNIVERSITAT MUNCHEN.

Nguyen, V. V., V. L. Pham, and N. S. Vu. 2018. Study of information extraction in resume. Technical report, VNU-UET Repository.

Palan, M. 2019. resume_dataset.csv. Kaggle, Mar. https://www.kaggle.com/maitrip/resumes.

RapidParser. 2020. Cv parsing with rapidparser: Lightning-fast! www.rapidparser.com.

Tjong Kim Sang, E. F. and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Trilldata-Technologies. 2018. Dataturks - best online annotation tool to build pos, ner, nlp datasets. https://github.com/DataTurks/Entity-Recognition-In-Resumes-SpaCy.

Zhang, C., M. Wu, C.-G. Li, B. Xiao, and Z. Lin. 2009. Resume parser: Semi-structured chinese document analysis. pages 12–16, 01.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 59-66 recibido 17-03-2020 revisado 28-05-2020 aceptado 28-05-2020

Minería de argumentación en el Referéndum del 1 de Octubre de 2017

Argument Mining in the October 1, 2017 Referendum

Marcos Esteve, Francisco Casacuberta, Paolo Rosso
Universitat Politècnica de València
maresca@inf.upv.es, {fcn, prosso}@prhlt.upv.es

Resumen: La minería de argumentación permite, mediante herramientas software, obtener cuáles son los argumentos que expresan los autores en un determinado texto. En este artículo se pretende realizar un análisis de la argumentación expresada por los usuarios en Twitter en relación al referendum del 1 de octubre de 2017. Se utilizará para ello, el dataset MultiStanceCat proporcionado en la tarea organizada en el IberEval 2018. Dado que las herramientas de minería de argumentación trabajan en su mayoría en inglés, será necesario construir un sistema de traducción neuronal con postedición que permita realizar una traducción de los tweets del español y el catalán al inglés. Los resultados al realizar la minería de argumentación sobre los *tweets* traducidos ha demostrado obtener un porcentaje muy reducido de argumentación en todas las comunidades.
Palabras clave: Minería de argumentación, traducción automática, 1Oct2017

Abstract: Argument mining allows, through software tools, to obtain which are the arguments expressed by the authors in a given text. This article aims to make an analysis of the arguments expressed by users on Twitter in relation to the referendum of October 1, 2017 using the MultiStanceCat dataset provided in the shared task organized at IberEval 2018. Since the tools of argumentation mining work mostly in English, it was necessary to build a neural translation system with post-editing that allows to translate of tweets from Spanish and Catalan to English. The results of argumentation mining on the translated tweets have shown to obtain a minimum percentage of argumentation in all the communities.
Keywords: Argument mining, machine translation, 1Oct2017

1 Introducción

Una de las principales problemáticas que existen en la actualidad en redes sociales, y que se acrecenta cuando los temas son controvertidos, como el referéndum sobre la independencia de Cataluña, el Brexit o las elecciones estadounidenses, es la alta polarización existente entre sus distintas comunidades (Lai, 2019; Lai et al., 2019). La polarización de las comunidades implica la existencia de una alta comunicación entre los usuarios de la misma comunidad y, además, los mensajes que intercambian con otras comunidades suelen ser mensajes tóxicos donde destaca el odio. En este tipo de comunidades existen algunos comportamientos propios de la alta polarización existente. Por una parte en la comunicación dentro de la comunidad se puede observar el fenómeno de aislamiento, donde el usuario solo percibe información perteneciente a su comunidad viendo de esta forma una sola realidad. Relacionado con el anterior, también se observa el fenómeno de cámara de eco, donde, debido al aislamiento del usuario se produce un incremento de sus creencias.

El objetivo principal de este artículo consiste en determinar si existe argumentación en los *tweets* de la tarea MultiStanceCat (Taulé et al., 2018) presentada en el IberEval del 2018. Este dataset consta de un conjunto de *tweets* agrupados según su posicionamiento respecto al referéndum del 1 de octubre de

ISSN 1135-5948. DOI 10.26342/2020-65-7

2018, a favor, en contra, o neutral. Para realizar el análisis, será necesario hacer uso de herramientas de minería de argumentación, que permitan analizar dichos *tweets* extrayendo los argumentos expresados por los usuarios de las distintas comunidades. Dado que no existen tecnologías desarrolladas de minería de argumentación para *tweets* en español o catalán, se ha optado por utilizar la herramienta desarrollada por Iryna Gurevyc en la Universidad Técnica de Darmstadt[1]. Esta herramienta únicamente trabaja en inglés o alemán, por lo que será necesario traducir los *tweets* a dicha lengua. Para ello se ha propuesto un sistema de traducción con postedición basado en redes neuronales, con el que facilitar al operario humano la tarea de traducción de dichos *tweets*.

En los siguientes apartados se detallará el estado del arte para las distintas tecnologías y una introducción a los datasets utilizados. Se comentarán los sistemas desarrollados, así como la experimentación realizada para la traducción neuronal. Por último, se detallarán los experimentos desarrollados para la minería de argumentos sobre los *tweets* traducidos.

2 Estado del arte

2.1 Minería de argumentación

La argumentación permite a un escritor convencer a una audiencia sobre unas ideas. En esta área, un argumento es una estructura que consta de varios componentes donde, tal y como comentan los autores en (Stab y Gurevych, 2014), se fundamenta por una afirmación que está apoyada o atacada por, al menos, una premisa. La afirmación tomará por tanto, la parte principal del argumento, y le seguirán las premisas que trataran de sustentar la afirmación, persuadiendo de esta forma al lector. Por ejemplo, tal como detallan los autores y se puede observar en la Figura 1, la afirmación **1** tomará la parte principal del argumento y le seguirán las premisas **2** y **3**, que tratarán de sustentar la afirmación. Por último, la premisa **4** sustentará a su vez a la premisa **3**.

[1]http://www.argumentsearch.com/

"**(1) Museums and art galleries provide a better understanding about arts than Internet.** (2) In most museums and art galleries, detailed descriptions in terms of the background, history and author are provided. (3) Seeing an artwork online is not the same as watching it with our own eyes, as (4) the picture online does not show the texture or three-dimensional structure of the art, which is important to study."

Figura 1: Ejemplo de estructura argumentativa

La minería de argumentación permite, de manera automática, extraer estructuras argumentativas de grandes volúmenes de textos. Para ello, tal como indican los autores en (Eger, Daxenberger, y Gurevych, 2017), las investigaciones se centran en varias subtareas:

- La separación de unidades argumentativas frente a no argumentativas (segmentación en componentes)
- La clasificación de componentes argumentativas en premisa o afirmación
- Búsqueda de relaciones entre las estructuras argumentativas
- Clasificación de las relaciones según apoyen o ataquen a la afirmación

Un aspecto interesante a destacar es la transición existente en el desarrollo de dichas tecnologías. Mientras que, históricamente, la búsqueda de estructuras argumentativas se basaba en el uso de gramáticas construidas manualmente para un dominio especifico (Palau y Moens, 2009), en la actualidad las tareas de minería de argumentación se están centrando en sistemas *end-to-end* basados en redes neuronales recurrentes (Eger, Daxenberger, y Gurevych, 2017).

2.2 Traducción automática

La traducción automática trata de solventar el problema de la traducción de una lengua origen a otra destino, utilizando para ello herramientas automáticas. Esta área ha sufrido un alto auge gracias al incremento de textos bilingües paralelos donde, para cada oración en un idioma origen, se dispone de su correspondiente traducción en una lengua destino.

Históricamente, la traducción automática estaba basada en el conocimiento lingüístico, hasta la irrupción con gran éxito, de la llamada traducción estadística (Brown et al., 1990; Koehn y Knight, 2000), donde cada oración origen podía ser traducción de una oración destino y, por tanto, el objetivo consistía en asignar una alta probabilidad en caso de que, efectivamente, fueran traducción una de otra y, en caso contrario, asignarle una baja probabilidad. En la ecuación 1 se detalla la formula básica de la traducción estadística, donde x es la oración a traducir e y es una posible traducción.

$$argmax_{I,s} \prod_{i=1}^{I} p(y_i \mid y_1^{i-1}, x) \qquad (1)$$

Actualmente, los esfuerzos en traducción se están centrando en la traducción neuronal, ya que ha demostrado obtener mejores resultados frente a la traducción estadística (Koehn, 2009). Esto ha sido gracias, en parte, a la inclusión de las representaciones distribucionales conocidas como *word embeddings* (Bengio et al., 2003; Mikolov et al., 2013). Este tipo de modelado permitió representar las palabras como vectores densos que capturaban las relaciones semánticas y sintácticas con otros términos. Gracias a esto último ha sido posible aplicar potentes técnicas de redes neuronales a una gran cantidad de tareas textuales, obteniendo resultados del actual estado del arte. En el caso de traducción automática, las dos arquitecturas que fijaron el actual estado del arte, son por una parte, la arquitectura encoder-decoder con modelo de atención (Luong, Pham, y Manning, 2015) basada en redes neuronales recurrentes y, por otra parte, y más recientemente el uso de modelos basados en Transformers (Vaswani et al., 2017); basados en redes *feed-forward* y modelos de auto-atención entre las palabras de entrada, las palabras de salida y las palabras de entrada y salida.

Por otra parte, dado que los resultados en traducción siguen sin permitir una traducción totalmente automática existen numerosos esfuerzos por conseguir sistemas que permitan a un operador experto traducir texto con el mínimo esfuerzo posible (Peris, Domingo, y Casacuberta, 2017). En esta área se han desarrollado numerosas técnicas de postedición con las que se pueden obtener mejores resultados con un esfuerzo humano inferior a si se tuviera que traducir manualmente.

3 Datasets

Dado que no existen herramientas para la extracción de argumentos que trabajen en español o catalán y, que la herramienta desarrollada por la Universidad Técnica de Darmstadt solo trabaja en inglés o alemán; se ha decidido integrar un sistema de traducción que permita traducir los *tweets* proporcionados en el dataset MultiStanceCat del español y el catalán al inglés. Para ello se requiere de corpus paralelos que permitan la obtención de traductores tanto para el catalán como para el español.

3.1 Dataset MultiStanceCat

El corpus MultiStanceCat (Taulé et al., 2018) consiste en una recopilación de tweets multimodal (texto e imagen), utilizando para ello los *hashtags* #1oct, #1O, #1oct2017 y #1octl6. Tras la recopilación de 220.148 *tweets* entre el español y el catalán, los autores anotaron manualmente un subconjunto de estos, dependiendo de su posicionamiento respecto al referéndum (a favor, en contra o neutral) obteniendo de esta forma un total de 5853 *tweets* en catalán y 5545 en español. Además, con el fin de garantizar la calidad de las anotaciones, los autores realizaron dos procesos de acuerdo inter-anotador.

Posicionamiento	**Catalán**	**Español**
A favor	5106	2099
En contra	149	2231
Neutral	598	1215
Total	5853	5545

Tabla 1: Distribución de los *tweets* según idioma y según se posicionan a favor, en contra o neutral

El número de datos recopilados está balanceado en cuanto a número de *tweets* entre ambas lenguas. Además, tal y como se observa en la tabla, mientras para el español la distribución de *tweets* en las distintas clases está balanceada, para el catalán se observa un sesgo donde la mayoría de los *tweets* recopilados se han etiquetado con un posicionamiento positivo.

Por otra parte, los autores realizaron un estudio con el fin de determinar cuanto polarizadas estaban las distintas comunidades respecto al referéndum. El estudio determi-

naba que las comunidades estaban desconectadas entre si, ya que únicamente el 13.64 % de los usuarios seguían a usuarios de otras comunidades y, el número de usuarios que solo seguían a usuarios con una polarización, positiva o negativa, era del 51.44 % y 25.03 % respectivamente. Además, únicamente el 9.89 % de los usuarios manifestaban una postura neutral, por lo que existía una alta polarización.

3.2 Datasets para la traducción automática

Para la construcción de los traductores ha sido necesario el uso de corpus paralelos de oraciones español-inglés y catalán-inglés. Se han utilizado el corpus Europarl para traducir del español al inglés y el Opensubtitles para traducir del catalán al inglés.

3.2.1 Europarl

El Europarl (Tiedemann, 2012) se trata de un corpus paralelo extraído de la web del Parlamento Europeo por Philipp Koehn. Este corpus incluye versiones en 21 lenguas europeas. Destacar que el corpus que nos interesa consta de un total de 1.9 millones de pares de frases español-inglés y alrededor de 50 millones de palabras, tanto para el español como para el inglés. A estos pares será necesario aplicarles un preproceso, así como una tokenización.

3.2.2 OpenSubtitles

El corpus OpenSubtitles (Lison y Tiedemann, 2016) se trata de un corpus paralelo extraído de subtítulos de películas. Este corpus ha sido extraído de la pagina web http://www.opensubtitles.org/. El corpus incluye versiones para 62 lenguas distintas. La versión catalán-inglés consta de un total de 428.000 pares de oraciones y alrededor de 3 millones de palabras. Estos pares catalán-inglés deberán ser preprocesados y tokenizados para poder ser utilizados por el traductor.

4 Traducción neuronal

4.1 Propuesta de solución

A la hora de construir los traductores se ha propuesto un sistema basado en redes neuronales, utilizando para ello el *toolkit* NMT-Keras (Peris y Casacuberta, 2018). Dado que los traductores que se van a entrenar están basados en corpus con un lenguaje más formal y el objetivo final es obtener traducciones de un corpus de *tweets*, los cuales poseen un lenguaje coloquial lleno de errores, abreviaturas etc. Se ha propuesto un sistema basado en postedición que permita a un humano experto traducir de una manera eficiente e ir mejorando las prestaciones de los traductores de forma iterativa (Domingo et al., 2019; Peris y Casacuberta, 2019). En la figura 2 se detalla la implementación seguida para dicho sistema.

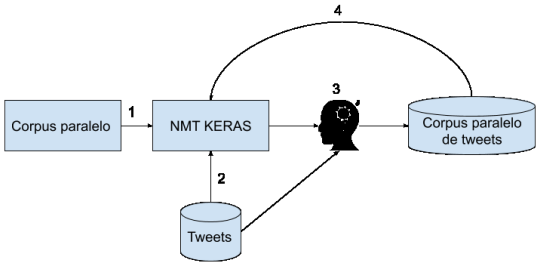

Figura 2: Descripción del sistema utilizado

Tal y como se detalla en la figura 2, el sistema consta de distintas fases:

1. Entrenamiento de los traductores neuronales con el *toolkit* NMT-Keras y utilizando los corpus paralelos expuestos en 3.2.1 y 3.2.2.
2. Extracción de un subconjunto de *tweets* del dataset expuesto en 3.1 y traducción por el sistema entrenado.
3. Revisión por parte del operario experto de las traducciones generadas por el sistema, rectificándolas y generando de esta forma un corpus paralelo de *tweets*.
4. Ajuste de los parámetros del traductor neuronal, utilizando el corpus paralelo de *tweets* generado por el operario.

4.2 Experimentacón

Para realizar la experimentación se ha cogido un subconjunto de los datasets paralelos y se ha explorado la variación del BLEU al cambiar los parámetros entre los modelos de **encoder-decoder con atención** y **transformer**. Concretamente, para las pruebas se ha utilizado 180.000 muestras para entrenamiento, 1.000 muestras para desarrollo y 2.500 muestras para test.

4.2.1 Resultados Europarl

En primer lugar, se ha explorado cual es la contribución del tamaño del *embedding* en el modelo de *encoder-decoder*; para ello se ha fijado el tamaño de las redes LSTM a 64 unidades y se han variado los tamaños de los

embeddings, tanto para el *encoder* como para el *decoder*.

Input	Target	BLEU
64	64	22.81
128	**128**	**30.99**
256	256	30.01

Tabla 2: Evolución del BLEU al modificar el tamaño de los embeddings en el corpus Europarl (3 epochs)

Como se puede observar en la tabla 2 los mejores resultados se obtienen cuando el tamaño de *embedding* es 128; por esta razón se fijará el tamaño del *embedding* a este tamaño a la hora de realizar la experimentación, y así observar la contribución del tamaño de la red LSTM a la variación en el BLEU.

Input LSTM	Target LSTM	BLEU
64	64	30.99
128	**128**	**32.12**
256	256	31.43

Tabla 3: Evolución del BLEU al modificar el tamaño de las capas LSTM en el corpus Europarl (3 epochs)

Tal y como se puede observar en la tabla 3, para nuestro caso el mejor modelo basado en *encoder-decoder* con modelo de atención se obtiene cuando el tamaño de los *embeddings* y de la red coinciden con 128 unidades.

Por otra parte, se ha explorado la arquitectura de traducción basada en Transformers, aunque las experimentaciones han revelado que los resultados obtenidos son significativamente inferiores a los obtenidos con la arquitectura encoder-decoder.

Por último, una vez determinada la mejor arquitectura (encoder-decoder) y parametrización (tamaño de *embeddings* 128 y unidades LSTM 128) se ha entrenado un traductor con un mayor volumen de datos, obteniendo un BLEU de 33.91. Este será, por tanto, el traductor que traducirá los *tweets* y se irá mejorando de forma iterativa utilizando el sistema de postedición propuesto.

4.2.2 Resultados OpenSubtitles

Se ha seguido un enfoque similar al expuesto anteriormente, fijando primeramente para la arquitectura encoder-decoder el tamaño de la red LSTM a 64 unidades y explorando la variación del BLEU al variar el tamaño del *embedding*.

Input	Target	BLEU
64	64	22.28
128	**128**	**22.59**
256	256	21.61

Tabla 4: Evolución del BLEU al modificar el tamaño de los embeddings en el corpus OpenSubtitles (3 epochs)

A la vista de los resultados expuestos en la tabla 4, el mejor BLEU se obtiene cuando el tamaño de embedding de entrada y salida coincide con 128 unidades. Se fijará, por tanto, este tamaño a los *embedding* a la hora de explorar la evolución del BLEU al aumentar el número de unidades en las redes LSTM.

Input LSTM	Target LSTM	BLEU
64	64	22.59
128	128	23.73
256	**256**	**23.91**
512	512	22.71

Tabla 5: Evolución del BLEU al modificar el tamaño de las redes LSTM en el corpus OpenSubtitles (3 epochs)

Como se puede observar en la tabla 5, los mejores resultados se obtienen cuando el tamaño de embedding es de 128 y el tamaño de las redes LSTM coincide con 256 unidades.

Además de la arquitectura encoder-decoder, se ha explorado la variación del BLEU al modificar el tamaño del modelo basado en la arquitectura Transformer, aunque los resultados no han conseguido mejorar al modelo basado en la arquitectura encoder-decoder.

De nuevo, una vez determinada la mejor arquitectura (encoder-decoder) y parametrización (tamaño de *embeddings* 128 y LSTM 256), se ha entrenado un traductor con un mayor volumen de datos obteniendo un BLEU de 27.02. Este será el nuevo traductor neuronal (catalán-inglés), el cual se irá mejorando de forma progresiva en la traducción de tweets empleando el sistema de postedición propuesto.

4.2.3 Traducción de los *tweets*

Una vez entrenados los traductores neuronales para la traducción del español y el catalán al inglés, se ha seguido la propuesta comentada en el apartado 4.1 para traducir los *tweets*. Se observó un incremento de las prestaciones de los traductores neuronales al incrementar

el número de muestras posteditadas en el proceso de adaptación; obteniendo traducciones más eficientes a si los tweets se tuvieran que traducir de forma manual.

5 *Argument mining*

Una vez traducido el corpus de *tweets* del dataset comentado en el apartado 3.1, se ha hecho uso de la herramienta, desarrollada por la Universidad Técnica de Darmstadt, ArgumenText (Stab et al., 2018) y disponible en el enlace http://www.argumentsearch.com/. Se trata de una herramienta que permite buscar argumentos escritos en inglés o alemán en grandes colecciones de documentos, haciendo uso de aprendizaje automático. Más concretamente hace uso de redes neuronales con modelos de atención para la clasificación de las oraciones según si poseen argumentos o no y, posteriormente redes recurrentes BILSTM, para clasificar el posicionamiento a favor o en contra del tema que está siendo analizado. Dado que la herramienta utilizada necesita un tema sobre el que realizar el análisis, se ha utilizado la palabra clave *catalonia* y se han agrupado los *tweets*, dependiendo de su lengua origen y del posicionamiento que tuvieran en el corpus MultiStanceCat.

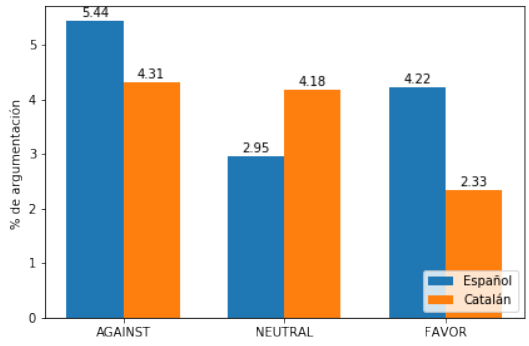

Figura 3: Porcentaje de argumentación según el idioma y el posicionamiento

En la figura 3 se puede observar como el porcentaje de argumentación es muy bajo, máximo de 5.44 % en el caso de la comunidad que está en contra y habla en español. Además, se puede observar que los hablantes de español en el caso de las comunidades a favor y en contra del referéndum, tratan de argumentar en mayor medida su postura, en comparación con los hablantes de catalán. En cambio, en el caso de la comunidad neutral se observa como los hablantes de catalán son los que tratan de argumentar en mayor medida sus ideas.

Algunos de los argumentos encontrados por la herramienta son:

- El de @cupnacional diu que tindrem més drets indepes. Potser si,BarcelonaWorld i escoles OPUS De moment recolzament partit corrupte. #1octL6[2] **(En contra)**
- #1O Catalá diu que especular amb la detenció de Puigdemont té poc sentit i insisteix que és decisió del jutge[3] **(Neutral)**
- #1octL6 que nivel de irresponsabilidad solo por interes electoral, de unos y otros. Politicos rompiendo la sociedad y un pais. Que pena...**(En contra)**
- Dice el Gobierno español q si renunciamos al referéndum #1O nos darán más dinero, autonomía y reformarán la Constit... **(A favor)**

6 *Conclusiones y trabajo futuro*

El objetivo principal de este artículo consistía en determinar si existía argumentación en los *tweets* del dataset MultiStanceCat acerca del Referéndum del 1 de Octubre de 2017 sobre la independencia de Cataluña. Para ello, puesto que no existen herramientas que realicen minería de argumentos sobre el catalán o el español, en este trabajo se ha propuesto un sistema de traducción neuronal construido con en el *toolkit* Nmt-Keras y basado en postedición para traducir el corpus de *tweets* de la tarea MultiStanceCat del catalán y el español al inglés. Esto ha permitido adaptar un dominio formal, como pueda ser un traductor entrenado sobre el corpus Europarl a la traducción de un dominio informal lleno de errores, como pueda ser los *tweets*. Esta fase ha sido especialmente ardua debido a la propia jerga informal utilizada en Twitter, donde por lo general existen una gran cantidad de errores en la escritura, abreviaciones, coletillas y características propias de la comunicación espontanea en medios sociales. Para solucionar este problema, ha sido necesario realizar un gran número de post-ediciones con el fin de

[2]El de @cupnacional dice que tendremos más derechos Indep. Quizás si, BarcelonaWorld y escuelas OPUS de momento apoyemos al partido corrupto. #1octL6

[3]#1O Catalá dice que especular con la detención de Puigdemont tiene poco sentido e insiste en que es decisión del juez

adaptar los traductores iniciales a la traducción de los *tweets*.

Por otra parte, una vez traducidos los *tweets* y, a la hora de analizar la argumentación existente en el corpus MultiStanceCat, se ha observado que los usuarios, en su mayoría no tratan de argumentar su postura acerca del referéndum celebrado el 1 de Octubre de 2018 y, además, existe un pequeño incremento en la argumentación de los usuarios castellanoparlantes.

Como trabajo futuro, puede ser interesante realizar un estudio del origen de los errores, ya que pueden existir errores en la traducción de los tuits y en la extracción de los argumentos. Por último, también puede ser interesante estudiar cuál es el porcentaje de argumentación en un corpus de tweets genérico. Esto permitiría realizar mejores comparaciones, permitiendo determinar si es habitual el uso de argumentación en Twitter. Además, se puede destacar la necesidad que existe de desarrollar herramientas que permitan realizar minería de argumentación para el español y las demás lenguas del estado español. Esto permitiría obtener mejores resultados al no tener que aplicar técnicas de traducción automática.

Agradecimientos

Dicho trabajo ha sido desarrollado en el marco de la asignatura Traducción Automática del Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital (MIARFID) de la Universitat Politècnica de València. El trabajo de los ultimos dos autores se ha desarrollado en el marco del proyecto *Misinformation and Miscommunication in social media: FAKE news and HATE speech* (MISMIS-FAKEnHATE) del Ministerio de Ciencia, Innovación y Universidades (PGC2018-096212-B-C31) y del proyecto PROMETEO/2019/121 (DeepPattern) de la Generalitat Valenciana.

Bibliografía

Bengio, Y., R. Ducharme, P. Vincent, y C. Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3:1137–1155.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, y P. S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Domingo, M., M. García-Martínez, Á. Peris, A. Helle, A. Estela, L. Bié, F. Casacuberta, y M. Herranz. 2019. Incremental adaptation of nmt for professional post-editors: A user study. *arXiv preprint arXiv:1906.08996*.

Eger, S., J. Daxenberger, y I. Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. En *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Koehn, P. 2009. *Statistical machine translation*. Cambridge University Press.

Koehn, P. y K. Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. En *AAAI/IAAI*, páginas 711–715.

Lai, M. 2019. *On language and structure in polarized communities*. Ph.D. tesis, Universitat Politècnica de València.

Lai, M., M. Tambuscio, V. Patti, G. Ruffo, y P. Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124 (online).

Lison, P. y J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Luong, M.-T., H. Pham, y C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

Palau, R. M. y M.-F. Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. En *Proceedings of the 12th international conference on artificial intelligence and law*, páginas 98–107.

Peris, Á. y F. Casacuberta. 2018. NMT-Keras: a very flexible toolkit with a focus on interactive NMT and online learning. *The Prague Bulletin of Mathematical Linguistics*, 111:113–124.

Peris, Á. y F. Casacuberta. 2019. Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language*, 58:98–126.

Peris, Á., M. Domingo, y F. Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Stab, C., J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, y I. Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.

Stab, C. y I. Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 46–56.

Taulé, M., F. M. R. Pardo, M. A. Martí, y P. Rosso. 2018. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. En *IberEval@ SEPLN*, páginas 149–166.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En *Advances in neural information processing systems*, páginas 5998–6008.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 67-74 recibido 03-04-2020 revisado 26-05-2020 aceptado 09-06-2020

Edición de un corpus digital de inventarios de bienes

Edition of a digital corpus of inventories of goods

Pilar Arrabal Rodríguez
Universidad de Granada
pilararrabal@ugr.es

Resumen: En este trabajo se pretende dar a conocer el proceso de elaboración de un corpus diacrónico digital a partir de la selección y edición digital de inventarios de bienes de los siglos XVIII y XIX de las provincias de Madrid y Almería. Los recuentos de bienes, de estructura repetitiva y abundantes en distintos puntos de la geografía hispánica, facilitan la comparación regional y cronológica de los documentos. Este corpus forma a su vez parte de *Oralia diacrónica del español* (ODE), un corpus que toma como inspiración el modelo tecnológico empleado por el proyecto europeo *P.S. Post Scriptum* para ofrecer en línea un corpus anotado a partir de la herramienta TEITOK (Janssen, 2016).
Palabras clave: Inventario de bienes, Lingüística histórica, Lingüística de corpus, español, XML, TEITOK

Abstract: The aim of this paper is to show the process of preparing a digital diachronic corpus based on the selection and digital edition of inventories of goods from the 18th and 19th centuries in the provinces of Madrid and Almeria. Inventories of goods which have a similar structure and are abundant in various areas of the Spanish geography, facilitate the comparison of the documents. This corpus is part of *Oralia diacrónica del español* (ODE), a corpus that takes as inspiration the technological model used by the European project *P. S. Post Scriptum* to offer an annotated corpus online using the TEITOK tool (Janssen, 2016).
Keywords: Inventories of goods, Diachronic linguistics, Corpus Linguistics, Spanish, XML, TEITOK

1 Introducción

En este trabajo se explica la metodología que se está siguiendo para elaborar un corpus diacrónico digital constituido por inventarios de bienes de los siglos XVIII y XIX de las provincias de Madrid y Almería.

Son cada vez más numerosos los corpus basados en documentación archivística que han sabido aprovechar las ventajas que ofrecen este tipo de documentos. En lengua española merece especial mención el corpus CHARTA[1], que integra numerosos subcorpus de muy diferentes tipologías, desde los albores del castellano hasta la época moderna. Entre otras ventajas, la documentación archivística posibilita con sobrada fiabilidad el estudio de la variación diatópica y diacrónica, lo que permite establecer diferencias geográficas claras en el uso de las palabras, así como delimitar la zona y extensión de los términos a lo largo del tiempo. De manera particular, los inventarios de bienes se benefician también de estas condiciones favorables para el estudio de fenómenos dialectales reflejados por los escribanos locales.

La gran abundancia de este tipo de documentos por todo el territorio hispánico permite comparar los resultados tanto diatópica como cronológicamente en la medida en que es posible ubicar con certeza el uso de los vocablos en un momento histórico y lugar geográfico concretos. Un claro ejemplo de ello es el *Corpus Léxico de Inventarios*[2], (Morala

[1] http://www.charta.es/

[2] http://web.frl.es/CORLEXIN.html

ISSN 1135-5948. DOI 10.26342/2020-65-8

Rodríguez, 2014). El *CorLexIn* reúne inventarios de bienes del siglo XVII de prácticamente casi todas las provincias de España y de otras partes de América. De esta manera, queda registrada de forma minuciosa una gran cantidad de léxico referido a objetos de muy diversa naturaleza que eran de extendido uso en la época y en todas las regiones hispánicas favoreciendo así interesantes investigaciones[3].

La abundante documentación conservada y la similar estructura que mantienen los textos, al igual que su contenido, e indistintamente de cuál sea su origen, favorece la investigación dialectal. Estas características hacen de los inventarios una interesante fuente documental que favorece la comparación.

Dejando a un lado los aspectos lingüísticos y pasando a las consideraciones tecnológicas, la edición digital de los documentos que conforman el corpus ha seguido de cerca el modelo tecnológico propuesto por el proyecto europeo *P.S. Post Scriptum*[4]. Este corpus reúne una amplia colección de cartas privadas tanto en español como portugués de autores de diferentes categorías sociales. Con ello se prioriza el estudio de variedades lingüísticas que no suelen aparecer reflejadas en otros corpus de carácter culto o literario, ya que la correspondencia privada se presta a evidenciar un discurso cercano a la oralidad por el tratamiento de asuntos cotidianos (Vaamonde, 2015).

Adicionalmente, el principal objetivo de este proyecto ha sido el de ofrecer un corpus enteramente editado, lematizado y anotado lingüísticamente gracias a la herramienta TEITOK (Janssen, 2016).

En la actualidad ya son veintidós los proyectos que utilizan TEITOK para alojar sus corpus. De entre ellos, *P. S. Post Scriptum* es el corpus que se ha tomado como referencia para la realización del nuevo corpus de inventarios por tratarse también de un corpus histórico de la Edad Moderna del español.

Además del empleo de TEITOK, el uso de estándares internacionales es imprescindible para no dar cabida a textos en los que la presentación gráfica sea heterogénea debido a diferentes criterios que además son desconocidos por el usuario que consulta el corpus. Para la elaboración de este corpus se ha tomado el estándar propuesto por el consorcio TEI para la transcripción de manuscritos.

El corpus que aquí se presenta justifica su elaboración en el principio de la fiabilidad de los documentos que lo componen y pretende ser una herramienta interdisciplinar que facilita investigaciones en diversas áreas como pueden ser la etnografía, historia, cultura y lingüística.

En definitiva, la idea que ha motivado la elaboración de un corpus como el que se propone parte de dos conceptos que se consideran clave: el aprovechamiento de la tipología inventario de bienes tal y como se consolida en *CorLexIn*, y el empleo de un modelo tecnológico para ofrecer un corpus con los últimos avances como lo es el propuesto por *P. S. Post Scriptum*.

2 Antecedentes

El origen del corpus cuya realización se expondrá en los apartados siguientes supone la continuación del proyecto *Corpus diacrónico del español del Reino de Granada. 1492-1833, CORDEREGRA* (Calderón-Campos y García-Godoy 2010), que recoge, entre otras tipologías de documentos, inventarios de bienes de las provincias de Granada, Málaga y Almería. No obstante, el número de textos era considerablemente menor en el caso de esta última. Por tanto, el objetivo ha sido completar el volumen del corpus con la inclusión de documentos de Almería para mostrar así un corpus completo, más rico y representativo de esta región.

Como mejoras frente al corpus ya existente, la ampliación prevé el uso de un avanzado modelo tecnológico que permitirá ofrecer un corpus enteramente digital gracias a la herramienta TEITOK. El resultado tras este cambio en la metodología es el corpus *Oralia diacrónica del español* (ODE) (Calderón Campos, 2019). El corpus que aquí se presenta es por tanto un subcorpus dentro de ODE.

La aplicación de una de las tecnologías más punteras en lingüística de corpus es el objetivo principal de los proyectos en los que este trabajo se enmarca: *HISPATESD* y *ALEA-XVIII*, que pretenden además alcanzar un elevado impacto social con la visibilidad y recuperación del patrimonio documental de Andalucía (por ahora de sus provincias

[3] Morala Rodríguez (2014) ha evidenciado en numerosos trabajos cómo corpus de estas características pueden hacer útiles aportaciones al estudio del léxico y a la lexicografía histórica.

[4] http://ps.clul.ul.pt/es/index.php?

orientales y más adelante de su totalidad). El segundo proyecto, como una de sus aplicaciones tecnológicas más relevantes, persigue además documentar formas léxicas identitarias de todas las provincias andaluzas a partir de un cartografiado diacrónico.

Se pretende contrastar la modalidad lingüística de la región con otras zonas y épocas, por lo que se ha proyectado la creación de un corpus de control con el que poder establecer diferencias o similitudes comparativas. Este corpus será representativo de la Comunidad de Madrid y servirá como referencia en tanto que reflejará una modalidad de habla lo suficientemente alejada de la registrada en Almería. De aquí en adelante y salvo que se especifique lo contrario, se hará referencia de manera conjunta a la metodología empleada tanto en el corpus de estudio almeriense como en el madrileño.

3 El corpus documental

La primera tarea para la confección del corpus se fundamentó en localizar, seleccionar y digitalizar los manuscritos. Para ello se consultaron los fondos notariales del Archivo Histórico de Protocolos de Almería y de Madrid. La labor de selección de los inventarios de bienes ocupó los primeros meses desde el inicio de elaboración del corpus y actualmente está concluida para ambas provincias.

En el caso de los documentos almerienses seleccionados, estos proceden de los once municipios ya contemplados en *CorLexIn* con el fin de mantener lo más fielmente posible el principio de la comparabilidad diatópica. A estos once puntos se le han añadido ocho nuevas localizaciones extendiendo así el espacio estudiado.

A su vez, el corpus de Madrid incluye inventarios de la capital y de otros trece municipios de la provincia. De esta manera se consigue un corpus dialectalmente más amplio y representativo de la provincia en su totalidad de lo que podemos encontrar en *CorLexIn*[5].

La selección de los inventarios de bienes no ha seguido ningún tipo de restricción a excepción de que estos se incluyan en el marco cronológico (XVIII-XIX) y la zona geográfica (Madrid y Almería) objetos de estudio. Ante la abundancia de este tipo de documentación en los archivos, se han preferido los manuscritos en mejor estado de conservación y solo se han rechazado aquellos inventarios que han planteado alguna duda sobre su certera datación o localización.

Los documentos seleccionados son, en su gran mayoría, cartas de dote y recuentos de bienes realizados con motivo de la muerte de algún familiar y tras los que se realiza su consecuente reparto entre los herederos; pero también se incluyen entre los documentos escogidos almonedas o embargos de bienes. En todos ellos el escribano recoge listas pormenorizadas de las pertenencias del benefactor, entre los que se encuentran utensilios domésticos variados, ropas, mobiliario o animales. En la Tabla 1 se muestra una tipología de los documentos pertenecientes a cada provincia incluidos en el corpus:

	Almería	Madrid
Cartas de dote	5	26
Particiones	44	18
Almonedas	2	1
Embargos	1	0
Total	52	45

Tabla 1: Relación de documentos por provincia

A la tarea de selección, le ha seguido la reproducción fotográfica del documento. Cada inventario se ha identificado con una cabecera en la que se incluyen datos que registran su datación y localización geográfica entre otras características descriptivas propias del documento.

Un propósito inicial radicaba en alcanzar un corpus de aproximadamente 100.000 palabras transcritas. Actualmente, con los cerca de cien documentos que componen el corpus, el material digitalizado sobrepasa este tamaño por lo que no se descarta seguir ampliando las dimensiones del corpus en fases posteriores.

4 Transcripción en XML-TEI

Los criterios de transcripción escogidos han sido el resultado de un largo proceso de reflexión y de toma de decisiones con el fin de lograr una edición digital que asegure el rigor filológico de las transcripciones. Tan solo se han normalizado la puntuación, el uso de mayúsculas y minúsculas y la separación de palabras, todo ello conforme a las normas ortográficas actuales. Las grafías se han

[5] La Comunidad de Madrid en *CorLexIn* únicamente se ve representada por tres de sus localidades, entre las que se incluye la capital.

respetado conforme al original con el fin de permitir estudios de carácter gráfico o fonético.

En cuanto a cuestiones de carácter técnico, la edición del total de documentos que componen el corpus se ha llevado a cabo en un lenguaje informático y versátil como es XML adaptado al estándar TEI (Text Encoding Initiative, 2007), un esquema de codificación muy especializado y de alcance internacional.

A partir de este tipo de lenguaje, el consorcio TEI propone un estándar que ofrece una metodología de codificación para la edición de textos en el ámbito de las Humanidades, y más particularmente para la marcación de manuscritos, que es la que ha servido de base para la edición de este corpus.

La transcripción en XML permite combinar aspectos de contenido y formato del documento gracias al etiquetado de diferente información, tal y como se muestra en la Figura 1:

```
        <pb n="3v" facs="20180601_132858.jpg"/>
        <p><lb/> <add place="margin">Imbentario</add> Estando en
mortuoria de Pedro Sevilla Cortinas, <lb/> vecino q<ex>u</ex>e fue
de las Cuebas, sita en la calle de los Na<lb break="no"/>bos de su
el d<ex>ic</ex>ho día veinte de sep<ex>tiembr</ex>e de mil <lb/> s
ochenta y seis, d<ex>o</ex>n Luis Flores Navarro, alg<ex>uaci</ex>
m<ex>ayo</ex>r <unclear>d este</unclear> juzgado, en uso de la com
q<ex>u</ex>e le está con<lb break="no"/>ferida, con asistencia de
es<ex>criba</ex>no y de Indalecio de <lb/> Meca y Josef Sevilla Ga
interesadas, <lb/> procedió a la práctica del imbentario mandado d
vienes q<ex>u</ex>e se manifiestan, a dejado el referido difun<lb
en la forma siguiente:
            <lb/> Prim<ex>eramen</ex>te una colcha de lana, azul,
encarnada.
            <lb/> It<ex>em</ex>: siete sávanas de lienzo almarieta
            <lb/> It<ex>em</ex>: tres pares de calzoncillos del mi
            <lb/> It<ex>em</ex>: cuatro camisas para hombre.
            <lb/> It<ex>em</ex>: otro par de calzoncillos.
            <lb/> It<ex>em</ex>: dos chamarretas blancas.
```

Figura 1: Transcripción en XML-TEI de un fragmento de inventario

En el fragmento transcrito se visualiza cómo es posible marcar características referidas a la estructura principal del documento con las etiquetas <pb/>, <p> o <lb/>; correspondientes a los inicios de folio, párrafo o línea respectivamente. También características físicas o visuales como la que refleja la etiqueta <add> para las adiciones en el margen fuera de la caja de escritura; o bien características conceptuales por medio de las etiquetas <ex> y <unclear>, para el desarrollo de abreviaturas o conjeturas editoriales respectivamente.

La edición en XML se ha llevado a cabo utilizando el procesador de textos *oXygen XML Editor*[6] que incorpora multitud de plantillas para trabajar con los distintos esquemas existentes de codificación, entre los que se encuentran las directrices propuestas por el consorcio TEI. El uso de *Oxygen* ha facilitado sobremanera el etiquetado y ha agilizado el proceso de transcripción, ya que permite validar automáticamente los documentos de acuerdo con el esquema elegido y asiste al usuario en la codificación.

La transcripción de los inventarios se ha centrado exclusivamente en aquellas partes consideradas de mayor interés léxico. Estas son las coincidentes con las listas o recuentos de bienes donde se enumeran pormenorizadamente los objetos de los que consta el inventario. Se han excluido, por tanto, aquellas partes referidas a las relaciones de parentesco de los herederos con el difunto o la relación de parientes implicados en la tasación y reparto de bienes. En definitiva, se han obviado fragmentos donde abundan las fórmulas fraseológicas o los tratamientos protocolarios propios de textos de carácter notarial como son los documentos que nos ocupan. Estas partes resultan claves para la correcta identificación del inventario y por ello se han digitalizado y conservado internamente, pero no se suman al conjunto del material transcrito. De manera excepcional, sí se han transcrito cuando brevemente permiten la identificación del documento y aparecen inmediatamente antes de la relación de bienes.

5 Procesamiento lingüístico del corpus

La edición digital del conjunto de datos no es la meta final. El tratamiento lingüístico del corpus supone uno de los grandes objetivos de carácter tecnológico con el que poder ofrecer un corpus anotado. El procesamiento del corpus consta principalmente de cuatro tareas que tienen que ver con la tokenización, normalización, lematización y etiquetado morfosintáctico. Todas estas tareas se realizan en línea directamente desde el sistema web TEITOK.

TEITOK está ideado para combinar la edición filológica con los avances de la lingüística computacional y abarca dos recursos en uno. En primer lugar, es una plataforma de consulta en la que cualquier usuario externo puede visualizar los documentos y realizar búsquedas en el corpus ajustadas a sus necesidades. En segundo lugar, es también donde se llevan a cabo por parte del equipo de trabajo las tareas de tratamiento lingüístico y

[6] https://www.oxygenxml.com/

edición del corpus que se desarrollan a continuación.

5.1 Tokenización

La tokenización supone uno de los primeros pasos para el tratamiento automático de los textos y alude al proceso de identificación de tokens. Esta tarea se lleva a cabo automáticamente en TEITOK una vez que los documentos han sido importados a la plataforma. El proceso de tokenización consiste en asignar a cada forma ortográfica, incluyendo también los signos de puntuación, un número único de identificación dentro de un elemento <tok> </tok> que engloba la palabra y cualquier otra información que se añada en relación con ella. A partir de ahora, dentro de cada token se almacenará toda la información relativa a los distintos niveles de edición de una sola palabra a la que posteriormente se le adjudicarán los atributos @form (correspondiente a la forma original), @fform (forma expandida) y @nform (forma normalizada), de manera que se respetan las múltiples formas gráficas que tiene una misma palabra. Véase el ejemplo de la Figura 2 tomado del token "vezino":

```
<tok id= "w-428" form= "vezno" fform= "vezino"
    nform= "vecino" >vezno</tok>
```

Figura 2: Ejemplo de token

La transformación de palabras en tokens es también imprescindible para los procesos de búsquedas del corpus con el sistema CQP (Corpus Query Processor). Gracias a la tokenización quedan vinculadas todas las formas de una misma palabra en sus distintos niveles de edición. En la búsqueda, el usuario puede decidir qué forma ortográfica desea buscar (paleográfica o normalizada) y el sistema recuperará todas las soluciones que coinciden en su normalización independientemente de las múltiples variedades gráficas que estén presentes en el texto original. Esto permite recuperar todas las variantes formales, incluyendo aquellas poco predecibles, a partir de una sola búsqueda.

5.2 Normalización

La normalización del corpus abarca exclusivamente el nivel ortográfico del texto. La gran variedad ortográfica que presentan los manuscritos obstaculiza sobremanera la posibilidad de ofrecer un corpus anotado morfosintácticamente. La normalización ortográfica supera esta barrera, pero, además, ello hace accesible el corpus al público interesado que no posea conocimientos lingüísticos específicos.

En esta segunda fase se añade la normalización de aquellas grafías que previamente se respetaron en la edición de acuerdo con el manuscrito original por poseer interés filológico y lingüístico. En TEITOK es posible realizar automáticamente la modernización ortográfica del corpus de acuerdo con parámetros previos importados a la plataforma y que sirven de entrenamiento. Tras la normalización automática se añade a cada token un atributo @nform tal y como se ha visto en el ejemplo de la Figura 2.

Tras la normalización automática se realiza una revisión manual conforme a las normas ortográficas actuales de aquellas palabras que no han sido procesadas correctamente. Esta revisión es de vital importancia, pues evita que errores en este nivel se sucedan en las siguientes fases de anotación.

5.3 Anotación lingüística

La anotación del corpus se lleva a cabo a partir de la forma normalizada de cada palabra. El proceso de anotación automática consiste en la adición de dos nuevos atributos @pos y @lemma a cada token ya formado. A estos atributos se incorpora la etiqueta morfosintáctica y el lema correspondiente.

El etiquetario utilizado para anotar el corpus se ha basado en el ya manejado por el proyecto *P. S. Post Scriptum*, aunque con ligeras modificaciones que simplifican en algunos casos las etiquetas utilizadas. Este, a su vez, sigue las directrices del estándar propuesto por el grupo EAGLES para la anotación de lexicones en lengua española.

La anotación morfosintáctica en TEITOK se lleva a cabo con el anotador automático NeoTag (Janssen, 2012), un analizador probabilístico del tipo HMM que adjudica a cada palabra la etiqueta correspondiente según la función gramatical que cumple en el texto. NeoTag funciona calculando la probabilidad de la etiqueta lingüística POS (part-of-speech) para cada forma ortográfica teniendo como base un corpus de entrenamiento. NeoTag busca en él la

frecuencia de aparición de cada palabra con su respectiva etiqueta POS y en base a ella adjudicará una etiqueta u otra dependiendo de la probabilidad.

En aquellas palabras nuevas o desconocidas para las que no hay datos en el corpus de entrenamiento, el analizador escogerá la etiqueta lingüística según otros factores, como son la terminación de la palabra o contextos similares de aparición. Sin embargo, hay otros casos también propensos a errar en la anotación. Se trata de aquellas palabras que pueden presentar ambigüedad léxica: palabras que, aun estando presentes en el corpus, requieran una anotación morfosintáctica distinta a la registrada por hallarse en un contexto diferente.

Hasta alcanzar las 100.000 palabras, el corpus de entrenamiento que sirvió como base para la anotación ha sido el del proyecto *P. S. Post Scriptum,* constituido por cartas de la Edad Moderna. No obstante, dadas las características individuales del corpus epistolar frente al de inventarios, existía un alto porcentaje de error.

Los siguientes ejemplos reales tomados del corpus servirán para ejemplificar cómo funciona el analizador. En la oración "un perol de cobre", que se encuentra fácilmente en el corpus, *cobre* es etiquetado reiteradamente con la etiqueta VMSP1S0 y con el lema *cobrar*. Se trataría por tanto de un verbo. Esto se debe a que en el corpus de entrenamiento (heredado de *P. S. Post Scriptum*) es alta la frecuencia de *cobre* como una forma verbal y no como el elemento químico al que se hace referencia en nuestro corpus.

De manera paralela a la anotación morfosintáctica se lleva a cabo la lematización, también de forma automática con Neo Tag. Para las palabras desconocidas, delimitar la terminación de las que no lo son es la estrategia usada para asignar el lema. Una vez que el analizador detecta varias posibilidades, elegirá la más frecuente. Este caso se aplica a la palabra "céntimos", anotada automáticamente como un verbo y con el lema **centimar*. Esta palabra no se encuentra en los parámetros y se anota conforme a su terminación correspondiente a la flexión verbal de otras palabras que en cambio sí localiza en los parámetros. En muchos casos como este una mala lematización está vinculada a una etiqueta POS errónea, pero en otros, como el caso de *veces*, (correctamente anotado como sustantivo, pero lematizado como *vec), se debe únicamente al procedimiento de asignación del lema. El algoritmo para ello genera en el acto un patrón de análisis morfológico con el que modificar la forma original para obtener el lema (Janssen, 2012). Este sistema fallará por ejemplo en aquellos casos de verbos con flexión irregular o de formas como la que se acaba de mencionar[7].

Tal y como se deduce a partir de estos ejemplos, los casos con mayor probabilidad de error radican en aquellas palabras que pueden presentar diferentes categorías gramaticales, casos de desambiguación, lo que exige una revisión manual de las mismas.

El porcentaje de acierto en la anotación dependerá de varios factores, entre los que se encuentran el tamaño y la calidad del corpus de entrenamiento en cuanto a lo que etiquetas correctamente asignadas se refiere. Pero sin duda, la precisión del analizador mejora cuando este utiliza el propio corpus como entrenamiento y no otros parámetros importados. Esto solo será posible cuando el tamaño del corpus lo permita (Janssen, 2016).

Al presente, debido al tamaño ya considerable del corpus, este se ha desligado recientemente de los parámetros de *P. S. Post Scriptum* y el conjunto de inventarios ya incluidos en TEITOK conforman el propio corpus de entrenamiento, en constante actualización según aumenta su tamaño. Con los parámetros de los inventarios como base se ha aumentado considerablemente la tasa de aciertos tras una anotación automática. La cantidad de nuevos datos anotados no es aún suficiente para poder realizar una comparación objetiva respecto a la efectividad de los parámetros previos. Se espera que se reduzcan las labores de revisión manual y que errores como los mencionados más arriba queden solventados, aunque no se han realizado todavía estadísticas concluyentes que valoren los resultados de la anotación en esta nueva fase[8].

Con todo, la tarea de la anotación lingüística demanda todavía una atención considerable y

[7] Aun así, los fallos en los lemas son muy limitados respecto a la anotación morfosintáctica. La lematización ha sido testada con un 95% de acierto en corpus españoles (Janssen, 2012: 3).

[8] Test de eficacia para la evaluación de NeoTag en la anotación morfosintáctica se han realizado sobre varios corpus del español actual con una precisión del 97% (*vid*. Janssen, 2012). En un corpus de los siglos XVIII y XIX como el que nos ocupa, y con un tamaño considerablemente menor, este porcentaje se verá reducido.

exige necesariamente una revisión manual, en algunos casos semiautomática[9]. Hasta el momento, esta revisión se ha llevado a cabo entre dos anotadores del equipo de trabajo y se está elaborando una guía, todavía de uso interno exclusivamente, en la que se contemplen los criterios estipulados con el fin de garantizar la consistencia de la anotación en la totalidad de documentos. De esta revisión depende la calidad del corpus de entrenamiento para la anotación de futuros documentos conforme va creciendo el corpus. Asimismo, también en esta fase radica la posibilidad de realizar búsquedas complejas mediante el sistema CQP que utiliza TEITOK, restringiendo las consultas y ofreciendo también información relativa al lema y a la etiqueta lingüística de palabras en posiciones específicas de la oración, lo que favorece una potente sintaxis de búsqueda.

6 Estado actual y trabajo posterior

Desde inicios de 2019 hasta ahora la tarea de recopilación de los inventarios de bienes está terminada tanto para el corpus de estudio como para el de control. Actualmente, el tamaño del que constan ambos corpus supera las 117.000 palabras conforme a la distribución que se muestra en la Tabla 2:

	Almería	Madrid
S. XVIII	28.826	30.325
S. XIX	25.785	33.017
Total	54.611	63.342

Tabla 2: Número de palabras en el corpus

Ya están disponibles para su consulta pública el volumen de textos anteriormente recogidos en la Tabla 2. Es posible acceder a ellos a través de la dirección web http://corpora.ugr.es/ode/, donde se aloja el corpus en su totalidad. Junto a cada transcripción, se puede consultar la información metatextual que acompaña a cada documento y que conforma la cabecera de cada uno de ellos. Además, está disponible la triple visualización de los documentos. Esto es: la reproducción fotográfica del documento o facsímil digital, la edición paleográfica y la edición normalizada de cada uno de los manuscritos.

Todos los documentos ya se encuentran normalizados y anotados morfosintácticamente. En este momento, las tareas de elaboración del corpus están centradas en la revisión del etiquetado lingüístico con el objetivo de constituir un corpus de entrenamiento consistente y de calidad con el menor número de errores posibles y con el que se vea reducido el actual porcentaje de error del anotador automático.

El corpus de inventarios ya se entrena con sus propios parámetros. De ahora en adelante, una vez desligado de los indicadores de *P.S. Post Scriptum*, se sigue probando la efectividad automática de la normalización y del analizador tras este cambio. TEITOK facilita la edición de la anotación lingüística a partir de búsquedas que permiten aplicar cambios en bloque. Actualmente se está trabajando en una revisión del etiquetado según este método, que, si bien no es del todo automático, agiliza sobremanera el proceso sin la necesidad de hacerlo documento por documento. Se espera que esta tarea se concluya al finalizar el presente año.

Producto de la metodología empleada, el resultado es ofrecer un corpus donde el usuario puede realizar búsquedas cruzadas de los documentos que ya están, además de normalizados, anotados lingüísticamente, y descargar libremente los archivos en formato XML o TXT con la transcripción de cada documento en la versión que desee.

Entre las tareas más inmediatas que se realizarán próximamente se contempla la necesidad de ampliar el tamaño del corpus con nuevo material transcrito una vez que se haya trabajado en la mejora del etiquetado morfosintáctico de los textos anotados. Está prevista también una proyección cartográfica digital que refleje todas las localidades de las que se tienen datos y que conforman el corpus.

Otras tareas que merecen especial atención están relacionadas con las búsquedas en el corpus. Entre ellas, mejorar la interfaz para hacerla intuitiva y accesible a cualquier usuario no especializado. Al mismo tiempo, se proporcionarán otras opciones avanzadas que posibiliten la recuperación de la información por medio de búsquedas complejas que permitan explotar al máximo un corpus del que será posible explorar muy diversos aspectos pertenecientes a diferentes ámbitos de investigación.

[9] Para una descripción de las distintas posibilidades que ofrece TEITOK a la hora de agilizar las correcciones en el etiquetado morfosintáctico consúltese Janssen, Ausensi y Fontana (2017).

7 Conclusiones

En este trabajo se ha dado a conocer el proceso de elaboración de un corpus de inventarios siguiendo los últimos avances en edición digital y lingüística computacional para corpus históricos.

TEITOK es una de las herramientas existentes que posibilita crear corpus a partir de la combinación de las tareas de transcripción y lingüística computacional en una plataforma con diversas funcionalidades donde se permite al mismo tiempo el mantenimiento en línea del corpus. Gracias a la tokenización se conserva toda la información relativa al texto y a su grafía al igual que se incluye información lingüística. Las ediciones que se ofrecen de un mismo texto a partir de una sola transcripción solo son posibles mediante el etiquetado previo de los distintos niveles de edición.

Partiendo de archivos codificados en el estándar TEI para lenguaje XML se ofrece un corpus enteramente digital, de libre acceso y ya disponible en línea.

En otro orden de cosas, el objetivo de un corpus de inventarios de las provincias de Almería y Madrid de las características mencionadas es el de poder realizar estudios lingüístico-estadísticos comparativos y establecer diferencias significativas a partir de los análisis contrastivos resultantes. Por ahora, el tamaño del corpus permite hacer una primera cala en esta dirección, pero el propósito es continuar ampliando el corpus y abarcar otras zonas que continúen enriqueciendo los análisis.

Los inventarios de bienes permiten realizar interesantes investigaciones en términos culturales y lingüísticos, ya que permiten reflejar la modalidad dialectal de la zona. Esto se traduce en un corpus creado para satisfacer no solo los intereses de los filólogos investigadores, sino también los de lingüistas, historiadores, etnógrafos o público interesado sin conocimientos lingüísticos específicos. Este carácter interdisciplinar es posible a la proyección de un corpus con diferentes ediciones de entre las que el usuario puede elegir la que desee y acceder a los textos a partir de un potente motor de búsquedas.

Agradecimientos

Este trabajo se inscribe en el marco de los proyectos de investigación *Hispanae Testium Depositiones HISPATESD*, de referencia FFI2017-83400-P (MINECO / AEI / FEDER, UE) y *Atlas Lingüístico y Etnográfico de Andalucía, s. XVIII. Patrimonio documental y humanidades digitales* (ALEA XVIII) (proyectos I+D+i Junta de Andalucía - FEDER, P18-FR-695).

Bibliografía

Calderón-Campos, M. 2019. La edición de corpus históricos en la plataforma TEITOK. El caso de Oralia diacrónica del español. *Chimera*, 6:21-36.

Calderón-Campos, M. y M. T. García-Godoy. 2010-2019. *Oralia diacrónica del español* (*ODE*). En línea: http://corpora.ugr.es/ode/

CLUL (ed.). 2014. *P. S. Post Scriptum. Archivo Digital de Escritura Cotidiana en Portugal y España en la Edad Moderna*. En línea: <http://ps.clul.ul.pt>

Consorcio TEI, (eds.). 2007. TEI P5: Directrices para la codificación e intercambio electrónico de texto (Versión 1.5.). En línea: <http://www.tei-c.org/Guidelines/P5/>

Janssen, M., J. Ausensi y J. M. Fontana. 2017. Improving POS tagging in Old Spanish using TEITOK. En *Proceedings of the NoDaLiDa 2017 workshop on Processing Historical Language*, páginas 2-6, Gotemburgo, Suecia.

Janssen, M. 2016. TEITOK: Text-Faithful Annotated Corpora. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, páginas 4037-4043, Portoroz, Eslovenia.

Janssen, M. 2012. NeoTag: a POS tagger for grammatical neologism detection. En *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Estambul.

Morala-Rodríguez, J. R. 2014. El CorLexIn, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro. *Scriptum Digital*, 3:5-28.

Vaamonde, G. 2015. P. S. Post Scriptum: dos corpus diacrónicos de escritura cotidiana. *Procesamiento del Lenguaje Natural*, 55:57-64.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 75-82 recibido 28-04-2020 revisado 04-06-2020 aceptado 04-06-2020

Relevant Content Selection through Positional Language Models: An Exploratory Analysis

Selección de Contenido Relevante mediante Modelos de Lenguaje Posicionales: Un Análisis Experimental

Marta Vicente, Elena Lloret
Department of Software and Computing Systems
University of Alicante, Spain
{mvicente,elloret}@dlsi.ua.es

Abstract: Extractive Summarisation, like other areas in Natural Language Processing, has succumbed to the general trend marked by the success of neural approaches. However, the required resources—computational, temporal, data—are not always available. We present an experimental study of a method based on statistical techniques that, exploiting the semantic information from the source and its structure, provides competitive results against the state of the art. We propose a Discourse-Informed approach for Cost-effective Extractive Summarisation (DICES). DICES is an unsupervised, lightweight and adaptable framework that requires neither training data nor high-performance computing resources to achieve promising results.
Keywords: Summarisation, Positional Language Models, Discourse Semantics

Resumen: Como muchas áreas en el ámbito del Procesamiento de Lenguaje Natural, la generación extractiva de resúmenes ha sucumbido a la tendencia general marcada por el éxito de los enfoques de aprendizaje profundo y redes neuronales. Sin embargo, los recursos que tales aproximaciones requieren – computacionales, temporales, datos – no siempre están disponibles. En este trabajo exploramos un método alternativo basado en técnicas estadísticas que, explotando la información semántica del documento original así como su estructura, proporciona resultados competitivos. Presentamos DICES, un método no supervisado, económico y adaptable que no necesita recursos potentes ni grandes cantidades de datos para lograr resultados prometedores respecto al estado de la cuestión.
Palabras clave: Resúmenes automáticos, Modelos de Lenguaje Posicionales, Semántica del Discurso

1 Introduction

The explosive growth of data that today's society witness, puts in the front line the research and development of suitable technologies that facilitate not only the access to such data deluge, but its comprehension. To this effect, summarisation techniques become a crucial resource, aiming at facilitating information access and understanding by condensing data with no loss of meaning (Nenkova and McKeown, 2011).

Deep Learning (DL) approaches have become increasingly popular in most of the Natural Language Processing (NLP) tasks, also in text summarisation, with competitive results and promising developments[1] for enabling transformation in industry and academia. However, today shortcomings of DL technologies raise concerns at different levels depicting a landscape of affected scenarios: small companies that cannot access huge amount of data, direct absence of such volume of data due to the specificity of the problem (e.g. some medical realms, organisational documentation, etc.) even the complex processing involved in DL, which may be costly not only in terms of computational resources, but in environmental impact (Strubell, Ganesh, and McCallum, 2019). Drawbacks of this technology, as it exists today, which highlight the need to explore alternative methodologies and motivate our current investigation.

Refocusing attention on statistical methods, in this paper we present a Discourse-

[1] `nlpprogress.com` is a repository to track the progress in NLP, for the most common tasks.

ISSN 1135-5948. DOI 10.26342/2020-65-9

Informed approach for Cost-effective Extractive summarisation (DICES), and examine its performance in the task of single document extractive summarisation (SDS). Since it is conceived as an unsupervised method that does not require human annotation or intervention neither copious amounts of data to provide good results, our approach represents a more digitally inclusive tool for public and private sector organisations that have tighter budgets for accessing Information and Communication Technologies. In this sense, organisations with small scale production of digital data could benefit from this lightweight mechanism to obtain the desired summaries from their different types of information. Our approach is feasible with less resources than would be necessary for deep or machine learning perspectives.

Central to our approach is the fundamental integration of a specific type of language model known as Positional Language Model (PLM), which has proven to be useful and cost-effective in different areas of NLP, such as information retrieval (Boudin, Nie, and Dawes, 2010) or language generation (Vicente, Barros, and Lloret, 2018). Taking into account the positional information of relevant elements within the document, we outline the significance of understanding and processing the text not as a simple set of words but as a succession of messages, whose full meaning must be accessed beyond the sentence level, at the discourse dimension. We shape our methodology to incorporate semantic information gathered from the document, instead of merely using words. Thereby, rather than simply relate to word counts, we aim at overcoming the limitations of the *bag of words* approaches by considering the original texts as structurally meaningful sources of information in a discourse-informed process, showing that such strategy have a positive impact on both the selection of content and the consequent generation of quality summaries.

In short, our contributions are: 1) we define a discourse-informed statistical model which incorporates semantics from the original text, revealing an improvement of the resulting summary; 2) we implement an unsupervised, lightweight and adaptable framework, simple yet effective and 3) we conduct a series of experiments over standard benchmarks and, with no heavy load of computational resources, empirically verify the effectiveness of our approach.

2 Related Work

As our approach is primarily designed for extractive summarisation, perspective that includes salient sentences from the original text without modification, we focus specifically on representative methods within this context, for brevity. We also include related work that underline the importance of discourse, as this is a core aspect for DICES.

Neural approaches have become mainstream research in recent years, where summarisation is tackled as a classification problem which establishes sentence appropriateness. Reinforcement learning (Wu and Hu, 2018; Chen and Bansal, 2018) or encoder-decoder architectures (Cheng and Lapata, 2016; Nallapati et al., 2016) are common, and research into new combinations grows constantly. As opposed to DICES, these approaches still need huge amounts of training data, which is not always available.

Whereas a large part of existing research relies on occurrence frequency, a few studies have focused on including discourse and semantics in their approaches. (Liu, Titov, and Lapata, 2019) proposed a structured attention architecture to induce trees while, in the case of (Liu and Chen, 2019) and (Hirao et al., 2013), they relied in Rhetorical Structural Theory (Mann and Thompson, 1987). The difference with our approach is that these systems include an expensive linguistic component that requires dependency parsing or rhetorical analysis to obtain the relation between the units of the document. DICES represents the semantics and structure of the components from a statistical perspective which imply shallow features and resources.

3 DICES Approach

In this Section we first explain the statistical foundations of DICES to later describe how a middle representation of the document is built upon the PLMs, serving as basis for the method to obtain the required summary.

The fundamental assumption here is that the better the understanding of the original text, the more informative the summary becomes. And only considering the text as a structured discourse, whose semantic elements coherently relates to each other, can that understanding be leveraged.

3.1 Positional Language Models

The basic idea behind this model is that for every position i within a document D, it is possible to calculate a score for each element w that belongs to the document's vocabulary. This value displays the relevance of w in a precise position, based on the element's distance to other occurrences of the same element throughout the document. The closer the elements appear to the position being evaluated, the higher the score obtained. This behavior allows the model to express the significance of the elements considering the whole text as their context, rather than being limited to the scope of a single sentence. In this manner, one PLM is computed for each and every position of the document, which can be formulated as follows:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \quad (1)$$

where *c(w,j)* indicates the presence of w in position j; $|D|$ refers to the length of the document D; V, the vocabulary and *f(i,j)* is the propagation function rating the distance between i and j.

3.2 PLM for summarisation

Having explained the model foundations, and thus the PLM module, a more detailed procedure needs to be designed to adapt the model to the task of summarisation. This procedure comprises three stages. First, we need the **definition of the vocabulary** as a parameter for the PLM module. From this stage, we obtain a representation of the text that involves both the vocabulary and the positions of its elements. Second, we **create a seed**, i.e., a set of words that can be relevant for the text and whose constitution depends on the corpus of origin itself. Finally, the processing of the PLM against the seed allows us to establish scores for the text elements, which will be transformed into a **ranking of sentences** from which the highest scored ones could be selected to produce the final summary up to a specific length.

3.3 The Vocabulary Definition

First, it is necessary to define which type of elements will constitute the vocabulary for the PLM. A straightforward approach would be to select the words as they appear in the text. However, this could lead to unnecessary repetitions. Alternatively, lemmas could be selected to obtain more effective results and, at a deeper and more comprehensive level, we could also consider more abstract forms, as identifiers for synonyms or deeper semantic or syntactic constructs.

In our current configuration, the vocabulary is composed of the synsets corresponding to nouns, verbs and adjectives, together with the named entities (NE) that appear along the text. This decision aligns with our semantic goal, that in this case was to capture the meanings, and the semantic information they convey. Freeling (Padró and Stanilovsky, 2012) is an open-source tool which provides several layers of linguistic analysis. We use it also to obtain synsets from WordNet (Kilgarriff and Fellbaum, 2000).

3.4 Seed Creation

A seed is then created, which must contain elements that allow us to dislodge the irrelevant parts of the discourse. The process of creation can begin with a sentence or with a set of words, that need to be analyzed with the same tools as the source text. A second vocabulary is then built from those.

Let V denote the source vocabulary, with elements $\{w_1,..,w_{|V|}\}$, and V_s the vocabulary extracted from the seed, a filter vector F is generated with as many positions as elements V has. If the element w_i from V belongs to V_s, then *F[i] = 1* ; *F[i] = 0* otherwise. Now it is possible to obtain a Score Vector (SC) with values for every position j:

$$SC[j] = \sum_{w \in V} P(w \mid j) \times F \quad (2)$$

The general vocabulary has been reduced to relevant vocabulary and, thanks to the PLMs, for each position of the document we dispose of a reference value. SC would become a detector of important areas, with maximum values when the accumulation of relevant elements given by PLMs is higher.

3.5 Ranking and Selection

From the SC, we are able to obtain the positions of interest within the document. From those, we retrieve as candidate sentences the sentences where these positions belong to, and subsequently obtain a value S_{score} for each of those sentences:

$$S_{\text{score}} = \sum_{i \in S} SC[i] \quad (3)$$

with S being the sentence to be scored, and i indicating the positions within the document for that sentence. Since it is possible that the position belongs to a very short sentence, and considering that the value in each position comes from its context, it would be pertinent to include the neighboring sentences. In this way, we introduce a parameter in the processing that allows us to select if we want to recover only the sentence that includes the position, or also its neighbors.

Typically, a parameter needs to be set in order to determine the length of the summary required. We sequentially select the highest scoring sentences, until the established length, to gather the best set of sentences for the resultant summary.

4 *Datasets, Experiments and Evaluation*

In order to evaluate DICES, several experiments were conducted over different corpus. Although the process is similar for all them, some aspects had to be adapted. Next, the datasets are introduced together with some details on the implementation. We finish with some notes on the evaluation metrics.

4.1 Datasets Description

The datasets selected to evaluate DICES were chosen for their renown, to enable a quality comparison with previous systems. In this section, we describe these datasets.

DUC 2002 Some of the most popular datasets used to address summarisation tasks come from the Document Understanding Conferences[2]. (DUC). Among others, DUC2002 includes a task aimed at SDS, whose goal was to built summaries from one English news article of 100 words at most.

CNN/DailyMail We selected a second corpus, the CNN/DailyMail (CNNDM) (Hermann et al., 2015), which is more recent than DUC2002 and is widely used to evaluate DL approaches from both extractive and abstractive perspectives. Apart from the abstractive gold standard in the form of highlights, some authors have created purely extractive model summaries. Particularly, the authors in (Cheng and Lapata, 2016) tagged with a label 1 the sentences from a document which should appear in a gold standard summary, and made the resulting dataset available[3].

One particularity of the corpus is that the documents are presented in an anonimized mode (we call it M for *mentions*), so that the entities appearing in the text are substituted by an identifier or mention. Then, along with the text, a list of the correspondent entities is provided. We processed the corpus to obtain a non-anonimized version (we call it E for *entities*), with the entities in their place. In this manner, our evaluation is conducted on both versions, M and E.

Although the version processed by the authors contains more than 280K documents, we selected a portion of the test documents to evaluate our system, since the strength of DICES does not rely on the amount of examples. Originally, the test set is comprised of 10,397 Daily Mail and 1,093 CNN documents. We took the 1,093 documents from CNN and randomly selected the same amount from Daily Mail, thus creating a smaller, but balanced dataset of 2,186 documents. For this collection, we estimate an average of 17 sentences per document labeled with 1 (following (Cheng and Lapata, 2016)'s version), and 4 highlights per document, also on average.

4.2 Implementation details

As introduced in Section 3, three stages were required to create the final summary: the vocabulary definition, the seed creation and selection of sentences after ranking them. The constitution of the seed is specific for each corpus. DUC2002 provides one headline per article. In this case, the headline, the first sentence or the combination of both would work as seed for the summaries. Regarding CNNDM, headlines are not provided but entities or mentions are. Therefore, the seed could consist of the first sentence of the document, the entities/mentions provided or their combination. For both corpora, we experimentally determined that the best strategy was to select as seed the combined option, for which results are reported.

4.3 Evaluation

We adopt ROUGE (Lin, 2004) for performing the evaluation, a recall-oriented measure which has become one of the most common metrics in summarisation. Summaries were evaluated taking into account ground truth

[2] duc.nist.gov

[3] github.com/cheng6076/NeuralSum

models and their relation with the system summaries in terms of n-gram overlap.

Regarding the gold-standard summaries, DUC2002 provided up to 4 model summaries for each single document and CNNDM documents are paired with a set of abstractive highlights, that serve as reference summary.

Besides, we created a pure extractive gold summary for CNNDM, taking into account sentences labeled as 1, provided by (Cheng and Lapata, 2016). The gold summary was exclusively used to examine DICES capablility of efficiently retrieving important information from a document. We discuss this issue in the next section.

5 Results and Discussion

To explore and evaluate the effectiveness of our approach, the following analyses were conducted: 1) we used the labeled CNNDM corpus to establish the system's ability to retrieve relevant sentences and, 2) we applied our system to the SDS task.

We next present our results and compare them with several state-of-the-art models. We found out that ROUGE unigram (R1) and bigram (R2) overlapping were usually reported by other systems, but occasionally longest subsequence overlap (RL) was also included. Furthermore, some works used F-score and some employed recall. Taking into account this diversity, and in order to get the clearest idea of our system's performance, we have included in the comparison all the significant approaches, reporting on the measure required in each comparison.

5.1 Relevant Sentence Retrieval

The average number of sentences with label 1 on the selected subset of the CNNDM corpus was 17. In compliance with this restriction, summaries limited to that length were obtained using DICES. These summaries were evaluated against the pure extractive gold summary to prove that our approach successfully detected the sentences where the relevant information in the article resided. R1, R2 and RL are presented in Table 1, both for the anonimized (M) and non-anonimized (E) versions of the corpus.

The results show our model's success in retrieving relevant information. The balance between recall and F-score also indicates that the recovered elements are significant, in the different n-grams modalities. However, the outcomes obtain a much higher value than those achieved when an abstract summary is used as reference. Therefore, in this case, we do not compare them with the other systems.

CNNDM	%	R1	R2	RL
M	R	83.18	74.54	81.03
	F	72.00	63.96	70.01
E	R	80.72	71.01	78.27
	F	71.17	61.93	68.86

Tabla 1: DICES evaluation against the pure extractive gold summaries from CNNDM—anonimized (M) and non-anonimized (E)

5.2 System comparison

In this section we compare our experimental results with state-of-the-art systems[4]. Additionally, some baselines are included to provide evidence of our achievements.

5.2.1 DUC2002 Evaluation

We evaluated DICES against several summarisation approaches and report the results in Table 2. The top part of the table exhibits systems that reported recall. The best performing system for the competition *BestDuc02* and the Lead baseline the organisers provided are included. The Lead baseline, taking as summary the first 100 words, relies on the assumption that in news genre, the relevant information is located firstly. Although this baseline was only surpassed by 3 systems, it is highly genre-dependent and does not consider semantic knowledge, contrary to our approach, which is easily adaptable to other genres and domains.

Additionally, taking into account that graph-based and statistical methods represent a common ground within extractive tasks, we included results from LexRank (Erkan and Radev, 2004), a popular graph-based technique that uses the PageRank algorithm, and implemented two baselines based on frequency counts that constitute popular references among statistical approaches, performing a *bag of words* strategy: *Tfidf* (term and inverse document frequency involved), and and *Tfisf*, which is a variation of the former, considering the inverse sentence frequency, instead *idf*.

Systems reporting F-scores are placed at the bottom of Table 2. To the best of our

[4]Results taken from the respective literature. Only for Pointer-Gen in CNNDM task, the code available in `github.com/abisee/pointer-generator##looking-for-pretrained-model` was run.

Duc2002	R (%)		
System	R1	R2	RL
Tfisf	37.03	13.39	30.11
Tfidf	38.43	14.39	31.40
Lead	41.13	21.07	37.53
BestDuc02	42.77	**21.76**	38.64
LexRank	43.20	17.94	**38.91**
DICES02	**44.72**	20.02	37.22
	F-score (%)		
System	R1	R2	RL
Pointer-Gen	37.22	15.78	33.90
ChenBansal	39.46	17.34	36.72
DICES02	**45.97**	**20.56**	**38.25**

Tabla 2: Recall and F-score results on the single-document task of DUC2002

knowledge there were no neural approaches reporting recall measure. We only found *ChenBansal* (Chen and Bansal, 2018) system that presents the F-score for results. They propose a reinforcement learning approach for abstractive summarisation and test it on the DUC2002 task, and also present the results for the pointer-generator system *Pointer-Gen* (See, Liu, and Manning, 2017). Although their system is abstractive and ours is not, the model summaries against which all three systems are compared are the same.

5.2.2 CNN/DailyMail Evaluation

As previously stated, our main objective when working with CNNDM was to evaluate DICES' ability to retrieve salient information (Section 5.1), given that the objective is eminently abstractive summarization. Nevertheless, DUC2002 results showed DICES remarkable performance regarding traditional yet competitive models, and moreover, its comparison against neural approaches revealed promising outcomes. We therefore decided to also conduct our summarisation experiments on this corpus, a task that has been usually approached from neural frameworks. Specifically, we carried out the experiments not over the whole dataset but over the subset of 2,186 documents that we had previously studied. These results may not be strictly comparable due to this size factor, but they certainly give us an idea of the potential of our approach. Table 3 summarises the results.

It appears in the Table a baseline that was created taking the first four sentences of each document, as this was the average length of the highlights for that subset. We built a baseline for each version of the subcorpus: the anonimized (M) and non-anonimized (E) one.

CNNDM - E	Non-anonimized		
System	R1	R2	RL
Tf-Isf	27.97	8.37	22.84
Tf-Idf	30.12	9.68	24.64
BL-4sent	35.56	13.89	31.53
Pointer-Gen	35.64	15.08	32.52
LiuLapata_19	**43.85**	**20.34**	**39.90**
DICES-E	34.46	13.09	28.19
CNNDM - M	Anonimized		
System	R1	R2	RL
Tf-Isf	31.00	10.05	25.59
Tf-Idf	32.77	11.25	27.09
BL-4sent	38.02	15.71	33.67
WuHu_18	**41.22**	**18.87**	**37.75**
DICES-M	36.78	14.62	30.27

Tabla 3: F-score (%) results for CNNDM-E and CNNDM-M, against highlights

We have also included the scores from the state-of-the-art systems: *LiuLapata* (Liu and Lapata, 2019) for E and *WuHu* (Wu and Hu, 2018) for M. Finally, we show the score for the provided model of *Pointer-Gen* on our corpora. Although we do not beat its performance, our score is considerably close.

Compared to state-of-the-art systems, the significant differences between the results may be caused 1) by the variation in the amount of data being processed in each case, 2) by the extractive condition of our approach against the other abstractive systems. We also conducted the test with the *Tf-Idf* and *Tf-Isf* set ups, also performing with them extractive summarization. As in the other tasks, DICES performs better than those approaches consistently.

We conducted one last experiment in order to better understand the performance and possibilities of DICES. As mentioned above, our subset of 2,186 documents was not strictly comparable with the state of the art. Nevertheless, we found an experiment in (Cheng and Lapata, 2016) which evaluates their extractive approach on 500 samples from CNNDM, with the highlights paired to the documents as gold standard. We randomly extracted the same number of articles from our data and performed a similar evaluation. The results (F-score) are reported in Table 4, and indicate a substantial improvement as least of 54 %. Nevertheless, we think it would be interesting to evaluate DICES exactly in the same set of documents.

CNNDM	F-score (%)		
500 docs	R1	R2	RL
ChengLapata	21.20	8.30	12.00
DICES-E	**34.14**	**12.83**	**28.05**
	(+61 %)	(+54 %)	(+133 %)

Tabla 4: F-score results computed on a random CNNDM subsample (improvement in brackets)

6 Discussion

The results obtained both with DUC2002 dataset, and in the evaluation of the system's capacity to recover relevant sentences, demonstrate the effectiveness of DICES in achieving the objectives established, and reinforces our effort on enhancing the semantic structure of the discourse as catalyst for progress in summarisation.

However, the results DICES obtained in some of the evaluation settings were lower than expected, for example, when compared to DL approaches. In this case, the different sizes of data evaluated could well explain the variation in the results or the disparity could be attributed to the fact that some of those systems are trained on different huge datasets and just tested in *smaller* datasets—the ones we use to evaluate our approach, as DUCs. In any case, to better understand the lower performance of DICES in the summarisation of CNNDM we plan to deeper analyse the impact of some factors. An aspect to be considered would be the system's evaluation over the whole CNNDM dataset, to check if in this case size is relevant. It is also worth noting that the corpus highlights are originally abstract summaries. This could imply a disadvantage with respect to our approach whose results may be better contextualised performing a manual evaluation of the resultant summaries. A thorough study on how the distinct constitution of the seed affects the outcomes could also give us more insight on what is causing the discrepancies and a wider scenario to enrich the research.

Moreover, we carried out an analysis on the resulting summaries that allowed us to identify some errors originated in the preprocessing stage. We detected, for example, how punctuation marks, mainly quotes, harmed the language analysers. Besides, the inadequate disambiguation of terms from which the synsets proceed also had an impact on the creation of the vocabulary, either coming from the body or from the seed, and affecting both the size of that vocabulary and its semantic composition, thus having a negative effect on the summaries generated.

Finally, it is worth mentioning recent work on summarisation which outlines the benefits of individually dealing with content selection and realisation (Gehrmann, Deng, and Rush, 2018; Cho et al., 2019). DICES is able to perform these tasks separately due to its modular architecture. The PLM stage presents a basic mechanism to detect salient content within a document by means of a condensed meaning representation. Although in this work we have exclusively tested its performance for extractive summarisation, DICES modules could also be part, for example, of an abstractive pipeline by adding a different realisation module. Additionally, DICES is able to work at multiple granularity levels by focusing on the sentences as a whole, specifically on their semantic constituents or even down to the token level. And this represents a crucial difference regarding common extractive approaches that usually relies in the sentence as their basic unit.

7 Conclusion and Future work

This paper explores a methodology for single document extractive summarisation that exploits positional and semantic information to improve the generation of summaries. A novel model based on statistical grounds is proposed. One of the motivations that led us to devise and test an approach like DICES was to provide a competitive alternative against the general trend that exploits neural networks, in contexts where, for one reason or another, computational and temporal resources or data are less accessible. In general, DICES achieves satisfactory results without the need for a large amount of data, training or computational load, in contrast to more sophisticated DL approaches.

The experiments show the capability of the framework both in detecting relevant areas of the document and in retrieving the appropriate sentences to construct relevant summaries. Its performance was successfully evaluated in creating single document summaries in the news domain for English.

The DICES methodology could easily be adapted to other languages, whenever a linguistic analyser is available. Moreover, due to its unsupervised nature and the flexibility DICES exhibits, it can readily be applied also to

different domains and summarisation modalities as multi-document summarisation, query and user focused summarisation or headline generation.

Acknowledgments

This research results from work partially funded by Generalitat Valenciana (*SIIA* PROMETEU/2018/089) and the Spanish Government—*ModeLang* (RTI2018-094653-B-C22) and *INTEGER* (RTI2018-094649-B-I00). It is also based upon work from COST Action *Multi3Generation* (CA18231).

References

Boudin, F., J. Y. Nie, and M. Dawes. 2010. Positional language models for clinical information retrieval. In *Proc. of EMNLP*, pages 108–115.

Chen, Y.-C. and M. Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proc. of the ACL, Vol. 1*, pages 675–686.

Cheng, J. and M. Lapata. 2016. Neural summarization by extracting sentences and words. In *Proc. of ACL*, pages 484–494.

Cho, S., L. Lebanoff, H. Foroosh, and F. Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proc. of the ACL, Vol. 1*, pages 1027–1038.

Erkan, G. and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Gehrmann, S., Y. Deng, and A. Rush. 2018. Bottom-up abstractive summarization. In *Proc. of EMNLP*, pages 4098–4109.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hirao, T., Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proc. of EMNLP*, pages 1515–1520.

Kilgarriff, A. and C. Fellbaum. 2000. WordNet: An Electronic Lexical Database. *Language*, 76(3):706.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, Y. and M. Lapata. 2019. Text summarization with pretrained encoders. In *Proc. of EMNLP-IJCNLP*, pages 3721–3731.

Liu, Y., I. Titov, and M. Lapata. 2019. Single document summarization as tree induction. In *Proc. of the NAACL, Vol. 1*, pages 1745–1755.

Liu, Z. and N. Chen. 2019. Exploiting Discourse-Level Segmentation for Extractive Summarization. In *Proc. of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121.

Mann, W. C. and S. A. Thompson. 1987. Rhetorical Structure Theory: Description and Construction of Text Structures. In *Natural Language Generation*. Springer Netherlands, pages 85–95.

Nallapati, R., B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proc. of SIGNLL*, pages 280–290.

Nenkova, A. and K. McKeown. 2011. Automatic Summarization. *Foundations and Trends® in Information Retrieval*, 5(2):103–233.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality FreeLing project developer. *Proc. of LREC*.

See, A., P. J. Liu, and C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proc. of ACL*, 1:1073–1083.

Strubell, E., A. Ganesh, and A. McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proc. of the ACL*, pages 3645–3650.

Vicente, M., C. Barros, and E. Lloret. 2018. Statistical language modelling for automatic story generation. *Journal of Intelligent & Fuzzy Systems*, 34(5):3069–3079.

Wu, Y. and B. Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI Conf. on Artificial Intelligence*, pages 5602–5609.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 83-90 recibido 02-04-2020 revisado 27-05-2020 aceptado 27-05-2020

Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry

Rantanplan, silabación y escansión rápidas de poesía española

Javier de la Rosa[1], **Álvaro Pérez**[1],
Laura Hernández[1], **Salvador Ros**[1], **Elena González-Blanco**[2]
[1]Digital Humanities Innovation Lab, UNED, Madrid, Spain
[2]School of Human Sciences and Technology, IE University, Madrid, Spain
{versae, alvaro.perez, laura.hernandez, sros}@scc.uned.es,
egonzalezblanco@faculty.ie.edu

Abstract: Automated analysis of Spanish poetry corpora lacks the richness of tools available for English. The existing options suffer from a number of issues: are limited to fixed-metre hendecasyllabic verses, are not publicly available, the syllabification procedure underneath is not thoroughly tested, and their speed is questionable. This paper introduces new methods to alleviate these concerns. For syllabification, we contribute with our own method and manually crafted corpus. For scansion, our approach is based on a heuristic for the application of rhetorical figures that alter metrical length. Experimental evaluation shows that both fixed-metre and mixed-metre poetry can be successfully analyzed, producing metrical patterns more accurately (increasing accuracy by 2% and 15%, respectively), and at a fraction of the time other methods need (running at least 100 times faster).
Keywords: stress, metrical patterns, scansion

Resumen: El análisis automatizado de corpus de poesía española carece de la riqueza de las herramientas disponibles para el inglés. Las opciones existentes adolecen de una serie de problemas: se limitan a versos endecasílabos de métrica fija, no están disponibles públicamente, el procedimiento de silabación no está probado a fondo, y su velocidad es mejorable. Este artículo presenta nuevos métodos para contrarrestar estos problemas. Para la silabación, contribuimos con nuestro propio método, así como un corpus elaborado manualmente. Para la escansión, nuestro enfoque se basa en una heurística para la aplicación de figuras retóricas que alteran la longitud métrica. La evaluación experimental demuestra que tanto la poesía de métrica fija como la de métrica mixta se analizan con éxito, obteniéndose patrones métricos con mayor precisión (mejoras de un 2% y un 15%, respectivamente), y en una fracción del tiempo que otros métodos necesitan (ejecutándose al menos 100 veces más rápido).
Palabras clave: acentuación, patrones métricos, escansión

1 Introduction

Although different in nature, syllabification and scansion are loosely coupled by the underlying functioning of the prosody of a language. Syllabification is the splitting of words into their constituent units, syllables. Unlike English, where there is a weak correspondence between sounds and letters, spoken syllables in Spanish are the basis of the orthographic units of its words. These building blocks shape the stress patterns and rhythm of a language, as well as the poetic metre of its poetry. Once a word is split into syllables, Spanish orthography establishes somewhat rigid rules to assign stress and classifies the words according to the position of the last stressed syllable. There is generally only one stressed syllable per word[1], with few exceptions (RAE, 2010). Depending on the position of the stressed syllable, there are three categories of words:

[1]In this work we use hyphens as the syllabic separator for representation purposes, marking in bold the stressed syllable (e.g., 'a-mo-**ro**-so').

ISSN 1135-5948. DOI 10.26342/2020-65-10

- *oxytone* words, when the stressed syllable is the last syllable of the word: 'tam-**bor**'.
- *paroxytone* words, when the stressed syllable is the one before the last syllable of the word: '**plan**-ta'.
- *proparoxytone* words, when the stressed syllable lies two syllables from the end of the word: '**plá**-ta-no'.

Some word functions, such as prepositions, conjunctions, articles, and even some pronouns and determiners, are usually left unstressed for metrical purposes despite having stress assigned by orthographic rules (Caparrós, 1993).

This division of words into stressed and unstressed syllables is the basis for scansion, the process of determining the metrical pattern of a verse. It depends entirely on a correct assignment of stress to the syllables of the words of a verse. However, scansion is also affected by some rhetorical devices that might alter the counting of stresses and even syllables present in a verse, thus differentiating between metrical length and syllabic length. We can talk about phonological groups for the syllables in a metre, which may be affected by metrical phenomena. Possibly, the two most common of these figures in Spanish are synalepha and syneresis. While both imply the union of separate phonological groups, the former acts between the last syllable of a word and the first of the next, for example in 'la amaba', 'la' and 'a' would be joined together. For the latter, the union occurs between the adjacent vowels within a word, 'son-re-ír' can be then split as 'son-reír' after syneresis. After applying these alterations, the number of syllables effectively shrinks for metrical purposes. Dieresis, on the other hand, is the metric phenomenon in which two vowels within the same syllable forming diphthongs are separated into different syllables, increasing the syllable count. Diereses tend to be graphically marked with a diacritical sign (¨), although its use in modern poetry is becoming less common.

Following the definition and representation of Spanish metre given by Navarro-Colorado (2017), we consider the metre of a Spanish verse as a sequence of stressed and unstressed syllables (Quilis, 1969; Navarro Tomás, 1991; Caparrós, 1993), where stressed syllables are marked with the plus symbol '+' and unstressed ones use the minus '−'. An extra unstressed symbol is added to the metrical representation of a verse when its last word is an oxytone, removed if a proparoxytone, or left unchanged if a paroxytone. Example 1 shows a verse of 8 syllables and the resulting metrical pattern after applying the synalepha (denoted by '‿') and considering the stress of the last word.

(1) *Cuando el alba me despierta*
*Cuan-do‿el-**al**-ba-me-des-**pier**-ta*
− − + − − − +− 8
(Miguel de Unamuno)

We aim at automating and enhancing the extraction of these metrical patterns of stressed and unstressed syllables. The application of automated techniques enables corpus linguistic approaches over poetry corpora that would otherwise need to be annotated manually. At the pedagogical level, it would also allow for the generation of didactic resources for the teaching of poetry and its scansion procedures, as our method produces not only a single output but all the information it relies upon to making its decisions.

2 Related Work

Manuals for metrical analysis of Spanish poetry exist at least since the 18th century, although the foundational work and subsequent refined guides for modern analysis would take another century to appear (Navarro Tomás, 1991; Caparrós, 1993). Despite such a long and rich tradition, not many computational tools have been created to assist scholars in the annotation and analysis of Spanish poetry. With ever increasing corpora sizes and the popularization of distant reading techniques (Moretti, 2013), the possibility of automating part of the analysis became very appealing. Although solutions exist, they are either incomplete, not suitable for Spanish, or not reproducible (Hartman, 2005; Agirrezabal et al., 2016). The first of such methods was introduced by Gervás (2000) as part of a larger system for the automatic generation of metrical poetry. In his work, Gervás uses Definite Clause Grammars in the logic programming language Prolog to model the division of a word into its constituents syllables, adding additional predicates to handle synalepha and syneresis. Once a metrical pattern is calculated, is

matched against a repository of metrical templates and the best match is returned. There are two issues with this approach: first, all words are assigned their correct lexical stress regardless of the part of speech. Secondly, all synalephas are applied indiscriminately since the actual metrical pattern calculated is never returned. How this repository is built is not entirely clear. He reported 88.73% per-line accuracy on a corpus of poems from the Spanish Golden Age period. We could not reproduce the figure since neither the code nor the dataset are publicly available at the moment.

A more modern approach was introduced by Navarro-Colorado (2017) as a rule-based system leveraging the morphological analyzer in Freeling (Padró and Stanilovsky, 2012) and focused on resolving metrical ambiguities. In his method, after splitting words into syllables and assigning stress according to their PoS, the possible synalephas and dieresis are marked and applied, ignoring synereses. This happens according to a knowledge base with probabilities for the different metrical patterns. The knowledge base is built offline from a large corpus[2] and fed to the system, thus assuming a relationship between high probabilities and metricality. The system was evaluated on more than 1000 lines extracted from a corpus of 100 manually annotated sonnets from the Spanish Golden Age period as well. A considerable increase in per-line accuracy is reported at 95%, contributing further with the first human annotated baseline reporting an inter-annotator agreement of 96%. However, and setting aside the dependence of the system on a correct PoS tagging, as much as 20% of the errors in the evaluation are due to problems related to the use of synalephas and diereses, mostly when combined. Moreover, there is no evidence nor evaluation of the ability of Navarro-Colorado's approach to properly assign metrical patterns for lines of verses other than hendecasyllables.

Shortly thereafter, Agirrezabal, Alegria, and Hulden (2017) experimented with the idea of applying neural networks to predict the metrical pattern of lines of verses. He designed a character-based bidirectional long short term (BiLSTM) neural network with conditional random fields and trained it on an similar corpus. A prior process of feature engineering added to the syllabification transformed each line of verse into a feature vector that kept the syllabic split, the surroundings of each syllable, PoS tags, and even stresses. He reported a per-line accuracy of 90.84%. Unfortunately, his approach is solely focused on predicting a metrical pattern from a very rich transformation of a verse, loosing in the process all information about phonological groups, individual syllabic stress, and synalephas, diereses, and synereses if any.

Although all approaches rely on a syllabification algorithm, Gervás' system was not made public, and there is no evaluation of Navarro-Colorado's although all his code was made publicly available to experiment with. To the best of our knowledge, the only published syllabification algorithm for Spanish was introduced by Agirrezabal et al. (2014) as an extension of his work in the English language. It used a finite state machine to split words into syllables and assign stress following the sonority hierarchy and maximum onset principle. However, we found some issues in the syllables of words present in the syllabification corpus employed for evaluation. Based on Ríos Mestre (1998), we disagree in the form some of the words are split into syllables, which could bias the accuracy of his method.

3 Fast Scansion

The aforementioned limitations guided the design of our own syllabification and scansion system, Rantanplan, which is comprised of four modules that work together to perform scansion of both fixed-metre as well as mixed-metre poetry: PoS tagger, syllabification, stress assignment, and metrical adjustment. The general algorithm, described in algorithm 1, operates at the line level with a sequence of words. First, for each word in a line of verse the PoS information is extracted and the word split into syllables (lines 2-3 in algorithm 1). Combining the PoS information and the syllabified word, the stress for each syllable is assigned according to the rules for oxytone, paroxytone, and proparoxytone words, plus a few exceptions detailed below (line 4). In the process, all possible synalephas and synereses are marked at the syllable level. With the enriched syllabic data, a new sequence of phonological groups is created by applying all possible synalephas and

[2]It is not exactly clear how large this corpus must be for his system to work.

synereses and keeping the information about the stress positions (line 6). This sequence of phonological groups is translated directly into a metrical pattern (line 7), since each phonological group represents a prosodic unit of pronunciation. The only consideration to factor in is the stress of the ending word, so an extra symbol could be added or subtracted accordingly when necessary. From here, two situations can occur:

1. The expected metrical length is not known, in which case the calculated pattern is returned (line 14).
2. The expected metrical length is known and its value greater than the length of the calculated pattern (lines 8-13). This means some of the applied synalephas and synereses must be undone until both lengths match. The metrical adjustment module will try every option iteratively giving priority based on a heuristic. For each attempt, a new metrical pattern and its corresponding length is calculated and checked against the expected metrical length. If no match is found, the last pattern calculated is returned.

Algorithm 1: Scansion procedure

Input: A sequence $\mathcal{W}$ of words $\langle w_1, w_2, \ldots, w_n \rangle$
Input: A value *length* for the metrical length expected (optional)
Output: A sequence $\langle s_1, s_2, \ldots, s_{\mathcal{L}} \rangle$ of booleans expressing the metrical pattern

1 **for** $w_i \in \mathcal{W}$ **do**
2 | $tag_i \leftarrow \text{pos}(w_i)$
3 | $syllables_i \leftarrow \text{syllabify}(w_i)$
4 | $stresses_i \leftarrow \text{stress}(syllables_i, tag_i)$
5 **end**
6 $groups \leftarrow \text{phonological}(syllables, stresses)$
7 $pattern \leftarrow \text{transform}(groups)$
8 **if** *length* **then**
9 | **while** $|pattern| < length$ **do**
10 | | $g \leftarrow \text{generate\_phonological}(\mathcal{W})$
11 | | $pattern \leftarrow \text{transform}(g)$
12 | **end**
13 **end**
14 **return** *pattern*

3.1 PoS tagger

We built Rantanplan on top of the industrial-strength NLP framework spaCy for speed (Honnibal and Montani, 2017). As mentioned previously, in Spanish some words are stressed depending on their function in the sentence, hence the need for a proper part of speech tagger. AnCora (Taulé, Martí, and Recasens, 2008), the gold standard corpus many modern statistical language models are trained on for PoS tagging of Spanish texts, splits most affixes thus causing some failures in the tags it assigns on prediction. To circumvent this limitation and to ensure clitics[3] were handled properly, we integrated Freeling's affixes rules via a custom built pipeline for spaCy. The resulting package, spacy-affixes[4], splits words with affixes before assigning PoS, and can be plugged in to a regular spaCy pipeline loading one of the statistical models for Spanish. In our approach, only suffixes on verbs are enabled in spacy-affixes to guarantee clitics are handled adequately by spaCy and PoS tags are assigned correctly.

3.2 Syllabification

Our method then follows a rule-based algorithm inspired by Ríos Mestre (1998), Caparrós (1993) and Navarro Tomás (1991) to split words into syllables. The procedure relies heavily on regular expressions to extract the letter groups that form the syllables. It is comprised of three steps.

1. Pre-syllabification rules are applied, which include the detection of consonant groups other than double 'l', as in 'aislar', and the handling of the prefixes 'sin-' and 'des-' when followed by consonants, as in 'deshielo'.
2. Regular letter clusters are identified and separated from the rest.
3. Post-syllabification exceptions for consonant clusters and diphthongs are applied.

Apart from the official rules for syllabification (RAE, 2010), there are cases with more

[3]Syntactically independent but phonologically dependent morphemes that appear together in a word, e.g., in 'cógemelo', both 'me' and 'lo' are pronouns written together with the verb 'coge'

[4]See `https://github.com/linhd-postdata/spacy-affixes/`

than one correct way to proceed. The first of these cases was the 'tl' group. Let's take the word 'atlántico' for example, its syllabification changes according to the territory[5]. We decided not to split the group 'tl' since most of the Spanish speakers around the world do not separate it. In the case of words of Nahuatl origin this separation should not be made either. Compounds words and words with an 'h' in between were also challenging. As an example of the former let's take the word 'reutilizar'. Although intuitively it may seem that the prefix 're-' should be separated from the rest of the word, the Fundéu[6] recommends not to do it this way, splitting instead as 'reu-ti-li-zar'. Similarly, the 'h' in a middle position does not split diphthongs, so 'desahijar' would be syllabified as 'de-sahi-jar', which might feel odd at a first pass but it actually agrees with the pronunciation of the word. Moreover, we also included possible diereses as part of our alternative syllabification exceptions. One such word is 'hiato'[7] which can be split either as 'hia-to' or 'hi-a-to'. As noted by Navarro-Colorado (2017), another common case is the word 'suave', which poets tend to apply dieresis to thus resulting in 'sua-ve' instead of the default split as 'su-a-ve'. Therefore, our method relies on a list of words with alternative syllabifications compiled from Ríos Mestre (1998). These alternatives are only taken into account by the metrical adjustment module.

3.3 Stress assignment and phonological groups

Once syllables and part of speech of a word are extracted, stress can be assigned. The assignment of stress follows very closely the rules defined in RAE (2010), adding exceptions for certain parts of speech, consonant groups, and words that are usually stressed but are not for metrical reasons. The sequence of phonological groups that will be used to extract the metrical pattern is calculated by applying all possible synereses and synalephas to the list of syllables of words per line, and propagating the stress information when needed. For example, the words 'me ama' are split into the syllables 'me-**a**-ma', and after applying synalepha the resulting phonological groups, '**me͡a**-ma', keep the stress in place. Word ends are also marked since they are needed to adjust the length of the metrical pattern according to the position of the stress of the last word. Phonological groups are then transformed into a metrical pattern representation and returned if the expected metrical length of the verse is not known beforehand.

3.4 Metrical adjustment

However, there are situations where the expected metrical length is known, such as processing a corpus of sonnets which tend to be hendecasyllables. In cases like this, verses with applied synalephas or synereses but a metrical length lower than the expected would trigger the adjustment module. In example 2, the expected metrical length is 11 but our system returns 9, thus triggering the metrical adjustment module.

(2) *loor a mi autor, y al que leyere*
***lo͡or**-a-mi͡au-**tor**-ya͡l-que-le-**ye**-re*
$+--+----+-$ $9 < 11$
(Juan de Timoneda)

This means that $11 - 9 = 2$ of the applied synalephas and synereses must be undone until both lengths match. The metrical adjustment module tries every possible metrical pattern combining synereses, synalephas, and alternative syllabifications. Priority is given to keep the synalephas since they are rarely broken, and syneresis are usually undone. The same happens for the alternative syllabifications, which deals with dieresis and adds more combinations to check for. A special case adding possibilities to the search space is the handling of synalephas between words with an initial 'h' and vowel ending words. Up to the 16th century, the initial 'h' in words was aspired instead of silent. This depends on the etymology of some words. For example, in the verse '*cubra de nieve la hermosa cumbre*' (see example 3) there should not be synalepha between 'la' and 'hermosa' since 'hermosa' evolved from the Latin 'fermosa' and as such a synalepha was not possible at all. To this day, this remains an option to the author, who can decide whether or not to apply a synalepha in such cases.

[5] See `https://www.fundeu.es/consulta/at-lan-ti-co-o-a-tlan-ti-co-12213/`

[6] The Fundéu is a foundation created from the Department of Urgent Spanish of the EFE Agency. See `https://twitter.com/Fundeu/status/1182226555457724416`

[7] Several examples can be found at `http://elies.rediris.es/elies4/Fon8.htm`

(3) *cubra de nieve la hermosa cumbre*
***cu**-bra-de-**nie**-ve-la-her-**mo**-sa-**cum**-bre*
+ − − + − − − − + − + − 11
(Garcilaso de la Vega)

For each attempt, a new metrical pattern and length is calculated and checked against the expected metrical length. If no match is found, the last pattern calculated is returned. For the verse in example 2, the generated possible metrical patterns are shown in example 4. Pattern 4 a, with no synereses and one synalepha between 'y' and 'al' would be generated first and returned afterwards. Since the metrical pattern has the correct length it is returned as such and the generation stops, saving the time it takes to generate any other possible pattern. This is also a limitation of our approach since more than one correct metrical pattern can be generated that matches the desired length.

(4) *loor a mi autor, y al que leyere*

(a) *lo-**or**-a-mi-au-**tor**-y‿al-que-le-**ye**-re*
− + − − − + − − − + − 11

(b) *lo-**or**-a-mi‿au-**tor**-y-al-que-le-**ye**-re*
− + − − + − − − − + − 11

(c) ***l‿oor**-a-mi-au-**tor**-y-al-que-le-**ye**-re*
+ − − − + − − − − + − 11

4 Evaluation

One notably difficult aspect of benchmarking automated analysis of Spanish poetry is the lack of a gold standard reference corpus. In recent years, the Corpus of Spanish Golden-Age Sonnets by Navarro-Colorado, Lafoz, and Sánchez (2016) is being used as the baseline. For syllabification, the best option is the limited corpus by Agirrezabal et al. (2014)[8]. Unfortunately, it contains some errors thus making it a not reliable source of truth. All evaluations were run on an computer with an Intel® Core™ i7-8550U CPU @ 1.80GHz and 16GiB of DDR4 RAM memory. When reporting figures, accuracy is expressed in percentages and time in seconds.

4.1 Syllabification

Since the only resource for syllabification in Spanish contains errors, we were forced to build our own corpus for the evaluation of the syllabification algorithm. We collected more than 100k words using a combination of online resources[9] into a corpus we named EDFU, and are releasing it under a Creative Commons license[10]. All entries are manually reviewed for correction and compliance with Ríos Mestre and Fundéu recommendations. Table 1 shows the accuracy of the methods by Agirrezabal (2014), Navarro-Colorado (2016), and ours when run against EDFU. Our method performs almost perfectly, more than one percentual point of gain over the others. No time comparison is made since all times are fairly similar.

Method	Accuracy
Navarro-Colorado	98.35
Agirrezabal	98.06
Rantanplan (ours)	**99.99**

Table 1: Scores on EDFU syllabification corpus. Best score in bold

4.2 Scansion

In his original work describing his scansion approach, Navarro-Colorado uses a set of 100 poems (1,400 verses) extracted from the Corpus of Spanish Golden-Age Sonnets (2016) for the evaluation of his system. While the list of the exact 100 poems selected was not made public, the author of the paper kindly provided us with a copy[11]. Since the corpus is comprised entirely of hendecasyllable sonnets, we used it for the evaluation of fixed-metre poetry and compared our results against Agirrezabal's neural network approach, and Navarro-Colorado's rule-based algorithm. Gervás' logic programming method was not available but he kindly agreed to run its system against the fixed-metre corpora and report back the raw outputs. Table 2 summarizes the results of per-line accuracy (evaluated as binary accuracy, entire metrical pattern matches divided by total number of lines of verse), showing that Rantaplan scores better than all other methods. The increase in accuracy is rather

[8] See `https://bitbucket.org/manexagirrezabal/syllabification_gold_standard/src/master/`

[9] Namely, `https://educalingo.com`, `https://dirae.es/`, and `https://www.fundeu.es/`

[10] See `https://github.com/linhd-postdata/edfu`

[11] We are making this corpus available in our corpus downloader tool, Averell: `https://github.com/linhd-postdata/averell/`

small but significant, while our method executes about 150 times faster than Navarro-Colorado's. We are marking the execution times for Gervás and Agirrezabal methods as not available.

Method	Accuracy	Time
Gervás[12]	70.88	N/A
Navarro-Colorado	94.45	2,356s
Agirrezabal	90.84	N/A
Rantanplan (ours)	**96.23**	**21s**

Table 2: Scores on Navarro-Colorado's fixed-metre 1,400 verses corpus. Best scores in bold

When compared against the entire manually checked part of Navarro-Colorado's corpus (2016), around 10,000 verses from 730 poems, the difference in per-line accuracy increases. Execution time is also added to the comparison. Table 3 shows per-line accuracy of our approach and Navarro-Colorado's system, showing a similar increment in accuracy for our method, around 2% better in metrical pattern calculation, and more than 300 times faster in terms of execution time.

Method	Accuracy	Time
Gervás[13]	67.56	N/A
Navarro-Colorado	90.89	16,787s
Rantanplan (ours)	**92.75**	**53s**

Table 3: Scores on Navarro-Colorado's fixed-metre 10,000 verses corpus. Best scores in bold

Lastly, for the evaluation of mixed-metre poetry we are using our own corpus of over 4,300 verses obtained from Carjaval's annotated anthology (2003). Unfortunately, due to copyright issues we are unable to release our annotated corpus for mixed-metre poetry. Table 4 shows results comparing performance of our method against Navarro-Colorado's (2017), showing that our approach is over 250 times faster and better suited to handle metrical stress that differ from a fixed value with a 15% increase in accuracy over Navarro-Colorado's system.

[12]Only 1,291 verses of the 1,400 verses corpus were evaluated by Gervás' method.

[13]Similarly, Gervás' method was only evaluated on 9,643 verses of the 10,000 verses corpus.

In addition to the improvements in accuracy for the different corpora, execution times seem to grow approximately linear with corpus size once we take into consideration that the loading time for the statistical model of Spanish in spaCy is 18 seconds, which gives execution times of 3 seconds for 1,400 verses, 9 seconds for 4,300 verses, and 35 seconds for 10,000 verses.

Method	Accuracy	Time
Navarro-Colorado	49.38	7,484s
Rantanplan (ours)	**65.02**	**27s**

Table 4: Scores on Carvajal's mixed-metre 4,300 verses corpus. Best scores in bold

5 Limitations

Despite the good scores obtained by our method, it is still based on a heuristic. Although thoroughly tested against different corpora, it could be the case that the heuristic we developed cannot account for changes in poetic production over time, thus rendering our system unable to accurately assess metrical patterns in modern expressions of poetry. We would need a more recent corpus to test this issue, but unfortunately most of these texts are still under copyright.

A second important limitation of our method is the use of a PoS tagger that relies on a statistical language model optimized for speed, which in some cases assigns incorrect part of speech tags, thus impacting the stress of the words and producing inaccurate metrical patterns.

6 Conclusions and Further Work

In this paper we have proposed methods for the automatic syllabification and scansion of Spanish poetry. Our syllabification method benefits from a carefully crafted new corpus, which we are releasing to the public. For scansion, two are the main contributions. First, we used a modern language model optimized for speed for the extraction of part of speech tags, improving execution times by a couple of orders of magnitude. Lastly, when extracting the actual metrical pattern we took the opposite approach to the previous state of the art and decided to apply all possible synalephas and synereses by default, only breaking them up when needed to match

metrical length. This decision paid off well in terms of accuracy since our method outperformed the rest in both fixed-metre and mixed-metre poetry.

We plan to continue improving Rantanplan and explore alternatives, specially using statistical language models to produce end-to-end metrical patterns further improving speed. Moreover, the output produced by our method will eventually be machine readable, interoperable, and ready to be ingested into a triple store compliant with the POSTDATA Project network of ontologies.

Acknowledgements

This research was supported by the project Poetry Standardization and Linked Open Data (POSTDATA) (ERC-2015-STG-679528) obtained by Elena González-Blanco and funded by an European Research Council (`https://erc.europa.eu`) Starting Grant under the Horizon2020 Program of the European Union.

References

Agirrezabal, M., I. Alegria, and M. Hulden. 2017. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Ranlp 2017*, pages 18–23.

Agirrezabal, M., A. Astigarraga, B. Arrieta, and M. Hulden. 2016. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4.

Agirrezabal, M., J. Heinz, M. Hulden, and B. Arrieta. 2014. Assigning stress to out-of-vocabulary words: three approaches. In *International Conference on Artificial Intelligence, Las Vegas, NV*, volume 27, pages 105–110.

Caparrós, J. D. 1993. *Métrica española*. Síntesis Madrid.

Fernández-Carvajal, F. 2003. Antología de textos.

Gervás, P. 2000. A logic programming application for the analysis of spanish verse. In *International Conference on Computational Logic*, pages 1330–1344. Springer.

Hartman, C. O. 2005. The scandroid 1.1. [Online; accessed 20-July-2020].

Honnibal, M. and I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 7.

Moretti, F. 2013. *Distant reading*. Verso Books.

Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Navarro-Colorado, B., M. R. Lafoz, and N. Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.

Navarro Tomás, T. 1991. Métrica española. *Reseña histórica y descriptiva*, 50.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.

Quilis, A. 1969. *Métrica española*. Alcalá Madrid.

RAE, R. A. E. 2010. *Ortografía de la lengua española*. Espasa.

Ríos Mestre, A. 1998. *La transcripcion fonetica automatica del diccionario electronico de formas simples flexivas del español: un estudio fonologico en el lexico*. Ph.D. thesis, Universitat Autònoma de Barcelona.

Taulé, M., M. A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *International Conference on Language Resources and Evaluation*.

A Appendix: Availability

A demo of Rantanplan can be found online at `http://postdata.uned.es/poetrylab/`. All source code is available under an Apache License 2.0 in a public code repository (`https://github.com/linhd-postdata/rantanplan/`) and as a Python package in PyPI (`https://pypi.org/project/rantanplan/`).

B Appendix: Reproducibility

To reproduce the results in this paper, please, refer to the next code repository: `https://github.com/linhd-postdata/rantanplan-evaluation`.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 93-96 recibido 30-04-2020 revisado 26-05-2020 aceptado 10-06-2020

COST Action "European network for Web-centred linguistic data science" (NexusLinguarum)

Acción COST "Red europea para la ciencia de datos lingüísticos centrada en la web" (NexusLinguarum)

Thierry Declerck[1], Jorge Gracia[2], John P. McCrae[3]
[1]DFKI GmbH, Multilinguality and Language Technology Lab
[2]Aragon Institute of Engineering Research, University of Zaragoza
[3]Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
declerck@dfki.de, jogracia@unizar.es, john.mccrae@insight-centre.org

Abstract: We present the current state of the large "European network for Web-centred linguistic data science". In its first phase, the network has put in place several working groups to deal with specific topics. The network also already implemented a first round of Short Term Scientific Missions (STSM)
Keywords: linguistic data science, multilingualism, linguistic linked data, language resources

Resumen: Presentamos el estado actual de la "Red Europea para la ciencia de datos lingüísticos centrada en la Web". En su primera fase, el projecto ha establecido varios grupos de trabajo para tratar temas específicos. La red también implementó una primera ronda de Misiones Científicas de Corto Plazo (la sigla STSM en Ingles, para Short Term Scientifc Mission).
Palabras clave: ciencia de datos lingüísticos, multilingüismo, datos lingüísticos anlazadoss, recursos lingüísticos

1 Introduction

We report on the current state of development of the "European network for Web-centred linguistic data science" (NexusLinguarum), which is a recently started COST Action.[1]

The main aim of NexusLinguarum is to promote synergies between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science. The network understands linguistic data science as a subfield of the expanding "data science", which focuses on the systematic analysis and study of the structure and properties of data at a large scale, along with methods and techniques to extract new knowledge and insights from it. Linguistic data science is a specific case, which is concerned with providing a formal basis to the analysis, representation, integration, and exploitation of language data (syntax, morphology, lexicon, etc.). NexusLinguarum brings thus the topic of linguistic data into this big data context.

In order to support the study of linguistic data science, the network is aiming at the construction at Web scale of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data. Such an ecosystem is needed to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation, and quality control of linguistic data. For this, NexusLinguarum is investigating the combination of linked data (LD) technologies, natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.). This combination seems to offer the potential to enable such an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages, by addressing in a principled way the semantic interoperability problem.

This combination builds on and further extends the recent development of the so-called

[1]See https://nexuslinguarum.eu/. The action started in October 2019, for a duration of 4 years. "COST" stands for "European Cooperation in Science and Technology". See https://www.cost.eu/.

ISSN 1135-5948. DOI 10.26342/2020-65-11

Linguistic Linked Open Data cloud (LLOD).[2] LLOD grounds on linked data to share and interlink linguistically relevant data sources. Specifically, linked data refers to the recommended best practices for exposing, sharing, and connecting structured data on the Web[3] and builds on Semantic Web recommendations and World Wide Web Consortium (W3C) standards such as Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL). In this, NexusLinguarum closely cooperates with on-going H2020 projects dealing with LLOD topics, Elexis[4] and Prêt-à-LLOD.[5]

2 Structure of the Network

The network counts on a very large number of participants, with institutions from 37 so-called COST Member Countries, 3 institutions from Near Neighbour Countries (Belarus, Georgia, Kosovo), 2 International Partner Countries (United States, Singapore) and 1 Specific Organisation (the Translation Centre for the Bodies of the European Union).

NexusLinguarum is co-ordinated by the University of Zaragoza, supported by the Polytechnic University of Madrid for administrative and financial matters. The chair of the Action is Jorge Gracia (University of Zaragoza, Spain), the Vice Chair is John McCrae (National University of Ireland, Galway) and the Scientific Communication Manage is Thierry Declerck (German Research Center for Artificial Intelligence). The Grant Holder Scientific Representative for the network is Elena Montiel (Polytechnic University of Madrid). The co-ordinator for the Short Term Scientific Missions is Penny Labropoulou (Athena Research Center) and Vojtech Svatek (University of Economics, Prague) is the ITC conference manager.

At its kick-off meeting, the network structured itself in 5 working groups, which are briefly described in the following sections.

2.1 Working Group 1 - Linked data-based language resources (leader: Milan Dojchinovski, Czech Technical University, Prague and Julia Bosque-Gil, University of Zaragoza)

WG 1 is laying the foundations and is developing best practices for the evolution, creation, improvement, diagnosis, repair, and enrichment of LLOD resources and value chains.

In the first phase of the action, this WG follows the task of identifying the current state of the LLOD cloud, with its problems and challenges, as well as to identify the state of the art and challenges in the current linguistic data models, especially with respect to under-resourced languages and domains that could and should be integrated in the LLOD.

One of the outcomes of WG1 will result in a matrix of datasets and languages, which can be shared with related work done in the Elexis project, which is aiming at developing a Matrix Dictionary in the field of eLexicography.

2.2 Working Group 2 - Linked data-aware NLP services (leader: Marieke van Erp, KNAW Humanities Cluster, Amsterdam)

This WG focuses on the application of linguistic data science methods including linked data to enrich NLP tasks in order to take advantage of the growing amount of linguistic (open) data available on the Web.

A main task in the first year of the action consists in identifying a list of existing standards, datasets annotations, NLP services and experts to connect to the COST action.

Additional tasks to be supported by this WG are for example the detection and description of LLOD data sets that can support a series of NLP tasks, like Natural Language Generation or Machine Translation.

2.3 Working Group 3 - Support for linguistic data science (leader: Dagmar Gromann, Centre for Translation Studies, Vienna, and Amaryllis Mavragani, University of Stirling, UK)

This working group focusses in the first year on fostering the study of linguistic data by following data analytic techniques at a large scale in combination with LLOD and linked

[2] http://linguistic-lod.org/llod-cloud

[3] https://www.w3.org/DesignIssues/LinkedData.html

[4] https://elex.is/

[5] https://www.pret-a-llod.eu/

data-aware NLP techniques. Big data and linguistic information. In this task, big data sources and state-of-the-art statistical analysis will be studied in combination to LLOD in order to better understand the language.

Visual analytics will be also considered for this task. This will have an impact on all sub-domains of linguistics, from typology to syntax to comparative linguistics.

WG3 is touching issues related to big data and linguistic information, as well as deep learning and neural approaches for linguistic data, in a multilingual setting.

2.4 Working Group 4 - Use cases and applications (Leader: Sara Carvahlo, University of Aveiro, Portugal, and Ilan Kernerman, K Dictionaries, Israel)

WG 4 is dealing with the identification of use cases that will demonstrate the possible deployment of LLOD related technologies. Use cases identified are in the legal domain, as well as in the digital humanities and social sciences, in order to show how linguistic data science can deeply influence studies in those fields.

At the first meeting of the WG, there was an agreement that setting up use cases that are broader in scope, particularly at this initial stage, could help get things in motion. Currently, five main topics are considered, although others will be added in later stages: legal domain, humanities and social sciences, linguistics, life science, and technology.

2.5 **Working Group 5 - Management and dissemination** (leader: Jorge Gracia, University of Zaragoza)

This WG deals with the day-to-day management and administrative coordination, as well as cross-WG and external communication and capacity building. Recently, a new task was added for scientific communication strategy. While this WG seems to be classical management one, it has to deal with the fact that the COST Actions are funding networking activities and not directly research work.

Therefore a main challenge consists in establishing cooperative relations with Research and Development programmes ant also to identify the most relevant Short Term Scientific Missions that can contribute to this kind of cooperation.

3 Possible Impacts of the Network

NexusLinguarun contributes to knowledge creation and transfer in several aspects.

Firstly, through training programs, scientific/industry events, datathons[6], which serve to promote and teach linguistic data science and its related technologies to people from both academia and industry.

Secondly, NexusLinguarum is contributing to a series of W3C standardisation activities in the field of Language Resources.

Thirdly, the project is aiming at disseminating its topics to a broader public, including popular science publications, blog posts on the Web, and contributions to crowdsourced resources, like Wikipedia.

Finally, the Action is committed to the design of a common curriculum for a Europe-wide master degree in linguistic data science, as a means to ensure knowledge transfer to a new generation of researchers and practitioners coming from different disciplines and with different backgrounds, in the topics related to linguistic data science.

4 Conclusion

We briefly presented the current state of development of the European network for Web-centred linguistic data science” (NexusLinguarum), which we think could greatly benefit to be discussed at the SEPLN conference, as the network could get impulses from the SEPLN communities, which are dealing with large varieties of languages.

Acknowledgments

Work presented here was supported in part by the COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”, the project Prêt-à-LLOD, under grant agreement no. 825182, and the ELEXIS project, under grant agreement no. 731015.

[6] NexusLinguarum will for example organise the next SD-LLOD Datathon and the next edition of the Language, Data and Knowledge (LDK 2021) conference. See http://2019.ldk-conf.org/sd-llod-2019/ for the past edition of those events.

References

Bellandi, A., E. Giovannetti, S. Piccini, and A. Weingart. 2017. Developing LexO: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France, September. Association for Computational Linguistics.

Berners-Lee, T. 2006. Design issues: Linked data. http://www.w3.org/DesignIssues/ LinkedData.html (August 24, 2020).

Chiarcos, C., J. McCrae, P. Cimiano, and C. Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources.* Springer, Heidelberg, Heidelberg, Germany.

Cimiano, P., C. Chiarcos, J. P. McCrae, and J. Gracia. 2020. *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

McCrae, J., G. A. de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, J., L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–719.

Bosque-Gil, J., D. Lonke, J. Gracia, and I. Kernerman. 2019. Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pages 726–746, Brno, Czech Republic, October. Lexical Computing CZ s.r.o.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 97-100 recibido 30-04-2020 revisado 26-05-2020 aceptado 11-06-2020

MINTZAI: Sistemas de Aprendizaje Profundo E2E para Traducción Automática del Habla

MINTZAI: End-to-end Deep Learning for Speech Translation

Thierry Etchegoyhen[1], Haritz Arzelus[1], Harritxu Gete[1], Aitor Alvarez[1], Inma Hernaez[2], Eva Navas[2], Ander González-Docasal[1], Jaime Osácar[1], Edson Benites[1], Igor Ellakuria[3], Eusebi Calonge[4], Maite Martin[4]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]HiTZ Center - Aholab, University of the Basque Country (UPV/EHU)
[3]ISEA, [4]Ametzagaiña
{tetchegoyhen, harzelus, hgete, aalvarez, agonzalezd, josacar, eebenites}@vicomtech.org
{eva.navas, inma.hernaez}@ehu.eus, iellakuria@isea.eus
ecalonge@ametza.com, maite@adur.com

Resumen: La traducción automática del habla consiste en traducir el habla de un idioma origen en texto o habla de un idioma destino. Sistemas de este tipo tienen múltiples aplicaciones y son de especial interés en comunidades multilingües como la Unión Europea. El enfoque estándar en el ámbito se basa en componentes principales distintos que encadenan el reconocimiento del habla, la traducción automática, y la síntesis del habla. Con los avances obtenidos mediante redes neuronales artificiales y aprendizaje profundo, la posibilidad de desarrollar sistemas de traducción del habla extremo a extremo (end-to-end), sin descomposición en etapas intermedias, está dando lugar a una fuerte actividad en investigación y desarrollo. En este artículo, se hace un repaso del estado del arte en este área y se presenta el proyecto MINTZAI, que se está realizando en el ámbito.
Palabras clave: Traducción del Habla, Traducción Automática, Reconocimiento del Habla, Síntesis del Habla, Aprendizaje Profundo

Abstract: Speech Translation consists in translating speech in one language into text or speech in a different language. These systems have numerous applications, particularly in multilingual communities such as the European Union. The standard approach in the field involves the chaining of separate components for speech recognition, machine translation and speech synthesis. With the advances made possible by artificial neural networks and Deep Learning, training end-to-end speech translation systems has given rise to intense research and development activities in recent times. In this paper, we review the state of the art and describe project MINTZAI, which is being carried out in this field.
Keywords: Speech Translation, Machine Translation, Speech Recognition, Text to Speech, Deep Learning

1 Participantes y entidades financiadoras

MINTZAI es un proyecto de investigación subvencionado por el Gobierno Vasco a través de la convocatoria de ayudas ELKARTEK 2019 de la Agencia Vasca de desarrollo empresarial Spri.[1] Su principal objetivo es la investigación y el desarrollo de sistemas de traducción automática neuronal del habla, con particular énfasis en la traducción entre euskera y castellano.

El proyecto tiene una duración total de 21 meses, con comienzo el 1 de abril de 2019 y finalización el 31 de diciembre de 2020.

MINTZAI se está llevando a cabo por el siguiente consorcio: Vicomtech[2], grupo Aholab de la UPV/EHU[3], ISEA[4] y Ametzagaina[5], siendo empresas adheridas Argia, EiTB y MondragonLingua. El proyecto tiene asignado el código KK-2019/00065 y el sitio web asociado es: http://mintzai.eus/

[1]http://www.spri.eus

[2]https://www.vicomtech.org
[3]https://aholab.ehu.eus/
[4]https://www.isea.eus/
[5]https://www.ametza.com

ISSN 1135-5948. DOI 10.26342/2020-65-12

2 Contexto y motivación

Los métodos de aprendizaje profundo (Deep Learning) se han impuesto como el nuevo paradigma en el campo de las tecnologías de la lengua, con mejoras significativas logradas por ejemplo en traducción automática, conversión de texto en habla, y reconocimiento automático del habla. Actualmente, tanto la investigación científica como la explotación comercial en estos ámbitos se basan mayoritariamente en variantes de redes neuronales artificiales profundas.

Una de las aportaciones importantes del enfoque basado en redes neuronales es la posibilidad de diseñar y entrenar sistemas extremo a extremo (E2E), i.e., sistemas que convierten información de entrada en información de salida mediante un sistema de aprendizaje neuronal único. Los sistemas de traducción automática neuronal, por ejemplo, modelan así de forma conjunta los procesos de alineamiento entre palabras y de traducción (Bahdanau, Cho, y Bengio, 2015); los sistemas E2E de reconocimiento del habla, a su vez, aprenden a asociar directamente señales de sonido con transcripciones para modelar el proceso completo de reconocimiento (Graves, Mohamed, y Hinton, 2013), y los sistemas E2E de conversión de texto a habla generan la señal partiendo de una representación fonética u ortográfica de la entrada (Wang et al., 2017).

Se puede contrastar este tipo de arquitectura con su alternativa estándar, donde distintos componentes se encadenan para convertir información de entrada en información de salida. En un sistema de traducción habla-habla estándar, por ejemplo, destacarían un módulo de reconocimiento del habla, cuya salida en forma de texto sería tratada por un módulo de traducción automática que produzca un texto en el idioma de destino, a partir del cual se pueda generar un contenido leído mediante una voz sintética.

Las diferencias de arquitectura se trasladan en diferencias importantes en cuanto a ventajas y deficiencias respectivas; a continuación se describen las principales diferencias, centradas en las ventajas obtenibles con sistemas E2E:

- *Reducción de errores*: Los sistemas en cadena tienen como desventaja la acumulación de errores generados por los distintos módulos de la cadena. Estos errores se propagan en la cadena de procesamiento global debido a la independencia de los módulos de tratamiento de los distintos aspectos (reconocimiento, traducción o síntesis). Los sistemas E2E, en comparación, no sufren de esta limitación y eliminan por lo tanto esta clase de errores.
- *Optimización de modelado*: Los sistemas E2E modelan la transformación de la información de entrada en información de salida mediante una red neuronal única. Esta característica permite modelar de forma conjunta los diferentes aspectos necesarios para encontrar una solución óptima; en contraste, los sistemas en cadena delegan la determinación de las representaciones a cada módulo de forma independiente, lo cual puede resultar en un modelado subóptimo del problema global a solucionar.
- *Desarrollo y despliegue*: Al ofrecer una arquitectura reducida a una única red neuronal, los sistemas E2E permiten una simplificación significativa de la preparación de sistemas para distintos idiomas y dominios. En comparación, los sistemas en cadena requieren entrenamientos separados de módulos complejos, y un encadenamiento adecuado de los distintos módulos que requiere un esfuerzo específico de adaptación y desarrollo. Los sistemas E2E facilita así el despliegue ágil que requieren los numerosos escenarios distintos que ocurren en la práctica.
- *Eficiencia*: Por la simplificación de arquitectura, los sistemas E2E pueden ofrecer una mayor eficiencia computacional, en términos de espacio y tiempos de procesamiento. Aunque sea posible en teoría desarrollar redes para sistemas E2E similares en complejidad a la suma de los componentes de sistemas encadenados, no suele ser el caso en la práctica y la eliminación de los componentes intermedios suele proveer una mejora a nivel de eficiencia.

Estas ventajas potenciales de los sistemas E2E han impulsado su presencia cada vez mayor en aplicaciones de tecnologías del lenguaje. Los principales sistemas comerciales en reconocimiento del habla, así como alguno de los sistemas de conversión de texto

en habla, son actualmente de tipo E2E, como pueden serlo también los sistemas de traducción automática adoptados en los ámbitos académicos y comerciales.

En el campo de la traducción del habla se han desarrollado muestras del potencial de la tecnología, como se describe en más detalle en la siguiente sección. En ciertas condiciones, los sistemas E2E iniciales pueden lograr resultados superiores a los obtenidos con sistemas estándar en cadena, y se considera una tecnología clave para el desarrollo de sistemas de traducción automática habla-habla y habla-texto, cuyas aplicaciones son múltiples y en alta demanda. En comunidades multilingües como la Comunidad Autónoma Vasca o la Unión Europea, por ejemplo, las necesidades de comunicación multilingüe se extienden a toda la red socioeconómica, con una presencia impactante de barreras lingüísticas que frenan la presencia cultural, la igualdad de idiomas y los desarrollos económicos.

Pese a resultados preliminares de cierto éxito con sistemas E2E de traducción del habla, los sistemas en cadena suelen obtener actualmente mejores resultados que los sistemas E2E en cuanto a calidad de traducción habla-texto en la mayoría de los casos. Por lo cual, existe un reto importante en investigación y desarrollo de arquitecturas E2E que permitan alcanzar o superar de forma consistente los sistemas en cadena clásicos. Por otro lado, mientras los sistemas de traducción habla-texto dominan el campo de la traducción del habla, la traducción habla-habla neuronal directa constituye un campo de investigación poco explorado, con retos propios importantes. Por último, para ciertos pares de idiomas, la escasez de recursos lingüísticos y componentes tecnológicos son obstáculos significativos para el desarrollo de tecnología de traducción del habla de calidad.

El proyecto MINTZAI se propone responder a estos retos, con la investigación y el desarrollo de métodos avanzados en traducción neuronal del habla, y su validación en casos de uso centrados en el par de idiomas euskera-castellano.

3 Estado del arte

Los sistemas estándares de traducción automática del habla se basan en tres componentes principales encadenados: un sistema de reconocimiento del habla, un sistema de traducción automática, y un sistema de síntesis del habla. Estos tres componentes principales se entrenan por separado, y el procesamiento opera en cascada: la salida del reconocedor de habla alimenta el sistema de traducción automática, el cual produce una traducción en forma de texto que sirve de entrada al sistema de síntesis del habla.

Para optimizar el funcionamiento de los sistemas encadenados, la comunicación entre componentes se suele adaptar a la tarea, en particular explotando hipótesis múltiples (Ney, 1999; Matusov, Kanthak, y Ney, 2005). Otros enfoques clásicos se han centrado en métodos estadísticos y en autómatas de estados finitos, integrando los modelos acústicos y de traducción en transductores estocásticos (Vidal, 1997; Casacuberta et al., 2004).

Como fue mencionado previamente, los sistemas encadenados sufren de la acumulación de errores y los avances a este nivel han consistido en mejorar los componentes individuales, como el reconocedor de habla, o en mejorar la conectividad entre componentes mediante rasgos específicos a la frontera entre componentes (Kumar et al., 2015).

Para resolver el problema de la propagación de errores y mejorar la calidad general de los sistemas, en trabajos recientes se ha explorado el enfoque extremo a extremo neuronal para la traducción habla-texto. Los primeros resultados obtenidos han sido prometedores con sistemas codificador-decodificador y mecanismos de atención, en particular en cuanto a la reducción de errores acumulados (Duong et al., 2016; Bérard et al., 2016; Weiss et al., 2017). En cuanto a la traducción habla-habla, los trabajos son escasos, con primeros trabajos exploratorios (Jia et al., 2019).

Pese a estos primeros resultados, donde sistemas E2E pueden superar a los sistemas encadenados en condiciones similares, en la mayoría de los casos los sistemas encadenados siguen obteniendo los mejores resultados, como muestran por ejemplo los resultados de las tareas compartidas internacionales en traducción del habla (Niehues et al., 2019). Una de las principales razones es la escasez de datos de entrenamiento paralelos para la tarea de traducción del habla, en comparación con los datos disponibles para entrenar componentes individuales. Otro factor relevante es la necesidad de mejorar las arquitecturas E2E para la traducción del habla en general.

4 MINTZAI

El proyecto MINTZAI se propone contribuir a los avances en la investigación de sistemas E2E tanto para la traducción automática habla-texto como para la traducción habla-habla. El proyecto pretende además avanzar en el estado del arte para la traducción del habla en el par de idiomas euskera-castellano, en las dos direcciones de traducción.

Tras su puesta en marcha en 2019, el proyecto ha logrado los primeros resultados resumidos a continuación:

- Creación de corpus paralelos habla-texto y habla-habla euskera-castellano y castellano-euskera a partir de los contenidos de las sesiones del Parlamento Vasco. Los corpus se compartirán con la comunidad científica bajo licencia Creative Commons BY-NC.
- Investigación y desarrollo de sistemas de traducción del habla encadenados en euskera y castellano, con componentes neuronales E2E para reconocimiento del habla, traducción automática y síntesis del habla.
- Investigación y desarrollo de sistemas E2E para traducción habla-texto y habla-habla en euskera y castellano.

Los primeros resultados del proyecto son satisfactorios, en particular con la creación de un corpus relativamente amplio para la traducción del habla en un par de idiomas con bajo soporte a nivel de recursos, y de primeros sistemas encadenados y E2E de traducción del habla para este par de idiomas.

Durante el 2020, el esfuerzo se centrará en extender y mejorar los primeros sistemas desarrollados, y en validar los resultados obtenidos.

Bibliografía

Bahdanau, D., K. Cho, y Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. En *Proc. of ICLR*.

Bérard, A., O. Pietquin, C. Servan, y L. Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. En *Proc. of NIPS*.

Casacuberta, F., H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcıa-Varea, D. Llorens, C. Martınez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, y C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Comput. Speech Lang.*, 18(1):25–47.

Duong, L., A. Anastasopoulos, D. Chiang, S. Bird, y T. Cohn. 2016. An attentional model for speech translation without transcription. En *Proc. of NAACL*, páginas 949–959.

Graves, A., A.-r. Mohamed, y G. Hinton. 2013. Speech recognition with deep recurrent neural networks. En *Proc. of ICASSP*, páginas 6645–6649.

Jia, Y., R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, y Y. Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv:1904.06037*.

Kumar, G., G. Blackwood, J. Trmal, D. Povey, y S. Khudanpur. 2015. A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation. En *Proc. of EMNLP*, páginas 1902–1907.

Matusov, E., S. Kanthak, y H. Ney. 2005. On the integration of speech recognition and statistical machine translation. En *Proc. of Eurospeech 2005*, páginas 467–474.

Ney, H. 1999. Speech translation: Coupling of recognition and translation. En *Proc. of ICASSP 1999*, páginas 517–520.

Niehues, J., R. Cattoni, S. Stuker, M. Negri, M. Turchi, E. Salesky, R. Sanabria, L. Barrault, L. Specia, y M. Federico. 2019. The IWSLT 2019 Evaluation Campaign. En *Proc. of IWSLT*.

Vidal, E. 1997. Finite-state speech-to-speech translation. En *Proc. of ICASSP*, páginas 111–114.

Wang, Y., R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, y R. A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv:1703.10135*.

Weiss, R. J., J. Chorowski, N. Jaitly, Y. Wu, y Z. Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv:1703.08581*.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 101-104 recibido 02-04-2020 revisado 24-05-2020 aceptado 24-05-2020

MISMIS: Misinformation and Miscommunication in social media: aggregating information and analysing language

MISMIS: Desinformación y agresividad en los medios de comunicación social: agregando información y analizando el lenguaje

Paolo Rosso[1], Francisco Casacuberta[1], Julio Gonzalo[2], Laura Plaza[2], J. Carrillo[2], E. Amigó[2], M. Felisa Verdejo[2], Mariona Taulé[3], Maria Salamó[3], M. Antònia Martí[3]
[1] PRHLT-Universidad Politécnica de Valencia
[2] NLP&IR- UNED
[3] CLiC-Universitat de Barcelona
{prosso,fcn}@dsic.upv.es; {julio, lplaza, jcalbornoz, enrique, felisa}@lsi.uned.es;
{mtaule, maria.salamo, amarti}@ub.edu

Abstract: The general objectives of the project are to address and monitor misinformation (biased and fake news) and miscommunication (aggressive language and hate speech) in social media, as well as to establish a high quality methodological standard for the whole research community (i) by developing rich annotated datasets, a data repository and online evaluation services; (ii) by proposing suitable evaluation metrics; and (iii) by organizing evaluation campaigns to foster research on the above issues.
Keywords: Fake news, biased information, hate speech, social media

Resumen: Los objetivos generales del proyecto son abordar y monitorizar la desinformación (noticias sesgadas y falsas) y la mala comunicación (lenguaje agresivo y mensajes de odio) en los medios de comunicación social, así como establecer un estándar metodológico de calidad para toda la comunidad investigadora mediante: i) el desarrollo de *datasets* anotados, un repositorio de datos y servicios de evaluación online; ii) la propuesta de métricas de evaluación adecuadas; y iii) la organización de campañas de evaluación para fomentar la investigación sobre las cuestiones mencionadas.
Palabras clave: Noticias falsas, información sesgada, mensajes de odio, medios de comunicación social

1 Introduction

Social media have become the default channel for people to access information and express ideas and opinions. The most notable and positive effect is the democratization of information and knowledge and the capacity of influence on the public opinion. Instead of a few media that control which information spreads and how (radio and TV channels, news media, etc.), any citizen can now share her views with the world and, potentially, influence the public state of opinion on any topic. But there are also undesired effects of this democratization of knowledge that is becoming increasingly important. One of them is that social networks foster information bubbles: every user may end up receiving only the information that matches her personal biases, beliefs, tastes and viewpoints. A perverse effect is that social networks are a breeding ground for the propagation of fake news: when a piece of news outrages us or matches our beliefs, we tend to share it without checking its veracity; and, on the other hand, content selection algorithms in social networks give credit to this type of popularity because of the click-based economy on which their business are based. Another harmful effect is that the relative anonymity of social networks facilitates the propagation of toxic, hate and exclusion messages. Therefore, social networks contribute, paradoxically, to misinformation and miscommunication. The main aim of this project is to alleviate these perverse effects via language analysis and the computational modelling of the propagation mechanisms for factual information, ideas and opinions in social networks.

ISSN 1135-5948. DOI 10.26342/2020-65-13

1.1 Participants

The MISMIS project (PGC2018-096212-B) is coordinated by Paolo Rosso (Universidad Politécnica de Valencia, UPV) and funded by the Spanish Ministry of Science, Innovation and Universities (R+D Knowledge Generation program). It started in 1st January 2019 and it will run until 31st December 2021. These are the groups involved in this project:

- The PRHLT[1] research center at UPV, which leads the MISMIS-FAKEnHATE (PGC2018-096212-B-C31) subproject. This is an experienced group in pattern recognition, machine learning, multimodality and NLP (e.g. author profiling, stance detection and automatic misogyny identification).
- The NLP&IR[2] research group at Universidad Nacional de Educación a Distancia (UNED), which leads the MISMIS-BIAS (PGC2018-096212-B-C32) subproject. They contribute to the project with their experience in the area of evaluation metrics and methodologies for Information Retrieval and NLP.
- The CLiC[3] research group at Universitat de Barcelona (UB), which leads the MISMIS-LANGUAGE (PGC2018-096212-B-C33) subproject. This group will contribute to this project with the development of theoretically founded linguistic models for the detection of misinformation and miscommunication, as well as by the creation of the datasets necessary for training and evaluating the systems that will be carried out.

2 Hypotheses and Objectives

The two main hypotheses that will guide our research are:

H1. Misinformation and miscommunication should be linguistically modelled, not only to improve the performance of algorithms but also to provide a better understanding of these phenomena.

H2. Linguistic resources, algorithms and metrics properly developed to address misinformation and miscommunication are essential for a consistent advancement related to the following general objectives of the project.

These objectives are the following:

- **O1 - Misinformation.** We will model and develop techniques to overcome the problem of misinformation from three angles: (i) by developing methods to extract unbiased knowledge from biased sources (what we call the wisdom of biased crowds); (ii) by developing techniques to identify fake news; and (iii) by designing summarization and visualization techniques for controversial topics, in which biases and manipulated information are exposed and the different points of view are identified and summarized contrastively.
- **O2 - Miscommunication.** We will develop techniques to detect and monitor aggressiveness and hate speech in social media, which contribute to create increasingly polarized communities and to demote argumentation in favour of pure confrontation. We will model hate speech from a linguistic point of view in different languages and we will also use multimodal signals (images and video) to complement textual evidence.
- **O3 - Methodological Tools.** One of the distinguishing aspects of our proposal is that we will make a substantial effort to improve research methodologies in the field, with the goal of establishing a high methodological standard for the research community. This is a horizontal objective for which we will: (i) develop rich annotated datasets for each of the relevant tasks; (ii) perform a formal study of suitable evaluation metrics that appropriately define each of the tasks and provide adequate analytical tools for evaluating system performance; (iii) develop a data repository and an online evaluation service in which metrics and datasets are centrally available for the research community, and (iv) organize evaluation campaigns using all the above tools to foster the sharing and reuse of knowledge in the research community.

3 Work progress

The problem of misinformation detection (O1) has been addressed from several perspectives: rumour detection (Ghanem et al., 2019a), fact checking (Ghanem et al., 2019b), fake news detection (Ghanem et al., 2019c) and credibility detection (Giachanou et al., 2019). False claims, and false information in general, are intentionally written to evoke emotions to the readers in an attempt to be believed and be disseminated in social media. To differentiate between credible and non-credible claims in the above latter works, we incorporated emotional

[1] https://www.prhlt.upv.es/wp/
[2] https://sites.google.com/view/nlp-uned/home
[3] http://clic.ub.edu/en

features into a Long Short Term Memory (LSTM) neural network. In order to address fake news from an author profiling perspective, currently we are organizing a shared task on profiling fake news spreaders on Twitter[4] (O3).

Another way to overcome the problem of misinformation is developing methods to extract unbiased knowledge from biased sources. Recent literature on recommender systems has focused on the inherent biases present in the data, which are amplified by recommendation algorithms. We are currently working on real-world datasets in order to characterize the existence of a geographic bias in the data. We are analysing the bias of current state-of-the-art recommender systems and try to understand the visibility that is given to items. Next, we will address the mitigation of the bias in the recommendation algorithms. We are also analysing fairness in session-based recommender systems (Ariza et al., 2020). We study how the effectiveness of state-of-the-art algorithms (neural and non-neural approaches) is affected by specific dataset structure, due to different session lengths, in terms of accuracy and distribution of content provider exposure.

Regarding the problem of miscommunication, and more concretely of hate speech detection, we organised the HatEval shared task at SemEval with Twitter corpora in English and Spanish (O3) and immigrants and women as target of hate speech (Basile et al., 2019). The twin problem of identifying offensive tweets (O2) in general (OffensEval shared task at SemEval) was addressed with a deep-learning ensemble method (De La Peña and Rosso, 2019). Special attention was given to the problem where targets of hate speech were women (Frenda et al., 2019a; 2019b). An extensive study of the different forms in that sexism attitudes and behaviours are found in social media has been performed, and different methods for automatically detecting everyday sexism in online communications have been developed (Rodríguez et al., 2020).

Hate speech in social media contributes to create increasingly polarized communities (Lai et al., 2019; 2020). At the moment, we are investigating the impact of hate speech and polarization (O2) on demoting argumentation[5] when controversial issues are addressed (Esteve et al., 2020). The preliminary work carried out to detect fake news and hate speech was covered also by the press[6].

Regarding the corpora developed (O2 and O3), we have created the NewsCom corpus, which contains 2,955 comments in Spanish posted in response to eighteen different news articles from online newspapers. This corpus has been annotated with the focus of negation, scope and negation markers (Taulé et al., 2020). Currently, we are annotating the comments of this corpus indicating their degree of toxicity: not toxic, midly toxic, toxic, very toxic. We are enriching this annotation with new and more detailed criteria. This annotation allows us to establish fine grained criteria for analysing and better defining what can be considered as a comment with toxic, offensive, abusive or hate language. This corpus could be used for the development of algorithms for the detection of this kind of language. Another line of research is the creation of a corpus of fake news from data provided by the Newtral[7] start-up with the aim to be used for training and testing algorithms for textual-based and multimodal fake news detection. These corpora could be used in future evaluation campaigns.

Regarding our work on methodological foundations of evaluation and textual similarity (O3): (i) we have completed a study on the evaluation of ordinal classification tasks (Amigó et al., 2020). Such tasks are very relevant for NLP and for the project, but they are currently evaluated with metrics for different problems (classification and regression). We have defined formal restrictions for the task and we have proposed a new metric, based on Information Theory and Measurement Theory, which complies with all the formal requisites and behaves better empirically than available alternatives. (ii) We have also completed the first stages of a formal study on similarity functions, which includes a specific axiomatics for similarity in the context of NLP and a new similarity function, which is most robust than state of the art alternatives. (iii) We have addressed the task of semantic compositionality, comparing word2vec approaches with contextual embeddings and

[4] https://pan.webis.de/clef20/pan20-web/author-profiling.html

[5] http://www.argumentsearch.com/

[6] https://www.lavanguardia.com/tecnologia/20191207/472082286311/usan-la-ia-para-detectar-noticias-falsas-y-mensajes-de-odio-en-redes-sociales.html

[7] https://www.newtral.es/

exploring unsupervised alternatives with a formal foundation.

Acknowledgments

The MISMIS project (PGC2018-096212-B) is funded by the Spanish Ministry of Science, Innovation and Universities.

References

Amigó, A., J. Gonzalo, S. Mizzaro and J. Carrillo. 2020. Effectiveness Metrics for Ordinal Classification: Formal Properties and Experimental Results. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Ariza, A., F. Fabbri, L. Boratto and M. Salamó. 2020. From The Beatles to Billie Eilish: Connecting Provider Representativeness and Exposure in Session-based Recommender Systems. Submitted to *SIGIR 2020*.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso and M. Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), (NAACL-HLT 2019), Minnesota, USA, 54-63.

De La Peña, G. P. and P. Rosso. 2019. DeepAnalyzer at SemEval-2019 Task 6: A Deep Learning-based Ensemble Method for Identifying Offensive Tweets. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), NAACL-HLT 2019, Minnesota, USA, 582–586.

Esteve, M., F. Casacuberta and P. Rosso. 2020. Minería de argumentación en el Referéndum del 1 de Octubre de 2017. *Procesamiento del Lenguaje Natural*, 65. To appear.

Frenda S., N. Kando, V. Patti and P. Rosso. 2019a. Stance or Insults?. *Proc. of the Ninth International Workshop on Evaluating Information Access* (EVIA 2019), Satellite Workshop of the NTCIR-14 Conference, National Institute of Informatics, Tokyo, Japan.

Frenda, S., B. Ghanem, M. Montes-y-Gómez and P. Rosso. 2019b. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5): 4743–4752.

Ghanem B., A. Cignarella, C. Bosco, P. Rosso, and F. Rangel. 2019a. UPV-28-UNITO at SemEval-2019 Task 7 Exploiting Post's Nesting and Syntax Information for Rumor Stance Classification. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), NAACL-HLT 2019, Minnesota, USA, 1125–1131.

Ghanem, B., G. Glavas, A. Giachanou, S. Ponzetto, P. Rosso and F. Rangel. 2019b. UPV-UMA at CheckThat! Lab: Verifying Arabic Claims using Crosslingual Approach. L. Cappellato, N. Ferro, D. E. Losada and H. Müller (eds.) *CLEF 2019 Labs and Workshops, Notebook Papers*. Workshop Proceedings.CEUR-WS.org, vol. 2380.

Ghanem B., P. Rosso and F. Rangel. 2019c. An Emotional Analysis of False Information in Social Media and News Articles. (arXiv:1908.09951). *ACM Transactions on Internet Technology* (TOIT). To appear.

Giachanou A., P. Rosso and F. Crestani. 2019. Leveraging Emotional Signals for Credibility Detection. *Proc. of the 42nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '19), Paris, France.

Lai, M., M. Tambuscio, V. Patti, G. Ruffo and P. Rosso. 2019. Stance Polarity in Political Debates: a Diachronic Perspective of Network Homophily and Conversations on Twitter. *Data & Knowledge Engineering*, 124. https://doi.org/10.1016/j.datak.2019.101738

Lai, M., V. Patti, G. Ruffo and P. Rosso. 2020. #Brexit: Leave or Remain? The Role of User's Community and Diachronic Evolution on Stance Detection. *Journal of Intelligent & Fuzzy Systems*. To appear.

Rodríguez-Sánchez, F. 2019. *Desarrollo de un sistema para la detección del machismo en redes sociales.* Trabajo Final de Máster en Tecnologías del Lenguaje. UNED.

Taulé, M., M. Nofre, M. González and M.A. Martí. 2020. Focus of negation: its identification in Spanish. *Natural Language Engineering*,1-22. https://doi.org/10.1017/S1351324920000388

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 105-108 recibido 14-03-2020 revisado 30-05-2020 aceptado 30-05-2020

AMALEU: Una Representación Universal del Lenguaje basada en Aprendizaje Automático

AMALEU: A Machine-Learned Universal Language Representation

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Resumen: El objetivo del proyecto AMALEU es aprender una representación común para diferentes idiomas. Se pretende tener una representación común para la lengua oral y una para la lengua escrita. AMALEU, de dos años de duración, está financiado por el MINECO dentro del programa de *Europa Excelencia*
Palabras clave: Traducción Automática Multilingüe, Traducción de Voz y Texto

Abstract: The objective of AMALEU's project is learning a multilingual common representation for speech and a different multilingual common representation for text. AMALEU is a two-year project funded by the MINECO.
Keywords: Multilingual Machine Translation, Speech and Text Translation

1 Participantes del proyecto

El grupo de investigación que participa en el proyecto pertenece al grupo de investigación de Voz del Departamento de Teoría de Señal y Comunicaciones de la Universidad Politècnica de Cataluña y al centro de investigación TALP. La investigadora principal es la autora de este artículo, y como investigadores están Carlos Escolano, Gerard Gallego y Javier Ferrando.

2 Entidad financiadora

El proyecto está financiado por el Ministerio de Economía y Competitividad y el código del proyecto es EUR2019-103819. AMALEU comenzó el 1 de enero de 2019, finaliza el 31 de enero de 2020 por lo que tiene una duración de 24 meses. La financiación total es de 74.850 euros.

3 Contexto y motivación

¿Por qué la traducción automática entre inglés y portugués es considerablemente mejor que la traducción automática entre holandés y castellano? ¿Por qué los reconocedores automáticos funcionan mejor en alemán que en finés? El principal motivo es la variación en la cantidad de datos etiquetados para entrenar y modelar los sistemas. Aunque el mundo es multimodal y altamente multilingüe, la tecnología de voz y lenguaje no es capaz de absorver la alta demanda que hay en todos los lenguajes. Necesitamos mejores algoritmos de aprendizaje que puedan explotar el progreso de unas pocas modalidades y lenguajes para el beneficio de otros.

Este proyecto se focaliza en el reto de aprendizaje a partir de pocos recursos a partir de una aproximación para la traducción automática multilingüe.

AMALEU propone entrenar conjuntamente un modelo multilingüe y multimodal que aprenda una representación universal del lenguaje. Este modelo compensará la falta de datos etiquetados y mejorará significativamente la capacidad de generalización de los datos de entrenamiento a partir de observar una variedad de recursos no etiquetados. Este modelo reducirá el número cuadrático de sistemas de traducción a linear, lo cual tendrá un gran impacto en un entorno multilingüe.

El reto de este proyecto radica en entrenar una representación universal de manera automática y con aprendizaje profundo. Para esto, AMALEU utilizará una arquitectura basada en la estructura de codificador-decodificador. El codificador aprende una representación del lenguaje de entrada mediante una reducción de dimensionalidad, que será la representación universal del lenguaje; a partir de esta abstracción, el decodificador genera la salida. La arquitectura interna del codificador-decodificador se diseñará explicitamente para aprender la abstracción

ISSN 1135-5948. DOI 10.26342/2020-65-14

universal del lenguaje, que se integrará como función objetivo de la arquitectura.

AMALEU impactará comunidades altamente multidisciplinares incluyendo ciencias de la computación, matemáticas, ingeniería y linguística, que trabajan conjuntamente en aplicaciones del procesado del lenguaje natural de voz y de texto.

4 Proyecto AMALEU

El objetivo de AMALEU es aprender de manera automática una representación universal del lenguaje ya sea voz o texto, de manera que se pueda explotar en aplicaciones de inteligencia artificial. Adicionalmente, el proyecto se plantea utilizar fuentes de información no etiquetadas así como información lingüística. La Figura 1 muestra el diagrama de Gantt del proyecto y los principales paquetes de trabajo. El plan de trabajo incluye: gestión y diseminación, representación multilingüe común, integración de recursos lingüísticos y la integración y evaluación. A continuación, describimos brevemente cada uno de estos cuatro puntos del plan de trabajo, así como el grado de consecución de los objetivos a fecha 1 de junio de 2020 (mes 17).

4.1 Gestión y Diseminación

Esta parte del trabajo incluye la adecuada gestión del proyecto preparando los informes adecuados y controlando el presupuesto adecuadamente. Además, se gestionará una diseminación elaborada que incluya publicaciones y participaciones en eventos internacionales. Las publicaciones relacionadas con el proyecto se van citando cuando se describen la consecución de los objetivos de cada tarea.

Consecución de los objetivos (mes 17). Este objetivo está completado al 75 %, resultados del mismo se pueden ver en acciones de diseminación como el desarrollo de la página web del grupo de investigación y del proyecto[1], la participación en la organización de evaluaciones internacionales (Barrault et al., 2019), y artículos de diseminación del área (Costa-jussà et al., 2020).

4.2 Representación Común de la Voz o el Texto Multilingües

El objetivo en este punto es obtener una representación común del lenguaje en su representación textual y en su representación oral, de manera independiente. La representación común del texto o la voz se construirá a partir de una arquitectura de codificador-decodificador. Dado que múltiples lenguajes se pueden entrenar en un único modelo que pueda producir traducciones multilingües, queremos aprender la representación multilingüe del lenguaje a partir del uso de una función objetivo común.

Dentro de esta parte del trabajo planteamos dos actividades. La primera es el desarrollo de la arquitectura para el lenguaje textual. Aquí nos focalizaremos en investigar mecanismos de atención y en diferentes representaciones intermedias que sean de longitud fija o variable. Evaluaremos tanto la calidad de la representación intermedia como la calidad de la traducción. La segunda actividad consistirá en construir una representación común para el lenguaje oral. Aquí la investigación se centrará en adaptar el codificador de manera que pueda aceptar la entrada oral que es considerablemente larga. Asimismo, identificaremos mecanismos de atención adecuados para tales longitudes. La parte del decodificador se compartirá en ambas actividades.

Consecución de los objetivos (mes 17). La representación común en texto se da por completada a partir de los siguientes trabajos: (Escolano, Costa-jussà, y Fonollosa, 2019; Escolano et al., 2019; Escolano et al., 2020a; Escolano et al., 2020b), donde se demuestra que la arquitectura que hemos planteado es capaz de acercar las representaciones de oraciones similares en distintos idiomas. La parte de voz está en curso.

4.3 Recursos Lingüísticos

El objetivo en este punto es utilizar información lingüística externa de manera que ayude a encontrar la representación común del lenguaje. Para eso se investigará sobre cuales son las mínimas unidades (palabras, subpalabras o caracteres) más adecuadas para conseguir una representación universal del lenguaje. Por otro lado se explotarán recursos de diferentes naturalezas incluyendo bases de datos monolingües.

Consecución de los objetivos (mes 17). Hemos explorado las unidades mínimas y información lingüística (Casas, Costa-jussà, y Fonollosa, 2020; Casas, Fonollosa, y Costa-jussà, 2020; Armengol-Estapé, Costa-jussà, y Escolano, 2020) así como el uso de datos monolingües (Biesialska, Rafieian, y Costa-

[1] http://mt.cs.upc.edu

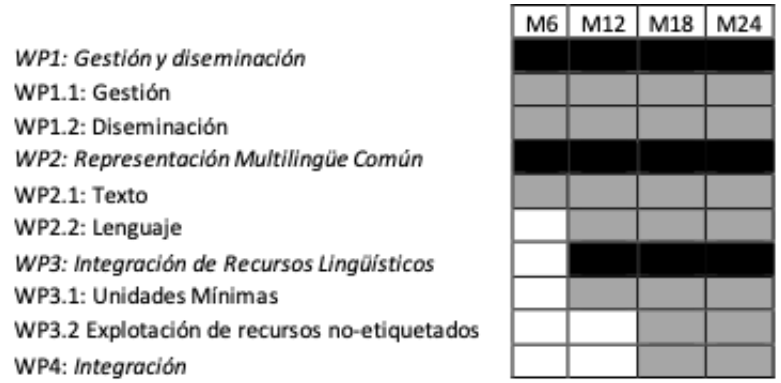

Figura 1: Diagrama Gantt. Plan de trabajo

jussà, 2020). Ahora falta su integración en la arquitectura multilingüe.

4.4 Integración/Evaluación

En cuando a datos, AMALEU usará datos multimodales de bases de datos existentes y conocidas como los que provienen de las evaluaciones del WMT[2] y del IWSLT[3] así como otros recursos disponibles desde el Linguistic Data Consortium[4].

Referente a la evaluación, como no existe un método establecido para evaluar la calidad de la representación intermedia, proponemos una nueva medida que utiliza la distancia de la representación entre oraciones similares. Esta distancia se combina con la evaluación de recuperación de la oración original a partir de su representación para calcular lo que denominamos Medida de Similitud con Recuperación. Asimismo, calcularemos con oraciones que son contradictorias, tienen una representación no similar (a partir de la medida de distancia entre oraciones) y combinaremos esta distancia con la recuperación de la frase original para calcular lo que denominamos Medida de Disimilitud con Recuperación. Estas medidas pretenden controlar que las oraciones similares tengan una representación similar y las oraciones contradictorias tengan una representación distante. La calidad de la traducción se evaluará con el método estandard BLEU (Papineni et al., 2002).

Consecución de los objetivos (mes 17). La transversalidad de esta tarea hace que su consecución se mida mayoritariamente en base a los objetivos previos. Además, queremos añadir que con nuestro sistema hemos participado en evaluaciones internacionales de prestigio (Casas et al., 2019) obteniendo siempre resultados competitivos.

5 Impacto

El impacto de AMALEU se refleja en una mejora de calidad y eficiencia de los actuales sistemas multilingües más allà de los traductores automáticos.

La novedosa integración de información multilingüe y multimodal en las aplicaciones de inteligencia artificial, que plantea AMALEU, permitirá mejorar la calidad de las mismas. Es clave que los sistemas pueden explotar múltiples recursos y se puedan entrenar a partir de datos no etiquetados. Como consecuencia, se consigue un aprendizaje zero-shot, que quiere decir que el sistema no necesita ver ejemplos de la tarea en concreto para aprenderla.

La extensión del entrenamiento a múltiples lenguas y modalidades contribuirá a mayores generalizaciones. Asimismo, la explotación de una alta variedad de recursos lingüísticos también lo hará.

La representación universal del lenguaje permite abrir nuevos horizontes en tareas como la búsqueda de información o los resúmenes automáticos cross-lingües. Asimismo, el uso de una representación común del lenguaje permitirá reducir el número de sistemas multilingües de N·(N-1) (donde N es el número de lenguas) a 2·N. Además nuestra aproximación permite añadir incrementalmente nuevas lenguas. Este logro consigue crear nuevas arquitecturas de aprendizaje profundo más efi-

[2]http://www.statmt.org/wmt20/
[3]https://workshop2020.iwslt.org/
[4]https://www.ldc.upenn.edu/

cientes, así como beneficiar las lenguas de pocos recursos a partir de las lenguas con más recursos.

Finalmente, mencionar que la investigación realizada en AMALEU tiene en cuenta producir resultados que sean justos y no reproduzcan sesgos sociales (Costa-jussà, 2019; Costa-jussà, Lin, y España-Bonet, 2020; Costa-jussà et al., 2020).

Agradecimientos

Este trabajo esta financiado por el Ministerio de Economía y Competitividad a través del projecto EUR2019-103819 y el programa Ramón y Cajal que incluye financiamiento del European Regional Development Fund.

Bibliografía

Armengol-Estapé, J., M. R. Costa-jussà, y C. Escolano. 2020. Enriching the transformer with linguistic and semantic factors for low-resource machine translation. *ArXiv*, abs/2004.08053.

Barrault, L., O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, y M. Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). En *Proceedings of the WMT*, páginas 1–61, Florence.

Biesialska, M., B. Rafieian, y M. R. Costa-jussà. 2020. Enhancing word embeddings with knowledge extracted from lexical resources. En *ACL Student Research Workshop*.

Casas, N., M. R. Costa-jussà, y J. A. R. Fonollosa. 2020. Combining subword representations into word-level representations in the transformer architecture. En *ACL Student Research Workshop*.

Casas, N., J. A. R. Fonollosa, y M. R. Costa-jussà. 2020. Syntax-driven iterative expansion language models for controllable text generation. *ArXiv*, abs/2004.02211.

Casas, N., J. A. R. Fonollosa, C. Escolano, C. Basta, y M. R. Costa-jussà. 2019. The TALP-UPC machine translation systems for WMT19 news translation task: Pivoting techniques for low resource MT. En *Proceedings of the WMT*, Florence, Italy.

Costa-jussà, M. R., R. Creus, O. S. Domingo, A. S. Dominguez, M. Escobar, C. I. López, M. Y. F. García, y M. Geleta. 2020. Mt-adapted datasheets for datasets: Template and repository. *ArXiv*, abs/2005.13156.

Costa-jussà, M. R. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Costa-jussà, M. R., C. España-Bonet, P. Fung, y N. A. Smith. 2020. Multilingual and interlingual semantic representations for natural language processing: A brief introduction. *Computational Linguistics*.

Costa-jussà, M. R., P. L. Lin, y C. España-Bonet. 2020. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. En *Proc of the LREC*.

Escolano, C., M. R. Costa-jussà, y J. A. R. Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. En *Proceedings of the ACL Student Research Workshop*, páginas 236–242, Florence.

Escolano, C., M. R. Costa-jussà, J. A. R. Fonollosa, y M. Artetxe. 2020a. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. *ArXiv*, abs/2004.06575.

Escolano, C., M. R. Costa-jussà, J. A. R. Fonollosa, y M. Artetxe. 2020b. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *ArXiv*, abs/2006.01594.

Escolano, C., M. R. Costa-jussà, E. Lacroux, y P.-P. Vázquez. 2019. Multilingual, multi-scale and multi-layer visualization of intermediate representations. En *Proceedings of the EMNLP-IJCNLP: System Demonstrations*, Noviembre.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of ACL*, páginas 311–318, Philadelphia, Pennsylvania, USA, Julio.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 109-112 recibido 01-04-2020 revisado 24-05-2020 aceptado 24-05-2020

Transcripción, indexación y análisis automático de declaraciones judiciales a partir de representaciones fonéticas y técnicas de lingüística forense

Transcription, indexing and automatic analysis of judicial declarations from phonetic representations and techniques of forensic linguistics

Pedro José Vivancos-Vicente[1], **José Antonio García-Díaz**[2], **Ángela Almela**[3], **Fernando Molina**[1], **Juan Salvador Castejón-Garrido**[1], **Rafael Valencia-García**[2]
[1]Vócali Sistemas Inteligentes S.L.
[2]Facultad de Informática, Universidad de Murcia
[3]Facultad de Letras, Universidad de Murcia
{pedro.vivancos, fernando.molina, juans.castejon}@vocali.net
{joseantonio.garcia8, angelalm, valencia}@um.es

Resumen: Recientes avances tecnológicos han permitido mejorar los procesos judiciales para la búsqueda de información en los expedientes judiciales asociados a un caso. Sin embargo, cuando técnicos y peritos deben revisar pruebas almacenadas en vídeos y fragmentos de audio, se ven obligados a realizar una búsqueda manual en el documento multimedia para localizar la parte que desean revisar, lo cual es una tarea tediosa y que consume bastante tiempo. Para poder facilitar el desempeño de los técnicos, el presente proyecto consiste en un sistema que permite la transcripción e indexación automática de contenido multimedia basado en tecnologías de deep-learning en entornos de ruido y con múltiples interlocutores, así como la posibilidad de realizar análisis de lingüística forense sobre los datos para ayudar a los peritos a analizar los testimonios de modo que se aporten evidencias sobre la veracidad del mismo.
Palabras clave: Reconocimiento de voz, word spotting, lingüística forense

Abstract: Recent technological advances have made it possible to improve the search for information in the judicial files of the Ministry of Justice associated with a trial. However, when judicial experts examine evidence in multimedia files, such as videos or audio fragments, they must manually search the document to locate the fragment at issue, which is a tedious and time-consuming task. In order to ease this task, we propose a system that allows automatic transcription and indexing of multimedia content based on deep-learning technologies in noise environments and with multiple speakers, as well as the possibility of applying forensic linguistics techniques to enable the analysis of witness statements so that evidence on its veracity is provided.
Keywords: Speech recognition, word spotting, forensic linguistics

1 Introducción

Los gobiernos de los países avanzados están llevando a cabo una serie de medidas de mejora y renovación tecnológica de los ministerios de justicia para mejorar los procesos judiciales y así reducir la posibilidad de errores, evitar la obsolescencia de equipos y mejorar la eficiencia en el uso de sus activos (Ballesteros, 2011). Algunas de estas medidas están enfocadas a mejorar la labor de los funcionarios de las administraciones públicas con el perfeccionamiento de sus servicios de comunicaciones internas, modernizar los sistemas informáticos, adaptarse al teletrabajo e incluir nuevas funcionalidades en sus sistemas

ISSN 1135-5948. DOI 10.26342/2020-65-15

internos.

Más concretamente, existen sistemas tecnológicos capaces de digitalizar en vídeo las vistas que tienen lugar en las salas de Justicia, como las descritas en (Zalman, Rubino, y Smith, 2019). Además, disponen de herramientas que permiten la búsqueda de información, tanto en el texto como en los metadatos y en los expedientes judiciales asociados a un caso. Sin embargo, cuando técnicos y peritos judiciales deben revisar pruebas almacenadas en documentos multimedia (como vídeos o fragmentos de audio), se ven obligados a realizar, o bien un visionado completo del vídeo, o bien probar su visualización desde cierto momento para localizar fragmentos que se consideren relevantes para el caso. Esta tarea es tediosa y consume una cantidad de tiempo considerable.

Para poder facilitar el desempeño de los técnicos es necesario poder clasificar y comprender el contenido de los recursos multimedia de manera eficiente. Los últimos avances en nuevas tecnologías permiten salvar estas dificultades con los consecuentes beneficios que ofrece disponer de una indexación eficiente del contenido multimedia. Con ello, se podrían reducir los ya dilatados procedimientos judiciales, suponiendo un ahorro a largo plazo y un beneficio para la ciudadanía.

Por otro lado, la lingüística forense, que es la rama de la lingüística aplicada que se encarga de estudiar los diversos puntos de encuentro entre lenguaje y derecho, ha demostrado su utilidad para aportar evidencias lingüísticas en los procesos judiciales. Esta disciplina se está implantando cada vez más en España y está mucho más extendida en países anglosajones, y pretende mostrar un conjunto de características lingüísticas que sea capaz de servir de apoyo en análisis del contenido de las declaraciones para determinar si ciertos testimonios o declaraciones tienen garantías de ser veraces.

El objetivo de este proyecto es el desarrollo de un sistema de transcripción, indexación y análisis automático de declaraciones judiciales a partir de representaciones fonéticas y técnicas de lingüística forense que permitan ayudar a los técnicos y peritos en el desempeño de su tarea, reduciendo así los ya dilatados procesos judiciales y mejorando la eficacia del sistema. El presente documento está estructurado de la siguiente manera. La sección 2 describe la arquitectura funcional del sistema, así como cada uno de los módulos que la componen mientras que la sección 3 describe su estado actual y nuevas vías de actuación.

2 Arquitectura del sistema

Este sistema está formado por tres módulos principales: 1) Sistema de transcripción de conversaciones de declaraciones judiciales basado en Deep-Learning (ver Sección 2.1), 2) Sistema de indexación y búsqueda de vídeos (ver Sección 2.2), y 3) Sistema de procesamiento de lingüística forense basado en características psicolingüísticas y fenómenos de duda en el discurso (ver Sección 2.3). En pocas palabras, el funcionamiento del sistema es el siguiente: A partir de un conjunto de vídeos y recursos multimedia introducidos de manera manual, el sistema transcribe automáticamente su contenido a ficheros WebVTT; en segundo lugar, estos ficheros se utilizan para construir un sistema de indexación para poder buscar sobre el vídeo a partir de citas textuales palabras relacionadas o bien a través de búsqueda fonética. Al final se provee a los peritos con una interfaz web donde pueden buscar sobre el fragmento exacto del vídeo a la vez que se genera un informe de lingüística forense que ayude a determinar la veracidad o duda de un determinado testimonio.

A continuación, se describen brevemente estos subsistemas.

2.1 Sistema de transcripción de conversaciones de declaraciones judiciales basado en Deep-Learning

El sistema de transcripción es capaz de aplicar un proceso de transcripción automática de los vídeos basada en tecnologías de Deep-Learning, mejorando su desempeño en situaciones de entornos con ruido de ambiente (Hannun et al., 2014), con varias personas hablando al unísono (Snyder et al., 2019) y sin la necesidad de un entrenamiento específico por parte de los interlocutores.

El resultado final de este primer sistema son ficheros en formato WebVTT (Pfeiffer y Hickson, 2013), que es un formato para mostrar pistas de texto cronometrado que se utilizan en sistemas de películas, vídeos en streaming, etc. Este formato permite incluir metadatos para añadir información asociada a los datos. De esta forma, se enriquecen las transcripciones realizadas añadiendo información

Marcador	**Ejemplos**
Vaguedad o exactitud	Cantidades inexactas, fechas, nombres, lugares, ...
Incertidumbre o certeza	Expresiones como *a menudo*, *hasta donde yo sé*, ...
Distanciamiento del hecho	Uso de pronombres personales en tercera persona, ...
Minimización	Expresiones como *lo que pasó*, *el incidente*, ...
Refuerzo de credibilidad	Adverbios como *honestamente*, *sinceramente*, ...
Generalizaciones	Referencias a colectivos o el uso de determinantes indefinidos, ...
Comunicación evitativa	Eufemismos, evasivas, interrupciones, ...
Actitud cooperativa	Expresiones como *tienes razón*, *es cierto*, *así es*
Egocentrismo	Uso de pronombres en primer personal del singular, ...
Repetición	Palabras duplicadas o expresiones como *ambos dos*
Exceso de lenguaje culto	Cultismos, latinismos, verbos en subjuntivo, ...

Tabla 1: Marcadores de discurso falaz

Marcadores	**Ejemplos**
Tiempo y espacio	países, capitales, lugares, fechas, rangos de tiempo, ...
Simplicidad	Uso de verbos en indicativo simple, frases cortas, ...
Experiencias subjetivas	Fenómenos percibidos a través de los sentidos
Orden	Expresiones como *de una parte*, *por un lado*, *para terminar*, ...
Detalles	Nombres propios, lugares, eventos, empresas, profesiones, ...
Naturalidad	Respuestas cortas

Tabla 2: Marcadores de discurso veraz

sobre el contexto situacional, así como qué hablante produjo la frase registrada.

2.2 Sistema de indexación y búsqueda de vídeos

El sistema de indexación y búsqueda de vídeos es el encargado de indexar y catalogar cada uno de los distintos vídeos en repositorios. Este sistema permite realizar búsquedas precisas de vídeos a partir de palabras clave, fonemas y frases parecidas e identificar y cargar el vídeo en el fragmento preciso. Para ello, este sistema es capaz de buscar en distintos tipos de formato multimedia (formatos AVI, MP4) y buscar entre la información transcrita a partir del sistema anterior y además realizar búsquedas directamente en los fragmentos de audio a través del modelado de fonemas.

2.3 Sistema de procesamiento de lingüística forense basado en características psicolingüísticas y fenómenos de duda en el discurso

El sistema de procesamiento de lingüística forense basado en características psicolingüísticas y fenómenos de duda en el discurso se encarga del análisis y extracción de características lingüísticamente relevantes relacionadas con análisis de veracidad del testimonio. Para ello, se ha desarrollado un sistema de análisis lingüístico diseñado en español basado en (Salas-Zárate et al., 2017) para evaluar distintas características para determinar si un determinado relato es veraz o, si por el contrario, es engañoso, como el estudio presentado en (Almela, Valencia-García, y Cantos, 2012). Para ello, el sistema analiza el discurso y ofrece al perito un informe que contiene fenómenos lingüísticos determinados que permiten ver si se minimizan los hechos, si hay negaciones excesivas, una actitud colaborativa, lapsus linguae, o bien si el discurso es coherente, consistente, simple, equilibrado y natural. Esta información se muestra a los peritos a través de una interfaz intuitiva que resalta dentro del texto cuáles son las palabras y frases clave identificadas por el análisis.

En la Tabla 1 y en la Tabla 2 se muestran algunos de los marcadores usados para identificar el relato falaz y veraz así como algunos ejemplos de cada marcador. Para poder extraer información del texto en base a estos marcadores se han empleado reconocedores de entidades, corpus anotados como los analizados en (Jiménez-Zafra et al., 2020) y el uso lexicones.

3 Trabajo futuro

Actualmente nos encontramos en la última anualidad del proyecto. Se está terminando el desarrollo de cada uno de los subsistemas y realizando distintas evaluaciones de rendimiento y fiabilidad.

Las tecnologías que hasta el momento nos han dado mejor resultado en la transcripción de conversaciones se basan en modelos DNN-HMM (modelos ocultos de Markov con redes neuronales profundas) (Ravanelli y Omologo, 2017). Los tipos de redes neuronales más efectivas son aquellas que, por tener en cuenta elementos pasados, se ajustan mejor a la clasificación de series temporales, como son Time Delay Neural Networks (TDNN) (Sun et al., 2017) y Long-Short Time Memory (LSTM) (Zhang et al., 2016). La indexación y búsqueda de vídeos se basa en tecnologías de Keyword Spotting, que trata de identificar una serie de palabras clave en señales acústicas. Además, se están estudiando otras variables de lingüística forense con el fin de tener en cuenta los metadatos de la conversación para medir la latencia, el tono y la velocidad del hablante.

Por último, se realizará la integración de los módulos en un prototipo final de la plataforma global. En este sentido, se planificarán distintas pruebas de campo para comprobar la calidad del sistema completo.

Agradecimientos

Este proyecto ha sido financiado por el Instituto de Fomento de la Región de Murcia con fondos FEDER dentro del proyecto con referencia 2018.08.ID+I.0025

Bibliografía

Almela, Á., R. Valencia-García, y P. Cantos. 2012. Detectando la mentira en lenguaje escrito. *Procesamiento del lenguaje natural*, 48:65–72.

Ballesteros, M. C. R. 2011. La necesaria modernización de la justicia: especial referencia al plan estratégico 2009-2012. *Anuario jurídico y económico escurialense*, (44):173–186.

Hannun, A., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, y others. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Jiménez-Zafra, S. M., R. Morante, M. Teresa Martín-Valdivia, y L. A. Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Pfeiffer, S. y I. Hickson. 2013. Webvtt: The web video text tracks format. *Draft Community Group Specification, W3C*.

Ravanelli, M. y M. Omologo. 2017. Contaminated speech training methods for robust dnn-hmm distant speech recognition. *arXiv preprint arXiv:1710.03538*.

Salas-Zárate, M. P., M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, y G. Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowl. Based Syst.*, 128:20–33.

Snyder, D., D. Garcia-Romero, G. Sell, A. McCree, D. Povey, y S. Khudanpur. 2019. Speaker recognition for multi-speaker conversations using x-vectors. En *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 5796–5800. IEEE.

Sun, M., D. Snyder, Y. Gao, V. K. Nagaraja, M. Rodehorst, S. Panchapagesan, N. Strom, S. Matsoukas, y S. Vitaladevuni. 2017. Compressed time delay neural network for small-footprint keyword spotting. En *INTERSPEECH*, páginas 3607–3611.

Zalman, M., L. L. Rubino, y B. Smith. 2019. Beyond police compliance with electronic recording of interrogation legislation: Toward error reduction. *Criminal Justice Policy Review*, 30(4):627–655.

Zhang, Y., G. Chen, D. Yu, K. Yaco, S. Khudanpur, y J. Glass. 2016. Highway long short-term memory rnns for distant speech recognition. En *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 5755–5759. IEEE.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 115-118 recibido 01-04-2020 revisado 24-05-2020 aceptado 13-06-2020

EmoCon: Analizador de Emociones en el Congreso de los Diputados

EmoCon: Emotions Analyzer in the Spanish Congress

Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, María Teresa Martín-Valdivia, Andrea López-Fernández
SINAI, Centro de Estudios Avanzados en TIC (CEATIC), Universidad de Jaén
{sjzafra, magc, maite, alfernan}@ujaen.es

Resumen: EmoCon es un prototipo de un analizador de emociones en el Congreso de los Diputados. Su objetivo es analizar el perfil emocional a nivel de sesión parlamentaria y a nivel de cada diputado, a partir de las intervenciones realizadas durante las sesiones parlamentarias que tienen lugar en el Congreso de los Diputados. Para ello, la demo cuenta con tres módulos principales: i) descarga automática de los documentos de las sesiones y extracción de las intervenciones realizadas por cada diputado, ii) análisis de las emociones expresadas a nivel de sesión y a nivel de diputado y, iii) visualización de la información en una aplicación web.
Palabras clave: Análisis de emociones, Congreso de los Diputados, política

Abstract: EmoCon is a prototype of an emotion analyzer in the Spanish Congress. Its objective is to analyze the emotions expressed by the deputies in the interventions made during the parliamentary sessions that take place in the Spanish Congress. To this end, the demo has three main modules: i) web scrapper for the session documents and processing, ii) emotions analyzer at the session level and at the deputy level, and iii) web application for visualization.
Keywords: Emotion analysis, Spanish Congress, politics

1 Introducción

Las emociones son un aspecto muy importante de nuestra vida. Lo que hacemos y lo que decimos refleja en parte las emociones que sentimos. La RAE define emoción como una "alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática". Desde el punto de vista psicológico las emociones se suelen definir como un complejo estado afectivo, una reacción subjetiva que ocurre como resultado de cambios fisiológicos o psicológicos que influyen sobre el pensamiento y la conducta. Estas emociones se pueden analizar a través de las expresiones faciales, de la voz, del texto, etc.

La detección de emociones ha sido un tema de gran interés en disciplinas como la neurociencia o la psicología y, actualmente, está atrayendo la atención de los investigadores en ciencias de la computación. El área que se encarga del estudio de las emociones a nivel computacional se conoce como *Computación Afectiva* (Cambria, 2016). En este trabajo nos centramos en esta área, en la tarea de reconocimiento de emociones en textos, que consiste en identificar de forma automática las emociones expresadas en un texto. En concreto, presentamos una demo que tiene como objetivo aplicar técnicas y herramientas de las Tecnologías del Lenguaje Humano para analizar y detectar las emociones que expresan los políticos en los diarios de sesiones del Congreso de los Diputados, fruto de de sus intervenciones. La finalidad de esta demo es mostrar el perfil emocional de los diputados del Congreso a partir de sus intervenciones.

Existen varias clasificaciones posibles de categorías de emociones (Ekman, 1992; Plutchik, 1986; Lövheim, 2012). EmoCon utiliza la clasificación de Paul Ekman (Ekman, 1992), al ser una de las más usadas y que más recursos tiene disponibles. Ekman define seis emociones básicas: enfado, miedo, asco, sorpresa, alegría y tristeza.

2 Motivación

En el ámbito de la política, el análisis de emociones se ha utilizado fundamentalmente para saber qué actitudes y propuestas de los candi-

ISSN 1135-5948. DOI 10.26342/2020-65-16

datos electorales provocan en los ciudadanos mejores respuestas emocionales, con el objetivo de realizar sondeos políticos en vista a unas elecciones. Nuestro enfoque es bien distinto y está centrado en el análisis de las emociones de los propios parlamentarios por sus propios contenidos, los cuales podrían estar relacionados con la reacción de los ciudadanos.

3 Descripción del sistema

EmoCon extrae y analiza los diarios de sesión de las intervenciones de los diputados en el Congreso de los Diputados (debate político), detectando las emociones en los mismos. Cuenta con tres módulos principales:

1. Descarga de documentos de las sesiones y procesamiento
2. Analizador automático de emociones
3. Aplicación web para la visualización de la información procesada

La Figura 1 muestra el esquema general de EmoCon.

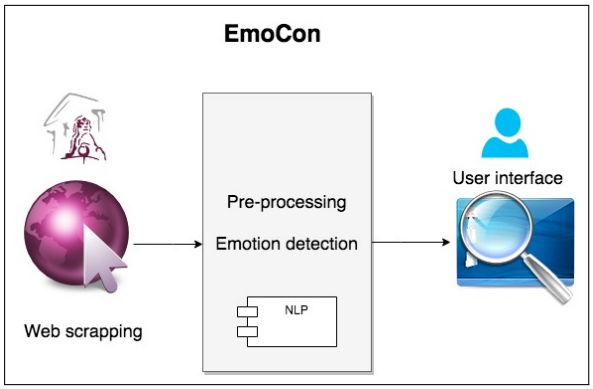

Figura 1: Esquema general de EmoCon

A continuación se detallan cada uno de estos módulos.

3.1 Módulo 1: descarga y procesamiento de datos

El sistema trabaja con información de las intervenciones realizadas en las sesiones parlamentarias que tienen lugar en el Congreso de los Diputados. Estos documentos están accesibles a traves del portal web del Congreso de los Diputados[1]. Los documentos están clasificados por legislatura, año, y mes, lo que facilita la localización de una sesión específica. En concreto, la demo presentada realiza un análisis de la XII legislatura, es decir la que transcurre desde el 19 de julio de 2016 hasta el 5 de marzo de 2019, y se continúan descargando y analizando documentos de la legislatura vigente. Actualmente el sistema trabaja con un total de 640 documentos de sesión de 384 diputados.

Para la descarga de la información se ha implementado un crawler, aprovechando que las URLs de estos documentos siguen un patrón[2], y se ha utilizado Python como lenguaje de programación. En la Figura 2 se muestra un ejemplo de un extracto de uno de los PDFs descargados.

La señora **PRESIDENTA:** Señorías, por favor, guarden silencio.

El señor **VENDRELL GARDEÑES:** ¿Es que a nosotros no nos preocupa el déficit y la deuda? Claro que nos preocupa. ¿Qué deuda nos preocupa? Los ingentes recursos destinados al rescate de las autopistas, a las estaciones de alta velocidad, al Castor, los 3.500 millones de la estafa eléctrica de los costes de transición a la competencia, los 26.300 millones de ayudas a la banca que no volveremos a ver más... y la corrupción, señores del Partido Popular. La corrupción también genera deuda porque con una mano ustedes han recortado servicios públicos y con otra han amparado la corrupción. Y nos preocupa otro tipo de déficit: el déficit social. El hecho de que en España, por ejemplo, 1.700.000 personas más son pobres en relación con el año 2009; o la pobreza infantil, que genera deuda social y traslada la desigualdad y la exclusión a las generaciones futuras.

La desigualdad crece con el desempleo, la precariedad y por la debilidad de las políticas públicas. ¿Cuál es nuestro problema? Que ingresamos ocho puntos menos de la media europea y somos el segundo país de la Unión Europea, después de Irlanda, con menos gasto público. Ese es el problema. Las políticas de austeridad son el eje del proyecto ideológico del Partido Popular: menos Estado del bienestar y más desigualdad. Pero, eso sí, más Estado centralizado. No podemos desaprovechar esta ocasión para realizar un cambio profundo de la ley que ponga fin a las políticas de austeridad injustas y recentralizadoras. Poner fin de una vez por todas a la hegemonía de la austeridad.

Gracias. **(Aplausos).**

La señora **PRESIDENTA:** Gracias.
Tiene la palabra el señor Montero Soler.

El señor **MONTERO SOLER:** Buenas tardes, señorías.
Hago mías las consideraciones que han hecho los miembros de mi grupo parlamentario que me han precedido en el uso de la palabra y plantearé algunas cuestiones de forma muy sintética por el tiempo tan restringido.

Figura 2: Extracto de un PDF de una sesión

De cada uno de los PDFs descargados el sistema extrae las intervenciones de los diputados, texto que posteriormente es analizado en el módulo 2 para determinar las emociones expresadas por cada diputado. Esto es posible ya que como se puede observar en la Figura 2, en estos documentos se sigue un patrón para mostrar cada una de las intervenciones de los distintos diputados: *El señor/La señora + nombre en mayúscula y negrita + dos puntos + intervención del diputado.*

3.2 Módulo 2: análisis de emociones

Este módulo es el encargado de analizar las emociones presentes en los textos extraídos de cada una de las sesiones parlamentarias. El sistema realiza dos tipos de análisis: uno a nivel de sesión y otro a nivel de diputado.

Para analizar las emociones utiliza el modelo de las seis emociones de Ekman comentado previamente y un clasificador automático basado en lexicón. El clasificador hace uso

[1] http://www.congreso.es/portal/page/portal/Congreso/Congreso/Publicaciones/DiaSes/Pleno

[2] http://www.congreso.es/public_oficiales/L12/CONG/DS/PL/DSCD-12-PL-[numsesion].PDF

del lexicón "Spanish Emotion Lexicon", conocido como SEL por su siglas en inglés[3] (Sidorov et al., 2012), e incorpora una heurística para el tratamiento de negadores y modificadores.

El analizador de emociones a nivel de sesión calcula el perfil emocional general de la sesión y para cada diputado determina en qué porcentaje ha expresado alguna de las seis emociones de Ekman. En primer lugar, realiza un preprocesamiento sobre el texto asociado a cada diputado: i) tokenización mediante expresiones regulares, ii) extracción de raíces utilizando el paquete NLTK Snowball stemmer[4], iii) eliminación de palabras vacías, y iv) conversión del texto a minúscula. En segundo lugar, realiza la clasificación de las emociones presentes en el texto de la siguiente forma:

```
diputado = {'ira': 0, 'asco': 0, 'miedo': 0,
'alegría': 0, 'tristeza': 0, 'sorpresa': 0}
for palabra in texto do
    if palabra in lexicon then
        if anterior is negador then
            peso = 0
        else if anterior is increment then
            peso = 1,75
        else if anterior is atenuante then
            peso = 0,75
        else
            peso = 1
        end if
    end if
    for emocion in lexicon[palabra] do
        fpa = lexicon[palabra][emocion]
        diputado[emocion] += peso * fpa
    end for
end for
```

El analizador de emociones a nivel de diputado calcula su perfil emocional en base a sus intervenciones en todas las sesiones y en qué porcentaje ha expresado cada una de las seis emociones a lo largo del tiempo.

3.3 Módulo 3: aplicación web para visualización

Una vez extraída, procesada y analizada la información, se desarrolló una aplicación web para mostrarla. Esta aplicación hace uso del framework *Flask*[5] de Python, de los lenguajes HTML, CSS y JavaScript, y de la herramienta Google Charts[6].

Tal y como se ha comentado en el módulo 2, el analizador de emociones permite realizar dos tipos de análisis: uno a nivel de sesión y otro a nivel de diputado (Figura 3).

Figura 3: EmonCon - Página principal

Mediante la opción *sesión*, la aplicación web permite filtrar por fecha de sesión, consultar el PDF de cada sesión y visualizar el perfil emocional completo de la sesión (Figura 4) y de un diputado concreto ese día (Figura 5). En el perfil general de la sesión aparecerá un diagrama de sectores que indicará el porcentaje en el que ha estado presente cada emoción ese día y seis rankings, uno por cada emoción, en los que se mostrará los cinco diputados que más han sentido cada emoción. En el perfil del diputado seleccionado aparecerá un diagrama de sectores representando los porcentajes en los que ha expresado las distintas emociones en la sesión celebrada. Además, se mostrará una gráfica de barras en la que se podrá comparar las emociones del diputado ese día concreto con las emociones que suele mostrar, es decir, con su perfil habitual.

Mediante la opción *diputados* es posible seleccionar un miembro de los grupos parlamentarios (Figura 6), accediendo a su perfil emocional habitual por medio de un gráfico de sectores y de un histórico en el que se mostrará la evolución de este perfil del diputado (Figura 7). Además, en el diagrama del histórico se puede seleccionar un día concreto y acceder a la correspondiente sesión.

4 Conclusiones y trabajo futuro

En este trabajo se ha presentado EmoCon, un prototipo para analizar las emociones expresadas en el Congreso de los Diputados, con el objetivo de obtener y mostrar el perfil emo-

[3]El lexicón SEL está formado por 2.036 palabras en español que están asociadas con al menos una emoción básica (ira, asco, miedo, alegría, tristeza y sorpresa) a través del factor de probabilidad de uso afectivo (FPA por su siglas en inglés).

[4]http://www.nltk.org/_modules/nltk/stem/snowball.html

[5]https://www.fullstackpython.com/flask.html

[6]https://developers.google.com/chart/

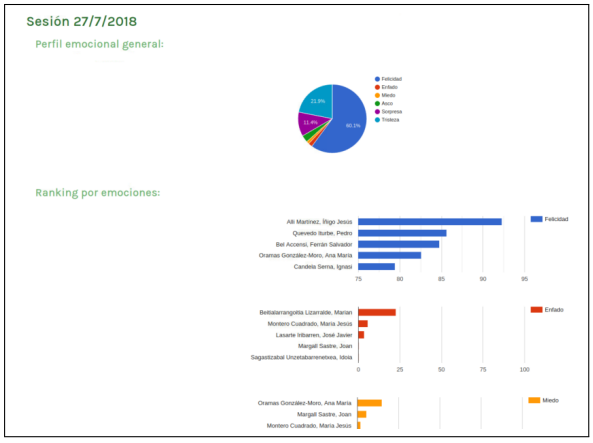

Figura 4: EmonCon - Análisis de una sesión: perfil completo de la sesión

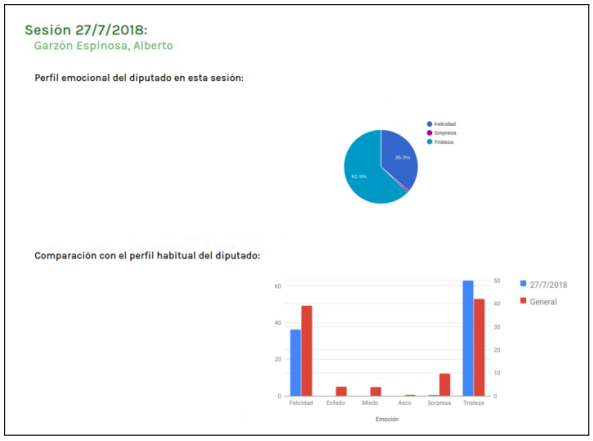

Figura 5: EmonCon - Análisis de una sesión: perfil de un diputado

cional a nivel de sesión parlamentaria y de cada diputado que interviene.

Como posible trabajo futuro de este sistema nos planteamos estudiar otros métodos de análisis de emociones y analizar la legislatura actual. Además, sería interesante añadir un apartado de temáticas para reflejar los temas tratados en cada sesión y así poder ver qué emociones se generan en función de lo que se esté hablando. Así mismo queremos analizar perfiles de usuarios y partidos políticos para comprobar si los miembros de un mismo partido tienen un perfil similar, analizando sus intervenciones en el Congreso y lo que publican en medios sociales.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto LIVING-LANG (RTI2018-094653-B-C21) del Gobierno de España.

Figura 6: EmonCon - Análisis de un diputado: selección del grupo parlamentario

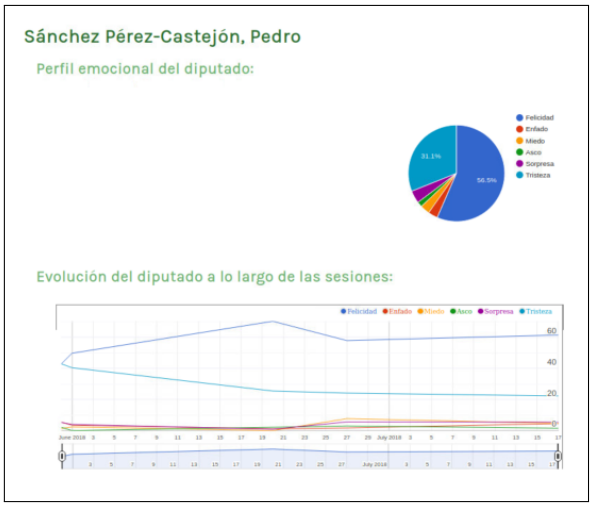

Figura 7: EmonCon - Análisis de un diputado: perfil general de un diputado

Bibliografía

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Lövheim, H. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2):341 – 348.

Plutchik, R., . K. H. 1986. *Emotion: Theory, Research and Experience*. New York: Academic Press.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, y J. Gordon. 2012. Empirical study of machine learning based approach for opinion mining in tweets. En *Mexican international conference on Artificial intelligence*, páginas 1–14. Springer.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 119-122 recibido 30-03-2020 revisado 24-05-2020 aceptado 12-06-2020

Nalytics: Natural Speech and Text Analytics

Nalytics: Analíticas del Habla y Texto Naturales

Ander González-Docasal[1], Naiara Pérez[1], Aitor Álvarez[1], Manex Serras[1]
Laura García-Sardiña[1], Haritz Arzelus[1], Aitor García-Pablos[1], Montse Cuadros[1]
Paz Delgado[2], Ane Lazpiur[2], Blanca Romero[2]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain) +34 943 309 230
{agonzalezd, nperez, aalvarez, mserras, lgarcias, harzelus, agarciap, mcuadros}@vicomtech.org
[2]Natural Vox S.A.U., Bulevar de Salburua Kalea 8, 01002 Vitoria-Gasteiz (Spain)
+34 945 227 200 {pdelgado, ane, bromero}@naturalvox.com

Abstract: Call centres have long demanded technology for analysing the data they manage. In this context, we present Nalytics, a platform that integrates Speech and Text Analytics in Spanish in a modular design, and capable of customising its models to the users' demands.
Keywords: Speech Analytics, NLP, Deep Learning, Call Centres

Resumen: Los centros de llamadas han demandado durante años soluciones tecnológicas para analizar los datos que gestionan. En este contexto, presentamos Nalytics, una plataforma que integra el análisis de voz y texto en español en un diseño modular capaz de personalizar sus modelos a demanda del cliente.
Palabras clave: Analítica del Habla, PLN, Aprendizaje Profundo, Centros de Llamadas

1 Introduction

Traditionally, call centres have addressed their conversational quality audits manually, by means of agents who manage to analyse $5 - 10\%$ of the recorded material at best. Therefore, the market has long demanded Speech Analytics tools for the automatic analysis of conversational content.

Nowadays, companies which offer Speech Analytics services can be categorised into two groups: *a)* those who offer a black box flow that handles the client directly and comes with numerous functionalities and advanced technology, but through solutions in the cloud; and *b)* those who offer more adapted services but rely on more traditional search technology, such as word spotting, and work on batch mode.

Nalytics combines the best of both cases. It includes advanced technology based on machine and deep learning techniques, operates in both batch and online modes, it can be deployed as a Software as a Service (SaaS) or on premise, and offers the possibility of adapting its technological modules to the application domain of the customers. Nalytics integrates technological modules to perform Speech Analytics, including speaker diarisation and rich transcription as its main components; and Text Analytics, including segmentation and normalisation, sentiment analysis, and text classification. Within its modular design, the solution allows different combinations of these modules considering the users' needs, resulting on a powerful platform for the automatic analysis of call centres' content in Spanish.

The rest of the paper is structured as follows: Section 2 introduces a general overview of the platform; Section 3 describes the developed front-end application for the use of the platform; Sections 4 and 5 explain the technological modules for speech and text analytics, respectively; Section 6 describes the process of including new models to the solution; and finally, Section 7 adds some conclusions to this work.

2 General architecture

Nalytics is a back-end service built in Python and hosted on a Linux system. It is integrated as a web service inspired on the REST architecture, in which the communication with the REST client is performed through HTTP requests and, if necessary, JSON objects sent to a specific address (i.e., IP and

ISSN 1135-5948. DOI 10.26342/2020-65-17

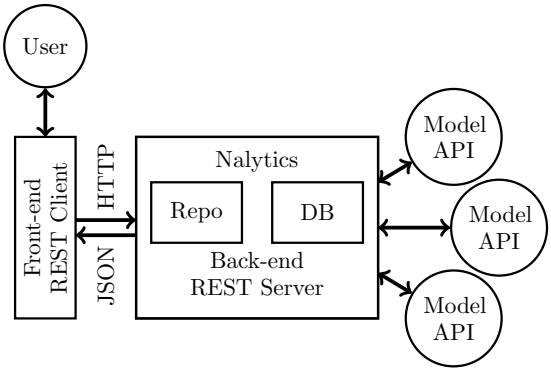

Figure 1: Minimal schematic diagram of the architecture of Nalytics

port). Once a request is received, the REST server responds with an informative JSON object and an HTTP status code. The platform is complemented by a database and a physical repository where all the information, data, and results are stored. In addition, informative web services retrieve information such as previously sent requests or status of ongoing processes. A minimal schematic diagram of the architecture of Nalytics is shown in Figure 1.

In order to service multiple customers that may have contracted different licences, all the HTTP requests are put in a priority queue inside Nalytics, where the management of the individual priorities is left to the main REST client.

Nalytics supports multiple request types for data analysis, including *offline* and *online* decoding over *synchronous* and *asynchronous* methods. For *offline* decoding, the platform loads the corresponding model in memory and frees it once the result is returned. On the contrary, in *online* decoding, the model is previously loaded into an API, which is assigned a free port and remains deployed to attend the incoming requests. It allows to operate in a real-time scenario, with a very low latency.

Additionally, in *asynchronous* methods the tasks are executed in the background depending on the available hardware resources in the server. If the request is successful, the result may be retrieved later using an informative service. In the *synchronous* requests, the master server waits for the result and returns it directly to the REST client as part of the HTTP response. However, if the task takes more than a previously defined and configurable time, it returns a timeout error. If so, the request is treated as *asynchronous* and the client is able to retrieve the result from the repository once it is completed.

Finally, the technological modules can be applied individually or as a pipeline, in what is known as a *combo* decoding. The longest *combo* decoding involves the execution of all the modules in chain from the audio preprocessing and speech transcription to text analysis (segmentation and normalisation, sentiment analysis or classification). In Figure 2, the accepted request module sequences are shown.

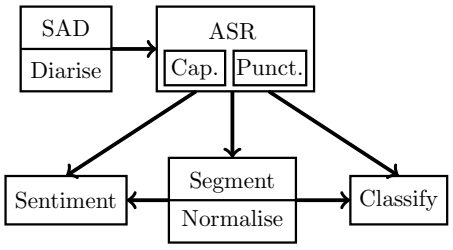

Figure 2: Accepted request submodule sequences in inference. Every module in the graph is a starting and ending node

3 *Front-end application*

Thanks to the RESTful API of Nalytics, it can be easily integrated with any front-end client. This application can serve as the main interface for communication between the user and the back-end. In this project, the client application was developed by the company Natural Vox and it is currently being served as a SaaS (Software as a Service) solution to a number of user clients. In the Figure 3, a screenshot of the current front-end application is presented.

The current front-end solution from Natural Vox is a web application that the users

Figure 3: The Natural Vox front-end application for Nalytics

can access through any web browser. The main objective is to provide a powerful tool to analyse speech and text data from call centres. The users can analyse audios by means of the rich transcription and speaker diarisation modules, whilst input texts or draft transcriptions can be classified or analysed at sentiment level. In addition, new categories for classification can be currently incorporated, adapting the modules to their domain and contents. It is worth mentioning that a rule-based anonymisation module is also included to deal with personal data. Finally, the results are generated in CSV format which are later converted into formal reports using Qlik[1].

4 Speech Analytics

This section presents the speech analysis modules in more detail, grouped as *Audio preprocessing* and *Speech transcription*.

4.1 Audio preprocessing

Firstly, the incoming audio data from the telephonic domain (e.g., raw format compressed with A-law, μ-law or G.729 codecs) is transcoded to a known audio format for Nalytics (PCM WAV 16 kHz and 16 bits). Then, the audio can be optionally processed by the Speech Activity Detection (SAD) and Speaker Diarisation modules, with the aim of discarding non-speech segments and performing speakers segmentation and clustering, respectively. Both modules have been developed using the Kaldi toolkit (Povey et al., 2011).

4.2 Speech transcription

The core of any Speech Analyitcs solution relies on the Automatic Speech Recognition (ASR) engine. Two different ASR architectures have been integrated into Nalytics. The first was estimated with the Kaldi toolkit and is composed of a hybrid Long-Short Term Memory-Hidden Markov Model (LSTM-HMM) acoustic model, where unidirectional LSTMs were trained to provide posterior probability estimates for the HMM states. In addition, a modified Kneser-Ney smoothed 3-gram language model, trained with the KenLM toolkit (Heafield, 2011), is used for decoding. The second system corresponds to an E2E neural network based on the Baidu's Deep Speech 2 architecture (Amodei et al., 2016), with an Kneser-Ney smoothed 5-gram used to rescore the hypothesis. Both acoustic and language models were estimated with the SAVAS corpus (del Pozo et al., 2014) transferred through land- and mobile-lines and augmented, as explained in (Bernath et al., 2018).

The automatic transcription can then be optionally sent through the capitalisation and punctuation modules, which take use of bidirectional Recurrent Neural Networks (RNNs) with an attention mechanism (Tilk and Alumäe, 2016) and a *recasing* model trained with the Moses toolkit (Koehn et al., 2007), respectively.

5 Text Analytics

This section introduces the different modules for text analysis available in Nalytics: *Text preprocessing*, *Text classification* and *Sentiment analysis*.

5.1 Text preprocessing

The text preprocessing module can be subdivided in two different tasks: *Normalising* the text, that is, correcting misspelling errors; and *Segmenting* or dividing each sentence into spans with potentially different polarities (e.g., "I loved the service but my problem was not solved" starts with positive polarity but ends with a negative remark).

The normalisation module is built on Hunspell[2]. Through configurable dictionaries and morphotactic rules, Hunspell detects possible typographic and orthographic errors and suggests correction candidates for each. Then, the actual correction is chosen using heuristics that involve Levenshtein's distance and knowledge about frequent orthographic errors in Spanish.

With respect to segmentation, the module first locates in the input text adversative conjunctions and phrases that may indicate contrast (e.g., "but", "on the contrary", etc.). Then, it obtains the dependency tree of the text using spaCy[3]. Finally, the text is segmented via heuristics based on these two pieces of information: each segment is a phrase or clause headed by the token an adversative phrase depends on. This module's behaviour can also be configured by incorporating new segmentation rules through configuration files.

[1] `https://www.qlik.com`

[2] `https://hunspell.github.io`

[3] `https://spacy.io/api/dependencyparser`

5.2 Text classification

Nalytics offers a diverse set of classification algorithms in an attempt to cover a wide range of application scenarios –read data availability, computational power, and so on–: *a)* Multinomial Naive Bayes (MNB) and Support Vector Machines (SVM), as implemented in scikit-learn[4]; *b)* spaCy's ensemble implementation[5] of Convolutional Neural Networks (CNN); *c)* finally, a learner for sequence-labelling tasks based on Conditional Random Fields (CRF), from sklearn-crfsuite[6].

5.3 Sentiment analysis

The sentiment analysis module is a specialisation of the Text classification module; that is, it offers the same text classification model learners and decoding functionalities, but oriented towards the sentiment analysis task. Specifically, the classifiers benefit from a rich set of features well-known for improving sentiment classification results, such as Brown clusters and gazetteer lookup. In addition, this module takes special advantage of the Text preprocessing module, which divides clauses with potentially different polarity.

6 Adapted models acquisition

As previously stated, Nalytics includes mechanisms for incorporating new models to the internal database. In the scope of speech recognition, externally created models can be transferred by placing them in the repository and executing a single HTTP request that registers the model in the database. This is implemented both for Kaldi and E2E models, whilst it is not supported for the SAD and diarisation models yet. New rules for the text preprocessing segmentation submodule can be registered in a similar way.

In the case of Text Analytics, new models can be trained directly from Nalytics. If the training process is successful, the resulting model is automatically registered in the database and placed in the repository for future use. Nalytics accepts all the training parameters exposed by the aforementioned libraries (Section 5.2).

These interesting functionalities allow Natural Vox to offer customised models to their user clients, considering that all the technological modules included in Nalytics perform better if they are adapted to the application domain.

7 Conclusions

This work presents a new solution for Speech and Text Analytics. Nalytics takes advantage of the last modelling paradigms to construct a back-end platform that can be easily integrated with any REST based client. Its operation modes, the available request methods, its modularity, powerful functionalities along with the variety of its technological modules turns Nalytics into an attractive solution that is already being exploited in the market.

References

Amodei, D., S. Ananthanarayanan, R. Anubhai, et al. 2016. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In *Proceedings of ICML 2016*, volume 48, pages 173–182.

Bernath, C., A. Alvarez, H. Arzelus, and C. D. Martínez. 2018. Exploring E2E speech recognition systems for new languages. In *Proceedings of IberSPEECH 2018*, pages 102–106.

del Pozo, A., C. Aliprandi, A. Álvarez, et al. 2014. Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling. In *Proceedings of LREC 2014*, pages 432–436.

Heafield, K. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of WMT 2011*, pages 187–197.

Koehn, P., H. Hoang, A. Birch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL*, pages 177–180.

Povey, D., A. Ghoshal, G. Boulianne, et al. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*.

Tilk, O. and T. Alumäe. 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *Proceedings of INTERSPEECH 2016*.

[4] `https://scikit-learn.org`

[5] `https://spacy.io/api/textcategorizer#architectures`

[6] `https://sklearn-crfsuite.readthedocs.io`

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 123-126 recibido 30-03-2020 revisado 24-05-2020 aceptado 02-06-2020

RESIVOZ: Dialogue System for Voice-based Information Registration in Eldercare

RESIVOZ: Sistema de Diálogo para el Registro de Información de Cuidado de Mayores mediante Voz

Laura García-Sardiña, Manex Serras, Arantza del Pozo, Mikel D. Fernández-Bhogal
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain), +[34] 943 30 92 30
{lgarcias, mserras, adelpozo, mfernandez}@vicomtech.org

Abstract: RESIVOZ is a spoken dialogue system aimed at helping geriatric nurses easily register resident caring information. Compared to the traditional use of computers installed at specific control points for information recording, RESIVOZ's hands-free and mobile nature allows nurses to enter their activities in a natural way, when and where needed. Besides the core spoken dialogue component, the presented prototype system also includes an administration panel and a mobile phone App designed to visualise and edit resident caring information.
Keywords: Gerontechnology, Spoken Dialogue Systems, Eldercare, Information Registration

Resumen: RESIVOZ es un sistema de diálogo orientado a ayudar a gerontólogos a registrar fácilmente información sobre sus cuidados a los residentes. En comparación con el uso tradicional de ordenadores instalados en puntos de control específicos para registrar la información, la naturaleza manos-libres y móvil de RESIVOZ permite al personal gerontológico registrar sus actividades de forma natural, donde y cuando se necesite. Además del principal componente de sistema de diálogo hablado, el prototipo de sistema también incluye un panel de administración y una aplicación móvil diseñada para visualizar y editar la información de cuidados a residentes.
Palabras clave: Gerontotecnología, Sistemas de Diálogo Hablado, Cuidado de Mayores, Registro de Información

1 Introduction

Spoken Dialogue Systems (SDS) are increasingly being integrated and deployed into a wide range of devices we can find in our daily lives. Many smart phones, speakers, and cars come with conversational assistants (e.g. Alexa, Siri, Google Assistant, Cortana) which can provide us with useful information (e.g. weather forecasts) and help us with various simple tasks (e.g. playing music). Because of their popularity for infotainment purposes, the application of SDS is also in demand and being explored in professional fields like healthcare (Laranjo et al., 2018). In particular, the use of conversational agents to support different kinds of patient populations (e.g. the elderly, the chronically ill) with health tasks is an emerging field of research with several projects underway, like EMPATHIC[1], MENHIR[2] or SHAPES[3].

The use case presented in this paper is also linked to healthcare and aims to facilitate the registration of resident information by geriatric nurses throughout a work shift. The implemented SDS provides them with a hands-free, natural, and flexible interface to register resident information in a nursing home management system, allowing them to devote more time to practice people-based care.

The paper is structured as follows: Section 2 presents our application use case; Section 3 describes the system's components; Section 4 displays a sample dialogue showing different interaction situations; and Section 5 gives some conclusions and future work directions.

[1] www.empathic-project.eu/
[2] www.menhir-project.eu/
[3] cordis.europa.eu/project/id/857159

ISSN 1135-5948. DOI 10.26342/2020-65-18

2 Use Case

The application scenario is that of an elderly care residence where geriatric nurses need to register information regarding the activities they carry out with each resident into a management system. Such activities include: oral hygiene, shaving, grooming, changing diapers, showering, checking overall appearance, cutting and checking fingernails, performing postural changes, checking bowel movements, and checking how much residents eat and drink in each meal. Sometimes, nurses also need to add observations to these activities and follow-up comments for each resident, so that other carers in following shifts are informed about particular incidents.

One of the issues with the traditional computer-based systems is that there is usually just one shared computer per floor or unit. Geriatric nurses usually do not have the time to stop by and register activities as they are completed, so they tend to wait until the end of their shift to enter all the activities carried out throughout their working day. This entails two main problems: on the one hand, queues are generated, since all geriatric nurses want to register all the information simultaneously at the end of their shift; on the other hand, the information introduced may not be entirely correct due to the difficulty of remembering all activities carried out with each resident throughout a shift.

The proposed solution integrates a mobile spoken dialogue system that allows geriatric nurses to introduce the information in an easy, hands-free way anywhere and anytime throughout their working day. Given the nature of the geriatric nurses' work, the solution's selected hardware needs to be: (i) **ergonomic** it needs to be lightweight, wireless and wearable, since it needs to be carried throughout the whole shift; (ii) **safe**, some wearable devices like smartwatches can cause abrasion injuries on residents' skin and should be avoided; and (iii) **discreet**, audio capturing and playing should not be invasive nor interfere with social interactions.

3 System Architecture

Figure 1 shows a diagram of RESIVOZ's architecture. To comply with the identified requirements, the selected hardware devices include (i) a single-earphone plus short microphone Bluetooth headset, and (ii) a lightweight, small smartphone. The microphone has soft noise cancelling and mute option, so geriatric nurses can disable audio recording when talking to someone other than the system. The smartphone allows mobile audio capturing, App usage, and response generation using its built-in Text-To-Speech (TTS) functionalities.

As for software, the developed prototype system consists of four main components: 1) a Spoken Dialogue System that allows user-system interaction, 2) a Mobile Phone App for information visualisation and edition, 3) a Control Panel Interface to manage users, shifts, and resident caring activities, and 4) a Domain Database where all the necessary information is stored.

3.1 Spoken Dialogue System

This is RESIVOZ's core component, acting as a voice-based interaction bridge between the final users and the rest of the system. The SDS receives the information to save and lets the users know whether the it has been correctly logged in or there has been some problem in the process: there is no identified resident, the information is incomplete or invalid for the current shift, there has been a database connection issue, or the system could not parse the user's input. The implemented SDS is a pipeline of four submodules:

(1) Automatic Speech Recognition (ASR) In this module, the audio stream captured by the headset goes to an energy-based Voice Activity Detector, which separates speech from non-speech segments and sends the former to a Speech-To-Text (STT) service. Here, Google Cloud's STT API service was integrated after introducing some custom domain phrases, words, and names as *contexts* for domain adaptation[4].

(2) Spoken Language Understanding (SLU) For this module, given that real domain data was not available, Phoenix grammars (Ward, 1991) were developed to parse the transcribed geriatric nurses' input into semantic codifications of *act-slot-value* sets. A hierarchical taxonomy was defined covering the domain. It includes four acts (*identify*, *inform*, *deny*, *delete*), 34 slots, and 28 fixed values (as named entities, resident names are considered variable values and not included in the taxonomy). Apart from parsing

[4] `cloud.google.com/speech-to-text/docs/context-strength`. Data sharing options were disabled in compliance with the GDPR

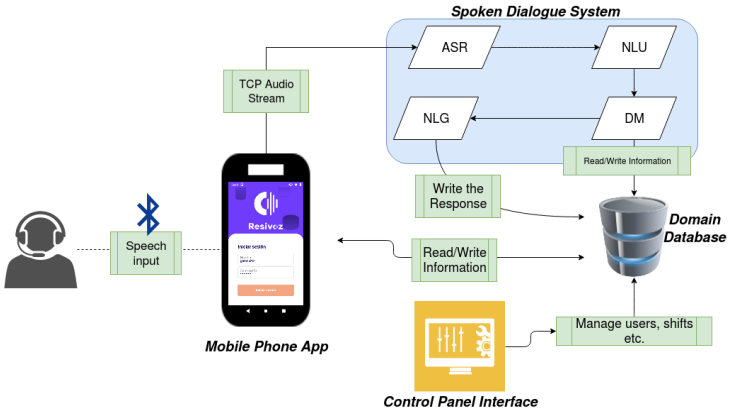

Figure 1: Architecture of the RESIVOZ system

the semantic actions from users' inputs, the NLU module also identifies named entities – resident names in this case– from users' utterances, using a fuzzy string-matching method and their registered full names. Residents are identified with their name and first surname[5], while their second surname is used for disambiguation when needed.

(3) Dialogue Manager (DM) Decision making and interaction planning is based on Attributed Probabilistic Finite State Bi-Automata (Torres, 2013), which store the latest user input, system output, and attributes (i.e. dialogue memory). The latter store task-sensitive information such as whether a resident is identified, or if there is any ambiguity with the provided information (Serras, Torres, and del Pozo, 2019). The DM is also responsible for reading and writing information in the Domain Database and for modifying attributes of the dialogue state. To leverage the lack of data, the next action is selected using *if this, then that* rules.

(4) Natural Language Generation (NLG) Natural language text templates were generated for each possible system action so as to transmit an adequate response to users in a human-interpretable manner.

3.2 Mobile Phone App

An Android App was developed to serve as a front-end and access point for the geriatric nurses. This is meant to be used at sporadic times, since nurses need to have both hands free while on their caring duties. When they log in, the information on their shift is shown: their assigned resident group, daily tasks, previous follow-up comments on residents, and so on. Also, a TCP connection is established with the server, starting the audio stream. The application has three purposes: 1) to be a gateway to easily display and modify the information that is being logged into the database; 2) to handle the socket connection and send the audio stream to the server where the back-end is located; and 3) to play the text responses given by the SDS using the phone's built-in TTS.

3.3 Control Panel Interface

A control panel web interface was developed to allow managers to easily administer shift-related data. From this panel, shifts and nurse and resident profiles can be created and deleted in the database in a user-friendly way.

3.4 Domain Database

The domain database is where all the information is registered and read from: resident and user profiles, the shifts scheduled by the manager, and the caring activities, observations and follow-up comments introduced by the geriatric nurses for each resident. It also stores the users' sessions during their assigned shifts. A series of customised RESTful API calls were implemented to communicate the database with the SDS, the Mobile Phone App, and the Control Panel.

4 Sample Dialogue

This section presents a simulated dialogue[6] between a geriatric nurse (U) and the system (S), showing some of the different situations that may happen during interaction.

[5]In Spain, unlike in other countries, people have two surnames.

[6]The names are made up and do not correspond to specific real residents. The dialogue is presented in English although RESIVOZ uses Spanish.

(1) U: Mouth cleaning okay
S: There is no resident identified
U: Ana Atxa had her teeth brushed
S: Oral hygiene done for resident Ana Atxa Cruz

In the first interaction, presented in example 1, no resident is identified, so the system will not register the information until one is. Then, the system is able to identify a resident given their name and first surname.

(2) U: Resident Pilar García
S: There are several residents with that name
U: Pilar García Zubia
S: Resident Pilar García Zubia has been identified

In example 2, the user tries to identify a resident by name and first surname, but there are several residents with that name. By using the second surname, the target resident is correctly disambiguated.

(3) U: She had half her meal for lunch
S: Intake of solids at lunchtime 50%
U: Add observation to intake of solids at lunchtime
S: Dictate
U: Pilar said she was too full and refused to eat more, she didn't want dessert either, *end of annotation*
S: Observations to solid intake at lunchtime saved

Although the user is not saying a name explicitly, in example 3 the identified resident is known from the session memory. In this example, the user adds an observation, i.e. a free text comment, to a registered control. Dictation ends when a pre-established command is uttered: "end of annotation".

(4) U: All my residents but Ander Goñi have been showered
S: Shower registered for all your residents but Ander Goñi Uria

RESIVOZ incorporates shortcuts to allow the selection of several residents and/or controls at the same time. Example 4 shows the use of a shortcut where all the residents in the nurse's assigned group but one are selected.

(5) U: Delete shower for Ander Goñi
S: No completed control with that name has been found

In example 5, the user tries to delete a control for a resident, but this control was registered before for all residents in the group except for the now identified. The system informs that it cannot delete a control that has not been saved as complete.

5 Conclusions and Future Work

We have developed and integrated a Spoken Dialogue System that helps geriatric nurses in an elderly care centre easily register resident caring information at any time and any place during their work shifts. Using RESIVOZ over the traditional system, only available in a few shared computers at specific control points, may lend geriatric nurses more quality time for their elderly caring activities. For future work, an evaluation by geriatric nurses is to be performed in their workplace, tackling different aspects such as system usability, points of improvement, difficulties found, adoption period, and other feedback regarding the system's usability and applicability in the real work environment.

Acknowledgements

This work was supported by the Adinberri Programme of the Provincial Council of Gipuzkoa, via the *RESIVOZ* project.

References

Laranjo, L., A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau, and E. Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Serras, M., M. I. Torres, and A. del Pozo. 2019. Goal-conditioned user modeling for dialogue systems using stochastic bi-automata. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, pages 128–134.

Torres, M. I. 2013. Stochastic bi-languages to model dialogs. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, pages 9–17.

Ward, W. 1991. Understanding spontaneous speech: The phoenix system. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 365–367.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 127-130 recibido 28-03-2020 revisado 30-05-2020 aceptado 30-05-2020

The Text Complexity Library

Biblioteca de Complejidad Textual

Rocío López-Anguita, Jaime Collado-Montañez, Arturo Montejo-Ráez
Centre for Advanced Information and Communication Technologies (CEATIC)
University of Jaén, Spain
{rlanguit, jcollado, amontejo}@ujaen.es

Abstract: This paper introduces a new resource for computing textual complexity. It consists in a Python library for calculating different complexity metrics for several languages from plain texts. The resource has been made available to the research community and provides all needed instructions for its installation and use. To our knowledge, it is the first time a resource like this is published, so we expect many researchers can profit from it.
Keywords: Demostration, linguistic resources, textual complexity, lexical analysis

Resumen: Este artículo presenta un nuevo recurso para el cálculo de la complejidad textual. Se trata de una biblioteca de programación en Python que facilita el cómputo de distintas métricas de complejidad para varios idiomas a partir de textos en lenguaje natural. El recurso se ha liberado para su uso por parte de la comunidad científica y proporciona todas las instrucciones necesarias para su instalación y aprovechamiento. Hasta donde sabemos, es la primera vez que un recurso así está disponible, por lo que esperamos sea de utilidad.
Palabras clave: Demostración, recursos lingüísticos, complejidad textual, análisis léxico

1 Introduction

Reading comprehension and reading competence are complex processes that are closely related, according to Pérez (2014), as is the concept of readability. The reading comprehension is close to the reader's capabilities and the latter is an objective view of the complexity of the text.

Determining the readability of a text is not a simple task, as each reader has different skills or limitations (Cain, Oakhill, and Bryant, 2004). It is usually determined by linguistic features, which are usually grouped into those related to grammar (in other words, syntax) and those related to the lexicon (i.e. vocabulary) (Alliende González, 1994).

Currently, we consider that measures of complexity can be a convenient way to model natural language in certain applications, such as authorship detection, text selection for people with difficulties associated with language disorders (autism, cerebral palsy...), or early detection of cognitive impairments, such as Alzheimer's. Therefore, in this article we present a "demo" paper, which consists of 12 of the most widely used metrics for lexical and syntactic readability. These measures and their interpretation are presented below, as well as details on the use of the library.

2 Complexity metrics provided

In this section, we introduce the different complexity metrics offered in this Python library, proposed by different authors, for different languages (Spanish, English, French...).

Lexical complexity: The lexical complexity of a text, determined by the frequency of use and lexical density, was proposed by Anula (2008). It is based on the number of different content words per sentence (*Lexical Complexity Index, LC*) and on measuring the number of low frequency words per 100 content words (*Index of Low Frequency Words, ILFW*). Consequently, the higher the LC index, the greater the difficulty in reading comprehension.

Spaulding readability: Commonly known as the SSR Index, it was proposed by Spaulding (1956). It focuses on measuring vocabulary and sentence structure to predict

ISSN 1135-5948. DOI 10.26342/2020-65-19

the relative difficulty of a text's readability. Its formula is an empirically adjusted measure to try to keep the score between 0 and 1.

Complexity of sentences: The Sentence Complexity Index (SCI) was proposed by Anula (2008), as a measure of the complexity of sentences in a literary text aimed at second language learners.

This syntactic complexity measure focuses on measuring the number of words per sentence, thus obtaining the sentence length index (*Average Sentence Length, ASL*), and the number of complex sentences per sentence, from a complex sentence index (*Complex Sentences, CS*).

Automated Readability Index (ARI): Senter and Smith (1967) proposes one of the most used indexes due to its ease of calculation, the Automated Readability Index, better known as ARI (*Automated Readability Index*). This index measures the difficulty of a text from the average number of characters (letters and numbers) per word and the average number of words per sentence.

Dependency tree depth: This measure was proposed by Saggion et al. (2015). It is a very useful metric to capture syntactic complexity: long sentences can be syntactically complex or contain a large number of modifiers (adjectives, adverbs or adverbial phrases). It complements the ASL measure, as it captures syntactic complexity in terms of recursive or nested structures.

Punctuation Marks: This measure was also proposed by Saggion et al. (2015). In the complexity of a text, the average number of punctuation marks is used as one of the indicators of the simplicity of the text.

Readability of Fernández-Huerta: Blanco Pérez and Gutiérrez Couto (2002) y Ramírez-Puerta et al. (2013) propose this measure of complexity as an adaptation to Spanish of Flesch's readability test (Flesch (1948)).

Readability of Flesch-Szigrist (IFSZ): The works of Barrio-Cantalejo et al. (2008) and Ramírez-Puerta et al. (2013) propose the Flesch-Szigristzt readability index as a modification of the Flesch formula (Flesch, 1948) adapted to Spanish by Szigriszt-Pazos in 1993. This index is currently considered a reference for the Spanish language. It focuses on measuring the number of syllables per word and the number of sentences per word in the text.

Comprehensibility of Gutiérrez de Polini: This metric, originally developed in 1972, is not an adaptation of English, but was created from the beginning for Spanish (Rodríguez, 1980). It focuses on measuring the average number of letters per word and the average number of words per sentence.

μ Readability: It is a formula to calculate the readability of a text. It provides an index between 0 and 100 and was developed by Muñoz (2006). This measure focuses on measuring the number of words, the average number of letters per word and their variance.

Minimum age to understand: In work of García López (2001) we can find another formula to measure the age needed to understand a text. It is, again, an adaptation into Spanish of Flesch's original formula (Flesch (1948)) for English. It measures the average number of syllables per word and the average number of words per sentence to obtain the minimum age needed to understand a text.

SOL Readability: Contreras et al. (1999) proposes the SOL metric as an adaptation to Spanish of the SMOG formula proposed by Mc Laughlin (1969). It measures the readability of a text by means of grade level, which is the number of years of schooling required to understand the text.

Years Crawford: This measure was proposed by *Alan N. Crawford* in 1989 (Crawford, 1984). It is used to calculate the years of school required to understand a text. Measures the number of sentences per hundred words and the number of syllables per hundred words.

3 How to obtain it

The library has been released under the General Public License (GPL v3.0) license[1] and can be downloaded or cloned from its public repository[2]. The library will be updated with new features in the future, and you can always get the latest version from that link.

[1] `http://www.gnu.org/licenses/gpl.html`

[2] `https://gitlab.ujaen.es/amontejo/text-complexity`

4 Installation

In order to use this library, you first need to install some previous requirements:

- NumPy, Scipy, Pandas, Matplotlib and Openpyxl for python3 have to be installed in your system.
- The FreeLing (Padró and Stanilovsky, 2012) package, which is a library providing language analysis services that our library makes use of. In order to install it, you have to follow its installation manual under the project's GitHub page[3]

5 Usage examples

In order to test the library and teach how to use it, we have prepared some testing texts for Spanish under the `./texts` folder (you can use your own texts by modifying these files or adding more to that location).

To compute complexity metrics on these text samples, modify the `FREELINGDIR` variable in the `TextComplexityFreeling.py` script (line 18) to your own FreeLing installation directory (`/usr/local` by default).

Then, if you run the Jupyter notebook `examples.ipynb` you should get some tables with the metrics for each text provided. The script will also generate three MS Excel files containing the results in your project's folder.

To use it in your Python scripts, this is as simple as follows:

```
import TextComplexityFreeling as TCF
# Create the text complexity calculator
tc = TCF.TextComplexityFreeling()
# Load text to analyze
text_processed = tc.textProcessing(text)
# Compute different metrics
pmarks = tc.punctuationMarks()
lexcomplexity = tc.lexicalComplexity()
ssreadability = tc.ssReadability()
sencomplexity = tc.sentenceComplexity()
autoreadability = tc.autoReadability()
embeddingdepth = tc.embeddingDepth()
readability = tc.readability()
agereadability = tc.ageReadability()
yearscrawford = tc.yearsCrawford()
```

For example, for a given sample text:

La última luna llena del año, que se observará completa este jueves en el cielo, será especial. Se produce estos días el fenómeno conocido como luna fría, una coincidencia astronómica que hace las delicias de quienes atribuyen al astro cualidades esotéricas. Sucede cuando la Tierra se encuentra ubicada exactamente entre el sol y la luna, de forma que la luna recibe directamente la luz. La luna llena será visible durante toda la noche, pero alcanzará su magnitud máxima cuando se encuentre a medio cielo, de forma que, al reflejar completamente la luz del sol que incide en la tierra, se verá especialmente grande y luminosa. Se llama luna fría porque marca la llegada del invierno en el hemisferio norte, aunque también se conoce como luna de las noches largas al ocurrir cerca del solsticio, informa National Geographic, que cita a un astrónomo de la NASA. La luna fría de 2019 coincide además con la lluvia de meteoros Gemínidas, visible entre el 7 y el 17 de diciembre pero que alcanzará su punto máximo de actividad entre el 11 y el 13. Es la lluvia de estrellas más masiva, lo que la hace mucho más brillante. El cielo augura todo un espectáculo esta noche.

This is the generated output:

```
Number of words (N_w):  207
Punctuation marks: 21
Number of low freq. words (N_lfw):  67
Number of content words (N_dcw):  71
Number of sentences (N_s):  8
Number of total content words (N_cw):  93
Lexical Distribution Index (LDI): 8.875
Index Low Frequency Words (ILFW):  0.72
Lexical Complexity Index (LC): 4.7977
Number of rare words (N_rw):  65
Spaulding Spanish Readability (SSR): 167.82
Average Sentence Length (ASL): 25.875
Complex Sentences (CS): 23.75
Sentence Complexity Index:(SCI): 24.81
Automated Readability Index (ARI): 13.66
Average embddings depth (MeanDEPTH): 7.25
Huerta's Readability index:  80.86
IFSZ Readability: 54.25
Polani's Compressibility (Polani's): 42.61
Mu Readability: 53.19
Minimum age to understand: 12.48
SOL Readability: 11.66
Years needed: 5.76
```

6 Conclusions and future versions

There are several studies that reflect the strong influence of the richness of the reader's vocabulary on reading comprehension, beyond symbols and grammar. However, we consider that complexity measures can be a convenient way to model natural language in

[3] `https://github.com/TALP-UPC/FreeLing-User-Manual`

certain applications, such as authorship detection, text selection for people with difficulties associated with language disorders (autism, cerebral palsy...), or early detection of cognitive impairments, such as Alzheimer.

Another future line of work is to define complexity from computed language models (like RNNs or BERT models). We believe that information measures on the parameters of these models may capture the inherent complexity of the texts they were trained on.

7 Acknowledgements

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

References

Alliende González, F. 1994. La legibilidad de los textos. *Santiago de Chile: Andrés Bello*, 24.

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. *La evaluación en el aprendizaje y la enseñanza del español como LE L*, 2:162–170.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. In *Anales del Sistema Sanitario de Navarra*, volume 31, pages 135–152. SciELO Espa{ na.

Blanco Pérez, A. and U. Gutiérrez Couto. 2002. Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. *Revista española de salud pública*, 76(4):321–331.

Cain, K., J. Oakhill, and P. Bryant. 2004. Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, 96(1):31.

Contreras, A., R. Garcia-Alonso, M. Echenique, and F. Daye-Contreras. 1999. The sol formulas for converting smog readability scores between health education materials written in spanish, english, and french. *Journal of health communication*, 4(1):21–29.

Crawford, A. N. 1984. A spanish language fry-type readability procedure: Elementary level. bilingual education paper series, vol. 7, no. 8.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

García López, J. 2001. Legibilidad de los folletos informativos. *Pharmaceutical Care España*, 3(1):49–56.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Muñoz, M. 2006. Legibilidad y variabilidad de los textos. *Boletín de Investigación Educacional, Pontificia Universidad Católica de Chile, 21*, 2:13–26.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479.

Pérez, E. J. 2014. Comprensión lectora vs competencia lectora: qué son y qué relación existe entre ellas. *Investigaciones sobre lectura*, (1):65–74.

Ramírez-Puerta, M., R. Fernández-Fernández, J. Frías-Pareja, M. Yuste-Ossorio, S. Narbona-Galdó, and L. Peñas-Maldonado. 2013. Análisis de legibilidad de consentimientos informados en cuidados intensivos. *Medicina Intensiva*, 37(8):503–509.

Rodríguez, T. 1980. Determinación de la comprensibilidad de materiales de lectura por medio de variables lingüísticas. *Lectura y vida*, 1(1):29–32.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Senter, R. and E. A. Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Spaulding, S. 1956. A spanish readability formula. *The Modern Language Journal*, 40(8):433–441.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 131-134 recibido 30-04-2020 revisado 26-05-2020 aceptado 03-06-2020

The Impact of Coronavirus on our Mental Health

El Impacto del Coronavirus en nuestra Salud Mental

Jiawen Wu[1], **Francisco Rangel**[1], **Juan Carlos Martínez**[2]
[1]Symanto, Germany
[2]Atribus, Spain
{jiawen.wu,francisco.rangel}@symanto.net, jcmartinez@atribus.com

Abstract: The Coronavirus represents the greatest threat to physical health in modern times. Simultaneously, fear of the unknown and the fear of the very real repercussions of the virus is threatening to impact the mental health of many around the world. To provide insights on the impact of Coronavirus on our mental health, we are constantly monitoring millions of conversations on Twitter each day, and analysing this enormous amount of data by means of psychological models trained with artificial intelligence techniques and deep neural networks.
Keywords: Coronavirus, COVID19, mental health, psychological models, deep learning

Resumen: El Coronavirus representa la mayor amenaza para la salud física en tiempos modernos. A su vez, el miedo a lo desconocido y a las repercusiones reales del virus, está amenazando con impactar en la salud mental de las personas alrededor de todo el mundo. Para analizar dicho impacto, estamos monitorizando millones de conversaciones en Twitter en tiempo real, y analizando esta gran cantidad de datos mediante modelos psicológicos entrenados con técnicas de inteligencia artificial y redes neuronales profundas.
Palabras clave: Coronavirus, COVID19, salud mental, modelos psicológicos, aprendizaje profundo

1 Introduction

Social media allow instant and borderless communication of what happens in people's daily lives, giving immediate access to knowledge that would otherwise be limited or biased at the discretion of a few. This is even more important when it comes to spreading information regarding threats. And we are living the greatest threat to physical health in modern times: the Coronavirus. Add into this the huge lifestyle changes we're all being asked to make and it's easy to see why our emotions and feelings are being affected. Our aim is at analysing the impact of the virus on the mental health of people around the world through our psychological models trained with artificial intelligence.

At Symanto[1], we have created two live trackers based on English and German[2] tweets with mentions of Coronavirus-related terms. Furthermore, together with our partner Atribus[3], we have created another tracker to monitor the mental health of people in Spanish speaking countries[4].

In this paper, we summarise some of our key findings in the development of mental health discussion in different countries, the triggering factors, the rising of digital psychotherapy and its perception, and much more.

2 Social Media Monitoring and Data Collection

We are constantly retrieving discussions about the Coronavirus on Twitter. We search for terms such as Covid-19, Coronavirus or SARS-Cov-2. We exclude spam (e.g., the same tweet changing only a user mention) and retweets. For this paper, we have analysed the data collected between March 10 and April 22. In Table 1, we present the to-

[1]https://www.symanto.net/
[2]https://www.symanto.net/live-insights/mental-health-coronavirus/
[3]https://www.atribus.com/
[4]https://www.atribus.com/covid19-esplatam/

ISSN 1135-5948. DOI 10.26342/2020-65-20

tal number of tweets and the total number of tweets referring to mental health issues[5].

Lang.	# Tweets	# Mental health
English	14M	811K
German	1.6M	75K
Spanish	11.2M	1.1M

Table 1: Number of tweets collected between March 10 and April 22

3 *Mental Health Issues amid Coronavirus Pandemic*

Within the conversations around Coronavirus, an essential volume of discussion is around mental health issues, which strongly indicates the negative impact of the pandemic on our mental health. We are closely monitoring both the volume of these conversations and the concrete issues in the discussion. *Anxiety* is by far the most dominant mental health issue, followed by *Stress*, *Depression*, *OCD*, and *Suicide*. Though less significant, there are also mentions of *Eating Disorder*, *Personality Disorder*, and *Self-harm* in relation to Covid-19[6].

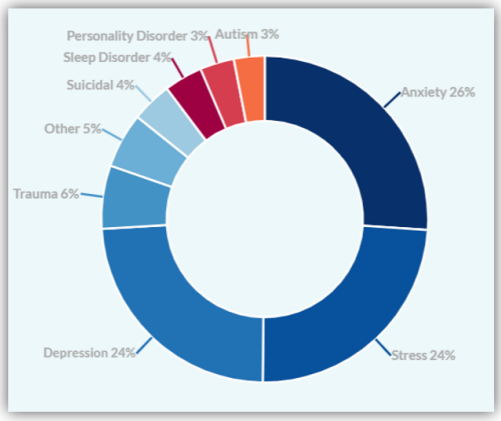

Figure 1: Mental Health Issues (18 Apr)

3.1 Why are People discussing Mental Health Illness on Social Media?

Based on Carl Gustav Jung's Psychological Types and the Communication model of Schulz von Thun, we have developed the personality and communication style identification model which allows to identify the communication purpose.

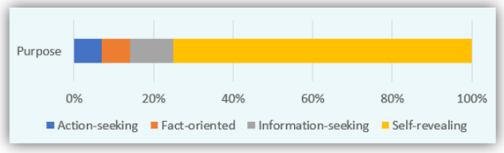

Figure 2: Communication Purpose

According to Figure 2, most of the users share their own experience and struggles (*self-revealing*), followed by users seeking for support and tips (*information-seeking*), participating in general discussions to obtain information and support (*fact-oriented*) and asking for attention, awareness campaigns (*action-seeking*).

3.2 A different perspective to look at Depression

Depression poses greater potential threat to our mental health than it may appear. Of the general conversation around Covid-19, less than 1% mentions explicitly "*depression*" and related terms. However, by applying our models trained on conversations from people diagnosed with depression, we find that up to 10% of the conversations show linguistic signs of depression[7].

3.3 How is Covid-19 affecting young people's mental state?

Over 35% of tweets about mental health struggles in Covid-19 discussion are from younger adults (18-24 yrs). Alarmingly, *Suicide* is mentioned more frequently (over 50%) by younger people than the entire older age group. Young people are talking about *Anxiety* as frequently as the older generations. *Stress* and *Depression* on the other hand, are significantly more prevailing among older age groups.

3.4 How is Covid-19 affecting different genders?

Around 60% of tweets about mental health struggles in Covid-19 discussion come from women. *Anxiety*, *Stress* and *Sleeping disorder* are much more prevailing among women

[5]Due to space limitatons, in the next sections, we use the English set to present the insights.

[6]Look at Figure 1 as an example for day 18 of April

[7]To notice that the prediction reflects the tendency and is by no means diagnostic.

than men. *Self-harm* and *Personality disorder* are more frequently mentioned by men than women.

3.5 What factors are triggering mental health issues during the pandemic?

Isolation is the biggest trigger across different mental health issues. *Financial issues*, *Job insecurity* are triggering *Suicidal* thoughts more drastically than other triggers. *Burnout*, though triggered by all various factors, is linked most to discussion around workload.

3.6 The Supporting of Mental Health

Appointment cancellation due to Covid-19 has lead to mounting health issue challenge. Using digital psychotherapy as alternative is so far creating a mixed sentiment, especially around anxiety.

For example, a user says *"My therapist wants to do the telehealth thing , but it actually gives me more anxiety to talk over video chat"* while another says *"I am SO thankful telemedicine is a thing . Being able to continue my weekly sessions with my therapist during the stay at home order has helped me keep my anxiety mostly under control over the past few weeks"*.

Though India and South Africa have higher volume of general mental health discussion, yet they are lagging behind in the conversation around digital psychotherapy. In other developing countries such as Philippines, Nigeria and Malaysia, there has not been much discussion around teletherapy.

4 Society Resilience

We are also monitoring the major topics of discussion to gain more insights on what is drawing our attention or keeping us occupied, in order to identify the main factors that are impacting mental health.

4.1 Coping in this Challenging Time

People of different cultures are engaged in various activities, some contributing to better mental health, and some less. *Playing* and *Watching TV* are equally frequently mentioned across different countries. In US and Canada, *Socialising* makes up to the most important activities in their daily life. *Exercising* has gained more conversations in countries such as UK and Germany. *Working* is more frequently mentioned by Germans than in any other country.

4.2 How is the Mood in the Discussion around Coronavirus?

We detect the emotions expressed in the tweets and calculate the Mood Index following Equation 1: the ratio of the number of tweets conveying positive emotions and the number of tweets conveying any emotion.

$$MoodIndex = 100 \cdot \frac{\#positive}{\#positive + \#negative} \quad (1)$$

We use the Mood Index to gauge the positivity and negativity of people. The more positive the mood, the higher the index and vice versa when the mood is more pessimistic. In Figure 3, we show an example of the Mood Index evolution in the range of a week.

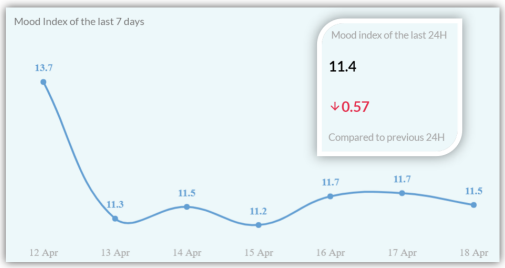

Figure 3: Mood Index (12 Apr-18 Apr)

As shown in Figure 4, African countries are showing a higher Mood Index than European and Asian countries.

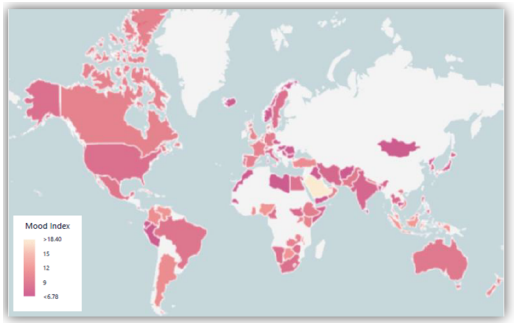

Figure 4: Mood Index around the World

Figure 5 shows the Mood Index in some countries. You can see that the negativity prevails all around the world; in these examples, the index ranges from less than 10 to around 20.

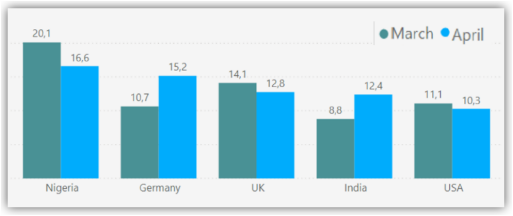

Figure 5: Mood Index in some Countries in the different Continents

4.3 Individualisctiv vs. Collectivistic attitudes

Collectivists attitude can be crucial in pandemic especially as conformity and self-sacrifice are key to the success of regulation and virus containment. It can also greatly impact our mental well-being state, especially when isolation and loneliness are one of the greatest challenges to our mental health.

Our preliminary results, although inconclusive, point out that individualistic cultures are more emotional, while collectivists are more information-oriented as they want to take advantage of rational information for the sake of "we".

5 Insights

The vast social media conversations have opened up a unique opportunity to understand and monitor the impact of Covid-19 on our mental health on a global scale. Our live Mental Health Trackers have raised the awareness around this topic and offered a first data-driven glimpse into the world of mood, emotions and struggles in this crisis time. Here, we summarise some of the key findings of this study:

- Nearly 6% of mental health issue discussion are related to Covid-19.
- Suicide is mentioned more frequently by younger people between 18-24 yrs than the entire older population.
- Digital Psychotherapy still out of reach for most in developing countries.
- Less prevailing issues such as Eating Disorder, Personality Disorder, and Self-harm can also be related to Covid-19.
- African countries are showing a higher Mood Index than European and Asian country.
- Depression poses great potential threat to our mental health and often understated.

6 Beyond Insights

The mental health of individuals and collectives is affected by physical threats such as the Coronavirus, as well as earthquakes or terrorist attacks. As an example, we can see in Figure 6 how the Mood Index correlated negatively (r=-0.831) to the rate of change in new Coronavirus cases in Spain for nine days.

Figure 6: Pearson Correlation between the rate of change in new Coronavirus cases in Spain and the Mood Index in a 9 days time span

Furthermore, mental health can also affect our daily life. Fear of the unknown, such as losing your job, affects the economy. If politicians capitalise fear, instead of working for the social good; if they seize the population's affiliation feeling by spreading fake news and misinformation, fostering hate; if they use technology to monitor and censor our opinions; they will only further polarise society. And this might raise civil conflicts, or even worse.

We plan to navigate through the pandemic and post-pandemic era, and we will also decode the mood development through the lens of psychology aiming to provide an indicator of the social dynamics as a whole. A holistic approach should give us the proper instruments to fight against these threats and to work together for the social good.

Acknowledgements

We want to warmly thank all people at Symanto and Atribus for their hard work and for making possible this project.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 135-138 recibido 29-04-2020 revisado 26-05-2020 aceptado 07-06-2020

AREVA: Augmented Reality Voice Assistant for Industrial Maintenance

AREVA: Asistente por Voz Dotado de Realidad Aumentada para el Mantenimiento Industrial

Manex Serras[1], **Laura García-Sardiña**[1], **Bruno Simões**[1], **Hugo Álvarez** [1], **Jon Arambarri** [2]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]VirtualWare Labs Foundation
{mserras, lgarcias, bsimoes, halvarez}@vicomtech.org, jarambarri@virtuawareco.com

Abstract: Within the context of Industry 4.0, AREVA is presented: a Voice Assistant with Augmented Reality visualisations for the support and guidance of operators when carrying out tasks and processes in industrial environments. With the aim of validating its use for the training of new operators, first evaluations were performed by a group of non-expert users who were asked to carry out a maintenance task on a Universal Robot.
Keywords: Spoken Dialogue System, Augmented Reality, Industry 4.0

Resumen: Dentro del marco de la Industria 4.0 se presenta AREVA: un Asistente por Voz dotado con visualizaciones de Realidad Aumentada para el apoyo y guiado del operario a la hora de realizar tareas y procesos en entornos industriales. Con el objetivo de validar su uso para el entrenamiento y capacitación de nuevos operarios, se ha realizado una primera evaluación con un grupo de usuarios no-expertos a quienes se les pidió que realizasen una tarea de mantenimiento sobre un Robot Universal.
Palabras clave: Sistema de Diálogo Hablado, Realidad Aumentada, Industria 4.0

1 *Introduction*

Industry 4.0 is one of the main paradigms of the latest years, where the industrial factories are augmented with wireless connectivity, sensors, and AI mechanisms that can help with different processes and tasks. Industrial 4.0 revolution has shown how technology can alter the way work is performed, the structure of organisations, and the role that workers play in the manufacturing process (Simões et al., 2019; Posada et al., 2018; Serras et al., 2020).

This paper describes an Augmented REality Voice Assistant (AREVA) to assist operators during the maintenance of Universal Robot's (UR) Grippers. To this end, the system guides the operator combining voice interaction and real-time 3D visualisations. Augmented Reality (AR) glasses with a built-in microphone are used to provide hands-free interaction and as a solution to overcome practical limitations in tasks that rely on intense manual work.

This paper presents the AREVA system for industrial maintenance tasks. The presented prototype system was tested by a set of 10 participants (Gender: 5 male, 4 female, 1 would rather not say; Age ranges: 50% 25-29, 30% 35-39, 10% 30-34, and 10% 18-24) who had little-to-none technical expertise on the field and no prior knowledge of similar maintenance tasks, using the presented assistant to support and guide them through the process. After the experimental sessions, participants filled an assessment form to judge the system's validity.

Section 2 describes the architecture of the system and the use case, Section 3 details the experimental framework and presents the evaluation results. Finally 4 presents the conclusions and sets the future work.

2 *System Architecture*

The system architecture follows the classical SDS schema, but incorporating communication channels with the necessary hard-

ISSN 1135-5948. DOI 10.26342/2020-65-21

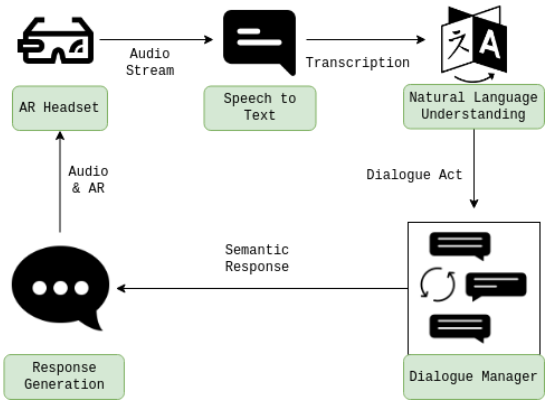

Figure 1: Architecture of the Augmented Reality enhanced Spoken Dialogue System

ware devices to allow for both visual and aural communication with operators, as shown in Figure 1. The core components are presented in more detail below.

- **AR Headset**: this device is responsible of capturing the users' speech, play AREVA's responses, and render AR animations. The selection of a head-mounted device is motivated so that operators have theirs hands free to perform the task while interacting with the system. For the proposed system, Microsoft HoloLens were used.
- **Speech to Text (STT)**: is the component in charge of converting the audio streams into texts. First, an energy-based automata is used as Voice Activity Detector (VAD) to segment audio streams into speech chunks, which are then sent to a speech transcription service to get the text utterance. Transcriptions were provided by the Google Speech API.
- **Natural Language Understanding (NLU)**: once the voice segments have been decoded into text, the NLU extracts the communicative intention of the transcribed text so it can be interpretable by the system. For the AREVA system, the grammar-based Phoenix parser was used (Ward, 1991).
- **Dialogue Manager (DM)**: the DM is the module that ensures the correct planning and interaction through the industrial process. The DM implemented for AREVA consists of an Attributed Probabilistic Finite State Bi-Automata (Torres, 2013; Serras, Torres, and Del Pozo, 2019) which uses a stack of expert-rules and simulated interactions to guide the operator through the task.
- **Response Generation**: this module receives the response selected by the DM to carry out with the interaction. It performs two actions: 1) it selects an adequate text template for the response which is then synthesised by a Text To Speech service (Hernaez et al., 2001), and 2) it checks if the action has an associated AR rendering. Then, these audio and visual responses are sent to the AR Headset to be conveyed to the operator.

2.1 Use Case

We considered as a use case the maintenance of the Universal Robot's gripper. The objective of the system is to assist untrained operators in real time with contextualised responses to the questions that may arise during the course of the task. AR is used to overlay animations that further enhance the experience.

The maintenance steps considered for the use case were retrieved from Subsections 7.1 and 7.3 of the official maintenance manual[1], which require the following actions:

1. Shutdown the UR, unplug the connected wires, and dismount the gripper.
2. Open the gripper to clean it and inspect it both visually and by moving its fingers to detect any wear, damages or incorrect joint articulations.
3. Close the gripper's fingers, mount it back to the UR, and attach the wires.

Once these four stages are completed, the monthly maintenance of the UR's gripper is finished. As it can be implied by the steps to perform the task, a heavy manual effort is required to perform the operation, so a hands-free system is critical to improve the User Experience (UX).

For the presented use case, the developed NLU module can understand a total of 10

[1] https://assets.robotiq.com/website-assets/support_documents/document/3-Finger_PDF_20190322.pdf?_ga=2.105997637.750885948.1563187207-1298495031.1563187207, last accessed 29/04/2020

Augmented Reality & Spoken Dialogue System	
Visualisations aside, how helpful has the voice interaction been in performing maintenance?	Mean score: 4,3/5
Voice interaction aside, how useful have the visualisations been to you?	Mean score: 3,7/5
Do you think you could have done the maintenance properly without the visualisations, just with the voice interaction?	Yes: 1, Yes, but it would have taken longer: 4, No, AR was a key feature: 4, I don't know: 1
Do you think you could have done the maintenance properly without the voice interaction, just with the visualisations?	Yes, but it would have taken longer: 4 No, voice interaction was a key feature: 6
If you had to say which technology was more relevant, which would it be?	AR: 1, SDS: 3, Both: 6
Do you think it was useful to combine both technologies?	Yes: 10, No: 0
Usability and perception	
Has the system made maintenance easier?	Mean score: 4/5
Have you felt frustrated at any point in the process?	Never: 1, Almost never: 4, Sometimes: 5
"The system has increased my confidence that I was performing the task correctly"	Mean agreement score: 3,8/5
How much has having your hands free made maintenance easier?	Mean score: 4,5/5
Task Completion and Recommendation	
Have you been able to complete the maintenance task?	Yes: 9, No: 1
Would you recommend using the assistant to someone who has never done the maintenance before?	Yes: 10, No: 0
Would you use the assistant to do maintenance a second time, the following month?	Yes: 8, No: 2

Table 1: Summary of the evaluation form filled by the participants

different intents, 20 entities, and more than 20 entity-values in total, using more than 2000 grammar rules to parse the operators' input into valid semantic actions.

The initial DM version was trained using 33 expert-rules. These DM rules map the dialogue state –which encodes the latest system's response, user action, and discrete memory attributes– with the possible actions to perform in the task, with physical objects (e.g. tools, wires, bolts), and with their properties, so that questions like "Which tool do I need?" and "Where are the bolts?" can be disambiguated according to the current step context. These rules also control the dialogue navigation aspects such as the welcoming prompt, repetitions, flow constraints (do not allow the operator to go to the next step if the current one is not satisfied), and so on.

The Response Generation had 48 possible response actions, amounting to 131 text templates (an action may have more than one associated template). A total of 8 possible AR animations were developed: 1) initial gripper position over the UR, 2) point shutdown button, 3) point the wires location, 4) point the bolts in the base of the gripper, 5) point the bolts under the gripper's fingers, 6) gripper's fingers opening, 7) gripper's fingers movement to check, and finally 8) gripper's fingers closing.

3 *Experimental Framework*

The objective of AREVA is to help and guide operators during maintenance tasks and make them more independent. Ten participants were recruited to validate the system and asked to perform a maintenance task in which they had no previous experience.

Figure 2: An operator using the AREVA system to carry out the gripper's maintenance

The robot arm was fixed on a pre-set position, with the gripper coupled and its fingers closed, as shown in Figure 2. The tools needed for the task were placed on the table. At the beginning of the session, while the participants were adjusting and getting familiarised with the AR Glasses, a researcher would present the UR, the system, and the task at hand that they should complete with no additional details.

The evaluation form had several questions regarding the technological components of the system, which were evaluated both separately and in conjunction. Besides,

additional questions about usability, participants' perception, and task completion were included. The results to these questions are summarised in Table 2.

The results reported in Table 2 show that the use of AR and SDS technologies combined in AREVA is useful for training and assisting applications for industrial processes.

The Spoken Dialogue System is perceived more positively by the participants, both in helpfulness and as a key feature, but, overall, the combination of both technologies is perceived as advantageous and both of them are relevant to assist and guide the operator through the gripper's maintenance task. Regarding the scores achieved in the system's usability and perception evaluation, all participants agreed that the system made the maintenance easier, especially by allowing them to have both hands free, and made them feel more confident that they were doing it correctly. Still, there were times when some users felt frustrated using the system, mostly due to STT or Voice Activity Detection errors.

Finally, 9 out of 10 users were able to complete the maintenance –note that no participant had made this process before and their knowledge of industrial processes was limited–. All of them would recommend the system to someone who has never done this task before, proving the usefulness of the system when training new operators. Moreover, 8 of them would use AREVA again for the following month's maintenance.

Additional observations made by the participants set the focus on improving the hardware ergonomics, as sometimes it would bother the participants when performing the maintenance. Also, the capability of adapting to each user profile (beginners/intermediate/advanced) to adjust the response verbosity was deemed as a useful feature.

4 Conclusions and Future Work

This paper describes an Augmented Reality Voice Assistant for natural and hands-free guiding for a Universal Robot Gripper's maintenance. The proposed solution facilitates the capture, distribution, and communication of domain specific knowledge to operators in training and production phases. The hands-free, technology-combining multimodal system was well accepted by a group of non-expert operators unfamiliar to the task. As future work, a wider sample of participants will participate as testers. In addition, UX factors will be further improved and evaluated, such as improvements in the STT and NLU modules and the use of Wake-up words to avoid ambient-noise activation of the VAD module. Also, the extrapolation to other industrial processes will be further investigated.

References

Hernaez, I., E. Navas, J. L. Murugarren, and B. Etxebarria. 2001. Description of the ahotts system for the basque language. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

Posada, J., M. Zorrilla, A. Dominguez, B. Simoes, P. Eisert, D. Stricker, J. Rambach, J. Döllner, and M. Guevara. 2018. Graphics and media technologies for operators in industry 4.0. *IEEE computer graphics and applications*, 38(5):119–132.

Serras, M., L. García-Sardiña, B. Simões, H. Álvarez, and J. Arambarri. 2020. Dialogue enhanced extended reality: Interactive system for the operator 4.0. *Applied Sciences*, 10(11):3960.

Serras, M., M. I. Torres, and A. Del Pozo. 2019. User-aware dialogue management policies over attributed bi-automata. *Pattern Analysis and Applications*, 22(4):1319–1330.

Simões, B., R. De Amicis, I. Barandiaran, and J. Posada. 2019. Cross reality to enhance worker cognition in industrial assembly operations. *The International Journal of Advanced Manufacturing Technology*, pages 1–14.

Torres, M. I. 2013. Stochastic bi-languages to model dialogs. In *Proceedings of the 11th international conference on finite state methods and natural language processing*, pages 9–17.

Ward, W. 1991. Understanding spontaneous speech: The phoenix system. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 365–367. IEEE.

Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 139-142 recibido 25-03-2020 revisado 30-05-2020 aceptado 30-05-2020

UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks

UMUCorpusClassifier: Recolección y evaluación de corpus lingüísticos para tareas de Procesamiento del Lenguaje Natural

José Antonio García-Díaz[1], **Ángela Almela**[2],
Gema Alcaraz-Mármol[3], **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia
[2]Facultad de Letras, Universidad de Murcia
[3]Departamento de Filología Moderna, Universidad de Castilla-La Mancha
{joseantonio.garcia8, angelalm, valencia}@um.es
gema.alcaraz@uclm.es

Abstract: The development of an annotated corpus is a very time-consuming task. Although some researchers have proposed the automatic annotation of a corpus based on ad-hoc heuristics, valid hypotheses cannot always be made. Even when the annotation process is performed by human annotators, the quality of the corpus is heavily influenced by disagreements between annotators or with themselves. Therefore, the lack of supervision of the annotation process can lead to poor quality corpus. In this work, we propose a demonstration of UMUCorpusClassifier, a NLP tool for aid researches for compiling corpus as well as coordinating and supervising the annotation process. This tool eases the daily supervision process and permits to detect deviations and inconsistencies during early stages of the annotation process.
Keywords: Corpus compilation, Document classification

Resumen: La construcción de un corpus anotado es una tarea que consume mucho tiempo. Aunque algunos investigadores han propuesto la anotación automática basada en heurísticas, éstas no siempre son posibles. Además, incluso cuando la anotación es realizada por personas puede haber discrepancias entre los mismos anotadores o de un anotador consigo mismo que influyen en la calidad del corpus. Por tanto, la falta de supervisión sobre el proceso de anotación puede llevar a corpus con baja calidad. En este trabajo, proponemos una demostración de UMUCorpusClassifier, una herramienta PLN para ayudar a los investigadores a compilar corpus y también a coordinar y supervisar el proceso de anotación. Esta herramienta facilita la monitorización diaria y permite detectar inconsistencias durante etapas tempranas del proceso de anotación.
Palabras clave: Compilación de corpus, Clasificación de documentos

1 Introduction

Supervised learning is the machine learning task which consists of building a model capable of predicting output for a specific problem based on prior observation of a previously labeled data set (Singh, Thakur, and Sharma, 2016). Supervised learning has applications for solving document classification tasks, which consist of matching a set of documents with a set of predefined labels. The main idea behind this is that supervised learning models can infer new knowledge by establishing associations between the examples provided and the expected tags. However, supervised learning requires a sufficient number of labeled examples that model the problem domain and, at the same time, the number of examples should be enough to cluster the examples in two subsets, one for model learning, and another for evaluating

ISSN 1135-5948. DOI 10.26342/2020-65-22

its accuracy based on samples that are not seen during the training stage.

The development of an annotated corpus is a very time-consuming process. To facilitate this task, some researchers have used distant supervision as a method of getting automatically annotated data (Go, Bhayani, and Huang, 2009). Distant supervision consists in the automatic tagging of the whole dataset based on certain assumptions or heuristics; however, valid hypotheses cannot always be made. Furthermore, automatic annotated data that have not thoroughly been reviewed can lead to noise in the data, which introduces information that do not follow the relationship with the rest of the elements of the dataset. These noisy documents require even greater volume of data to minimize its effects.

Furthermore, even though the annotation process occurs manually, the quality of the corpus cannot be guaranteed. In this sense, Mozetič, Igor et al. (Mozetič, Grčar, and Smailović, 2016) conducted an experiment to measure the quantity and quality of tagged Twitter datasets to evaluate the impact of measuring accuracy in sentiment analysis-based classification experiments, and they found that not all corpus annotated manually had a strong degree of agreement among the annotators, which indicated poor-quality. To solve these drawbacks, we present UMUCorpusClassifier, a tool that assists researchers in compiling text corpus, and helps to coordinate the annotation process.

The remainder of this paper is organized as follows: Section 2 describes our proposal, and Section 3 contains information regarding studies based on corpus compiled with this tool, as well as the description of future lines of action.

2 System architecture

UMUCorpusClassifier is designed to work with the social network Twitter, which is a widely used network for compiling corpus in various branches of Natural Language Processing (NLP) (Pak and Paroubek, 2010). Twitter provides a developer API through which you can query the social network. The free version of the Twitter API has some restrictions regarding the number of calls that can be made in a time interval. Therefore, to create an account in the platform, users are required to link a Twitter API key with their account. These credentials are generated through the Twitter API website. Once users have created and validated their account, they can create a corpus by entering a search string and, optionally, a geolocation. The search string has the same flexibility and limitations as the Twitter API; that is, it is allowed to nest terms, do exact searches, or search for specific user accounts. A list of these operators can be found here[1]. In addition, users can specify that only user-initiated messages are obtained by filtering messages that are responses to other users' tweets. Once the search query is specified, the corpus are compiled automatically. From time to time, a scheduled cron job process queries new tweets based on keywords. UMUCorpusClassifier works effectively with timelines, making use of *since_id* parameters to get new tweets and avoid duplicate results. However, this feature can be omitted in cases where the user deliberately changes the search string.

Identifying duplicate tweets is a complex task. Although Twitter provides a mechanism called *retweet* that allows the content of messages written by other users to be disseminated and the identification of these messages is trivial, many users in the social network use copy and paste mechanisms so it is possible to find duplicate or virtually the same tweets. Also, hyperlinks in tweets are encoded differently with each new tweet due to Twitter's own hyperlink shortening mechanism. For this reason, we have made the decision to replace URLs with a fixed token, which makes it easier to identify certain tweets. In addition, we have added a mechanism to calculate the similarity of the texts. Being an experimental technology, tweets are not removed, but administrators are given the opportunity to combine the responses of the tweets at their discretion.

The next step is to assign each corpus a set of independent labels. They can be made using a set of predefined labels, such as *out-of-domain*, *positive*, *negative*, *neutral*, *do-not-know-do-not-answer* or define a new set of tags for the corpus. Each label is identified with a color and a name.

The corpus labeling process can be carried out manually by the same user or allow access to the platform to a set of annotators and to supervise their work. Documents are

[1] https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/standard-operators

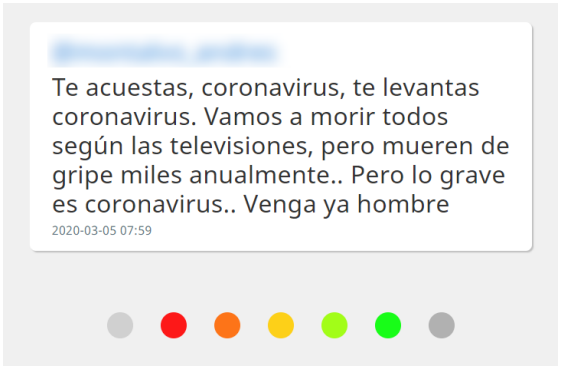

Figure 1: Screenshot of an annotator's view

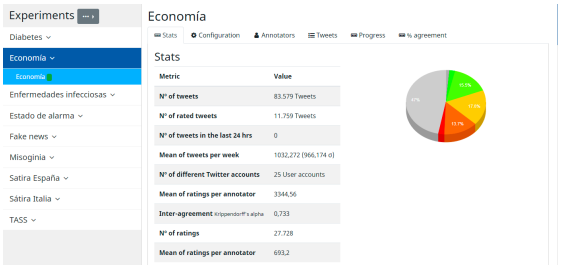

Figure 2: Screenshot of a researcher's view

annotated individually. When an annotator enters the web application, a tweet appears randomly from tweets that they had not previously classified. Figure 1 shows the view of an annotator at the platform.

To facilitate the labeling monitoring process, UMUCorpusClassifier provides a set of metrics and charts, such as the evolution of the annotations made by time, to evaluate that the work is being done constantly; the total rankings by tag; or (3) the average of annotations as well as their standard deviation. Figure 2 shows the researcher's view of the platform.

For each annotator, the degree of *self-agreement* is calculated, which measures how the same annotator classifies semantically similar documents. To cluster similar documents, we obtain the *sentence-embeddings* from FastText in its Spanish version (Grave et al., 2018) and then we have calculated the cosine distance among those vectors. Tweets that exceed a certain threshold distance are discarded.

For each tweet, one can see the degree of *inter-agreement* that measures how the same tweet is classified by different annotators by using the Krippendorff's alpha coefficient (Krippendorff, 2018). Moreover, inter-agreement has been proposed as upper bound

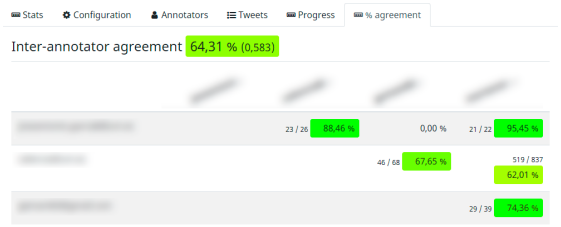

Figure 3: Screenshot of an annotator's view

to estimate the expected *accuracy* when finding a classifier to solve the classification problem (Mozetič, Grčar, and Smailović, 2016). Figure 3 shows the overall inter-agreement of the corpus, as well as the inter-agreement for each pair of annotators.

Once the corpus has been compiled and annotated, it can be exported to text formats. The export process is flexible and allows you to choose the number of classes to export. Combining classes is also allowed. This is useful, for example, when a classification has been made on a scale of very negative, negative, neutral, positive, and very positive type values, but one may want to combine the results to return the corpus grouped into positive, neutral, and negative. Corpora can be exported balanced, that is to say, the system automatically searches for the class with the largest number of instances and cuts instances from the other classes. These deleted instances are agreed by consensus. Finally, it is possible to export only the Twitter IDs in order to share them with the community as recommended by Twitter's privacy policies[2].

Furthermore, the number of instances to export can be selected and set. One of the advantages of this approach is that corpus can be exported by consensus: since the same tweet can be classified by different annotators, the number of tweets to export can be limited and retrieve those tweets that have achieved strong consensus among annotators. Thus, subsets of the corpus comprising the documents with common agreement can be retrieved, and the rest of the documents can be analyzed. Furthermore, the software is easily extensible. In this respect, it is relatively easy to include new strategies to export the data or to improve the platform to include new data sources other than Twitter.

[2]https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

3 Further work

In this study, we have presented UMUCorpusClassifier, a NLP tool that assists in the compilation and annotation of linguistic corpus. So far, we have used this application on several domains. Specifically, we have compiled tweets about different types of diseases to carry out infodemiology studies that involve measuring the population's perception of infectious diseases (Apolinardo-Arzube et al., 2019; García-Díaz, Cánovas-García, and Valencia-García, 2020; Medina-Moreira et al., 2018), diabetes (Medina-Moreira et al., 2019), and figurative language such as satire (Salas-Zárate et al., 2017).

In the current version of the platform it is only possible to assign a label to a document. We are working to enable the *multi-label* classification. Another line of research is the addition of a contextual feature extraction module, enabling the analysis of groups of Twitter accounts from which tweets are extracted. These features may include information on the time of publication, number of followers, etc. Lastly, with regard to semantic similarity, we are currently analyzing ways to distance the most different tweets from each other, so that we can export the tweets with strongest consensus and the most distant ones.

Acknowledgments

This demonstration has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the Doctorado industrial programme.

References

Apolinardo-Arzube, O., J. A. García-Díaz, J. Medina-Moreira, H. Luna-Aveiga, and R. Valencia-García. 2019. Evaluating information-retrieval models and machine-learning classifiers for measuring the social perception towards infectious diseases. *Applied Sciences*, 9(14):2858.

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:614–657.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Medina-Moreira, J., J. A. García-Díaz, O. Apolinardo-Arzube, H. Luna-Aveiga, and R. Valencia-García. 2019. Mining twitter for measuring social perception towards diabetes and obesity in central america. In *International Conference on Technologies and Innovation*, pages 81–94. Springer.

Medina-Moreira, J., J. O. Salavarria-Melo, K. Lagos-Ortiz, H. Luna-Aveiga, and R. Valencia-García. 2018. Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. In *Proceedings of the Technologies and Innovation: 4th International Conference, CITI*, page 229. Springer.

Mozetič, I., M. Grčar, and J. Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5).

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Salas-Zárate, M. d. P., M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, and G. Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowl. Based Syst.*, 128:20–33.

Singh, A., N. Thakur, and A. Sharma. 2016. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee.

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 Maíllo
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:

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

Demostraciones

EmoCon: Analizador de emociones en el congreso de los diputados
Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, María Teresa Martín-Valdivia, Andrea López-Fernández 115
Nalytics: natural speech and text analytics
Ander González-Docasal, Naiara Pérez, Aitor Álvarez, Manex Serras, Laura García-Sardiña, Haritz Arzelus, Aitor García-Pablos, Montse Cuadros, Paz Delgado, Ane Lazpiur, Blanca Romero 119
RESIVOZ: dialogue system for voice-based information registration in eldercare
Laura García-Sardiña, Manex Serras, Arantza del Pozo, Mikel D. Fernández-Bhogal 123
The text complexity library
Rocío López Anguita, Jaime Collado Montañez, Arturo Montejo Ráez 127
The impact of coronavirus on our mental health
Jiawen Wu, Francisco Rangel, Juan Carlos Martínez 131
AREVA: augmented reality voice assistant for industrial maintenance
Manex Serras, Laura García-Sardiña, Bruno Simões, Hugo Álvarez, Jon Arambarri 135
UMUCorpusClassifier: compilation and evaluation of linguistic corpus for natural language processing tasks
José Antonio García-Díaz, Ángela Almela, Gema Alcaraz-Mármol, Rafael Valencia-García 139

Información General

Información para los autores 145
Información adicional 147