Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Artículos

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universitat Pompeu Fabra
Año de edición: 2014

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

Consejo asesor

M. Antonia Martí	Universitat de Barcelona
M. Teresa Martín	Universidad de Jaén
Patricio Martínez-Barco	Universitat d'Alacant
Raquel Martínez	UNED
Lidia Moreno	Universitat Politècnica de València
Lluís Padro	Universitat Politècnica de Catalunya
Manuel Palomar	Universitat d'Alacant
Ferrán Pla	Universitat Politècnica de València
German Rigau	Euskal Herriko Unibertsitatea
Horacio Rodríguez	Universitat Politècnica de Catalunya
Emilio Sanchís	Universitat Politècnica de València
Kepa Sarasola	Euskal Herriko Unibertsitatea
Mariona Taulé	Universitat de Barcelona
L. Alfonso Ureña	Universidad de Jaén
Felisa Verdejo	UNED
Manuel Vilares	Universidad de A Coruña
Ruslan Mitkov	Universidad de Wolverhampton, UK
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues, France
Leonel Ruiz Miyares	Centro de Linguística Aplicada de Santiago de Cuba
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Alexander Gelbukh	Instituto Politécnico Nacional, México
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores, Portugal
Bernardo Magnini	Fondazione Bruno Kessler, Italia

Revisores adicionales

José Mariño Acebal	Universitat Politècnica de Catalunya
Rodrigo Agerri	Universidad del País Vasco
Laura Alonso	Universidad Nacional de Córdoba
Enrique Amigó	UNED
Alberto Barrón-Cedeño	Universitat Politècnica de Catalunya
Núria Bel	Universitat Pompeu Fabra
Luciana Benotti	Universidad Nacional de Córdoba
Stefan Bott	Universidad de Stuttgart, Alemania
Zoraida Callejas	Universidad de Granada
Jorge Carrillo-de-Albornoz	UNED
Juan Miguel Cigarrán	UNED
Joan Codina	Universitat Pompeu Fabra
Jesús Contreras	iSOCO
Victor Darriba	Universidad de Vigo
Iria Da Cunha	Universitat Pompeu Fabra
Manuel de Buenaga	Universidad Europea de Madrid
Adrià de Gispert	University of Cambridge
César de Pablo Sánchez	Universidad Carlos III
Alberto Díaz	Universidad Complutense de Madrid
Víctor J. Díaz Madrigal	Universidad de Sevilla
Mireia Farrús	Universitat Pompeu Fabra
Gabriela Ferraro	NICTA, Australia
Miguel Ángel García Cumbreras	Universidad de Jaén
Pablo Gervás	Universidad Complutense de Madrid
Carlos Gómez	Universidade da Coruña

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 53 de la revista de la Sociedad Española para el Procesamiento del Lenguaje Natural contiene trabajos correspondientes a tres apartados diferenciados:

comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de herramientas. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista que ha sido llevado a cabo según el calendario previsto. Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 47 trabajos para este número de los cuales 29 eran artículos científicos y 18 correspondían a resúmenes de proyectos de investigación y descripciones de herramientas. De entre los 29 artículos recibidos 14 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 48,27%. Autores de otros 7 países han participado en los trabajos publicados en la revistas. Estos países son: Alemania, Catar, EEUU, Irlanda, Reino Unido, Rusia y Singapur.

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.

Estimamos que la calidad de los artículos es alta. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando haya sido igual o superior a 5 sobre 7.

Septiembre de 2014
Los editores

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of 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 the 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 53rd issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, investigation projects and tools descriptions summaries. All of these were accepted by the traditional peer reviewed process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Forty-seven papers were submitted for this issue, from which twenty-nine were scientific papers and eighteen were either projects or tool description summaries. From these twenty-nine scientific papers, we selected fourteen (48.27%) for publication. Authors from other seven countries have submitted papers to the journal. These countries are: Germany, Qatar, USA, Ireland, United Kingdom, Russia and Singapore.

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 to those papers with a difference of three or more points out of 7 in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit was supervised by the editorial board.

We consider that the quality of the articles is high. The cut-off criteria adopted was the average of the three scores given, as long as this has been equal to or higher than 5 out of 7.

September 2014
Editorial board

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Artículos

Artículos

Extracción de Información

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 17-24 recibido 14-04-14 revisado 02-06-14 aceptado 10-06-14

Análisis morfosintáctico y clasificación de entidades nombradas en un entorno *Big Data* *

PoS tagging and Named Entitiy Recognition in a Big Data environment

Pablo Gamallo, Juan Carlos Pichel, Marcos Garcia
Centro de Investigación en Tecnoloxías da Información (CITIUS)
{*pablo.gamallo, juancarlos.pichel, marcos.garcia.gonzalez*}*@usc.es*

José Manuel Abuín y **Tomás Fernández-Pena**
Centro de Investigación en Tecnoloxías da Información (CITIUS)
{*tf.pena, josemanuel.abuin*}*@usc.es*

Resumen: Este artículo describe una suite de módulos lingüísticos para el castellano, basado en una arquitectura en *tuberías*, que incluye tareas de análisis morfosintáctico así como de reconocimiento y clasificación de entidades nombradas. Se han aplicado técnicas de paralelización en un entorno *Big Data* para conseguir que la suite de módulos sea más eficiente y escalable y, de este modo, reducir de forma signiticativa los tiempos de cómputo con los que poder abordar problemas a la escala de la Web. Los módulos han sido desarrollados con técnicas básicas para facilitar su integración en entornos distribuidos, con un rendimiento próximo al estado del arte.
Palabras clave: Análisis morfosintáctico, Reconocimiento y clasificación de entidades nombradas, Big Data, Computación Paralela

Abstract: This article describes a suite of linguistic modules for the Spanish language based on a pipeline architecture, which contains tasks for PoS tagging and Named Entity Recognition and Classification (NERC). We have applied run-time parallelization techniques in a Big Data environment in order to make the suite of modules more efficient and scalable, and thereby to reduce computation time in a significant way. Therefore, we can address problems at Web scale. The linguistic modules have been developed using basic NLP techniques in order to easily integrate them in distributed computing environments. The qualitative performance of the modules is close the the state of the art.
Keywords: PoS tagging, Named Entity Recognition, Big Data, Parallel Computing

1 Introducción

En la sociedad digital moderna, se estima que cada día creamos alrededor de 2,5 trillones de bytes de datos (1 exabyte), de tal forma que el 90 % de los datos en todo el mundo han sido creados en los últimos dos años[1]. Así por ejemplo, Twitter genera unos 8 terabytes de datos al día, mientras que Facebook captura unos 100 terabytes[2]. Una de las principales características de esta información es que, en muchos casos, se encuentra sin estructurar, ya que se trata en un porcentaje alto de información textual. Dado el ingente volumen de información textual generado a diario, se hace cada vez más necesario que el procesamiento y análisis lingüístico de esta información se efectúe de manera eficiente y escalable, lo que provoca que las tareas de PLN requieran de soluciones paralelas. Por consiguiente, el uso de la Computación de Altas Prestaciones (HPC) y de su derivación en el paradigma *Big Data*, se hace indispensable para reducir de forma notable los tiempos de cómputo y mejorar la escalabilidad de los sistemas de PLN. En este mismo sentido, cabe reseñar que la filosofía de los enfoques más recientes de la lingüística de corpus se basan en la *Web As Corpus*, línea de investigación donde se postula que con más datos y más

* Este trabajo ha sido subvencionado con cargo a los proyectos HPCPLN - Ref:EM13/041 (Programa Emergentes, Xunta de Galicia), Celtic - Ref:2012-CE138 y Plastic - Ref:2013-CE298 (Programa Feder-Innterconecta)

[1] IBM, Big Data at the Speed of Business: http://www-01.ibm.com/software/data/bigdata/

[2] http://hadoopilluminated.com/-hadoop_illuminated/Big_Data.html

ISSN 1135-5948

texto se obtienen mejores resultados (Kilgarriff, 2007). Y para procesar más corpus a la escala de la Web se requieren soluciones HPC.

En este artículo, nuestro principal objetivo es aplicar técnicas Big Data a un conjunto de tareas lingüísticas integradas en una suite de módulos PLN para el castellano y de código abierto, llamada *CitiusTool*, que incluye la etiquetación y desambiguación morfosintáctica (PoS tagging), así como el reconocimiento y clasificación de entidades nombradas (NERC). De esta manera, conseguimos que la suite de módulos de PLN sea más eficiente y escalable permitiendo reducir de manera signiticativa los tiempos de cómputo y abordar problemas de un tamaño aún mayor.

La arquitectura de la suite de módulos se basa en el paradigma de tuberías (o *pipeline*), y cada módulo lingüístico de la suite es una función escrita en Perl. La ventaja de este enfoque es que cada componente está directamente conectado con los otros a través de los tubos (o *pipes*), de tal modo que no es necesario esperar hasta que finalice un proceso antes de comenzar el siguiente de la *pipeline*. A diferencia de las arquitecturas PLN basadas en el flujo de trabajo (*workflow*), como GATE[3] (Tablan et al., 2013) o UIMA[4], en una tubería cuando un módulo comienza a producir algún tipo de salida esta se transmite como entrada del siguiente módulo sin producir ficheros de datos intermediarios. Una suite de módulos similar a la descrita en este artículo, para castellano e inglés, ha sido implementada en el proyecto opeNER (Agerri, Bermudez, y Rigau, 2014), y que ha dado lugar a IXA pipes (Agerri, Bermudez, y Rigau, 2014), cuyos módulos lingüísticos están basados en Java.

La simplicidad de los módulos PLN que se consideran en este trabajo, así como la clara independencia de las unidades lingüisticas de entrada de dicho módulos (frases, párrafos, textos, etc.), son factores que facilitan su integración en una arquitectura para *Big Data* que usa el modelo de programación *MapReduce*. En concreto, se utilizará la herramienta de código abierto *Hadoop*[5] que implementa dicho modelo. Por otro lado, debemos destacar que a pesar de la gran simplicidad de nuestros módulos lingüísticos, la calidad de sus resultados está muy próxima a la ofrecida por los sistemas considerados estado del arte.

Una vez integrados en una plataforma distribuida, nuestros módulos lingüísticos se podrán utilizar en aplicaciones más complejas y de alto nivel que verán así mejorar su eficiencia. Concretamente, las aplicaciones de ingeniería lingüística que pueden beneficiarse de estos módulos son: traducción automática, recuperación de información, búsqueda de respuestas, sistemas inteligentes de vigilancia tecnológica y, en general, sistemas relacionados con el paradigma de la analítica de textos (*text analytics*), tan en voga con el auge de las redes sociales.

El resto del artículo se organiza del siguiente modo. En la siguiente sección (2), se introduce la arquitectura de los módulos. A continuación, en las secciones 3 y 4, se describen los módulos de análisis morfosintáctico y clasificación de entididades, respectivamente. En ambas secciones se describen también experimentos y evaluaciones cualitativas de los módulos. Seguidamente, la sección 5 se centra en los experimentos realizados con la plataforma Hadoop y finalizamos con las conclusiones y trabajo futuro (sección 6).

2 Arquitectura

CitiusTool[6] es una herramienta lingüística de libre distribución (licencia GPL) concebida para ser fácil de usar, instalar y configurar. La figura 1 muestra la arquitectura en tubería de los diferentes módulos de la suite lingüística. La herramienta consta de varios módulos de análisis básico, un reconocedor de entidades (NER), un analizador morfológico (PoS tagger), que incluye un lematizador, y un clasificador de entidades (NEC). Hasta ahora, hemos adaptado los módulos al castellano, aunque estamos trabajando en su adaptación a otras lenguas peninsulares.

3 Herramientas de análisis

3.1 Análisis básico

Como se puede observar en la figura 1, el sistema consta de varios módulos de procesamiento básico del lenguaje:

- Separador de frases: toma en cuenta los puntos finales, líneas en blanco y un fichero externo de siglas y abreviaturas.

[3]https://gate.ac.uk/

[4]https://uima.apache.org/

[5]http://hadoop.apache.org/

[6]Disponible para descarga en: http://gramatica.usc.es/pln/tools/CitiusTools.html.

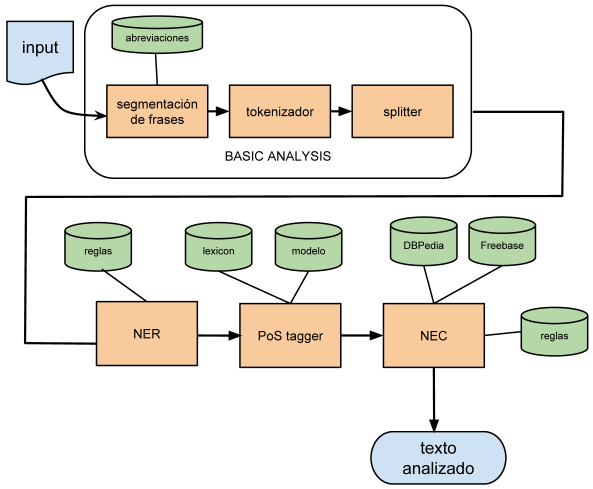


Figura 1: Arquitectura de la suite de módulos lingüísticos CitiusTool

- Tokenizador: separa las frases en tokens.
- Splitter: separa las contracciones en sus correpondientes tokens, por ejemplo, "del" = "de el", "comerlo = "comer lo".

El análisis básico sirve de entrada al Reconocedor de Entidades Nombradas (NER), que aplica un conjunto de reglas para la identificación de nombres propios formados por más de un token. Por ejemplo: *Universidad del País Vasco*, *Real Club Celta de Vigo*, etc. El formato de intercambio entre módulos es similar al de la suite lingüística FreeLing (Padró y Stanilovsky, 2012).

3.2 Análisis morfosintáctico

Hemos desarrollado un desambiguador morfosintáctico o *PoS tagger*, llamado CitiusTagger, utilizando el mismo léxico y etiquetario que FreeLing. Se fundamenta en un clasificador de bigramas bayesiano que asigna la etiqueta más probable a cada token ambiguo tomando en cuenta su contexto a la izquierda y a la derecha. Para desambiguar un token ambiguo, calculamos la probabilidad de cada etiqueta e_i asociada a ese token, dado un conjunto de atributos contextuales $A_1, ...A_n$:

$$P(e_i \mid A_1, ...A_n) = P(e_i) \prod_{i=0}^{N} P(A_i \mid e_i) \qquad (1)$$

El conjunto de atributos que, en experimentos preliminares, nos dieron mejores resultados es el formado por:

- e_{i-1} : la etiqueta que aparece inmediatamente a la izquierda de e_i
- e_{i+1} : la etiqueta que aparece inmediatamente a la derecha de e_i
- (t_i, e_{i-1}) : la copresencia del token ambiguo t_i junto con la etiqueta que aparece inmediatamente a la izquierda de e_i
- (t_i, e_{i+1}) : la copresencia del token ambiguo t_i junto con la etiqueta que aparece inmediatamente a la derecha de e_i

Tomando en cuenta estos cuatro atributos específicos, similares a los utilizados en (Banko y Moore, 2004), el producto de atributos genéricos de la ecuación 1 se puede especificar de esta manera:

$$P(e_{i-1} \mid e_i)P(e_{i+1} \mid e_i)P(t_i, e_{i-1} \mid e_i)P(t_i, e_{i+1} \mid e_i) \qquad (2)$$

Finalmente, se selecciona la etiqueta con el valor máximo de todas las posibles asignadas a un token ambiguo. Optamos por un algoritmo que desambigua los tokens de izquierda a derecha, lo que significa que el contexto a la izquierda de una palabra ambigua es un token ya desambiguado. Solo los contextos a la derecha pueden contenener tokens ambiguos susceptibles de asociarse a varias etiquetas, el conjunto de las cuales forma parte de los atributos contextuales de la etiqueta a desambiguar.

La estrategia bayesiana descrita aquí se encuentra conceptualmente próxima al formalismo de los modelos de Markov ocultos (HMM) (Brants, 2000), formalismo en el que se basa el desambiguador morfosintáctico de la suite lingüística FreeLing (Carreras et al., 2004; Padró y Stanilovsky, 2012), con el que comparamos nuestro sistema. La principal diferencia es que, en nuestro modelo, la desambiguación se hace token a token (como un problema de clasificación), en vez de buscar la mejor secuencia de etiquetas asociada a la frase de entrada mediante programación dinámica (algoritmo Viterbi). El motivo que nos ha llevado a utilizar un simple clasificador bayesiano como desambiguador es su eficiencia computacional.

3.3 Experimentos y evaluación

Para conocer la calidad de los resultados de nuestro sistema de análisis morfosintáctico, CitiusTagger, lo comparamos con otras

dos herramientas, FreeLing y Tree-Tagger (Schimd, 1995), entrenados con el mismo corpus de entrenamiento, el mismo léxico y el mismo conjunto de etiquetas (*tagset*). El corpus de entrenamiento utilizado fue generado con los primeros 300 mil tokens del corpus anotado Ancora (Taulé, Martí, y Recasens, 2008). El léxico es el utilizado por FreeLing, cuyo tagset coincide con el del corpus Ancora. El corpus de test consiste en otros 90 mil tokens extraídos del mismo corpus. El módulo de desambiguación morfosintáctica de FreeLing se basa en Modelos Ocultos de Markov (HMM), a partir de trigramas. El método de análisis de Tree-Tagger, por su parte, se basa en árboles de decisión.

La precisión se calcula dividiendo el número de tokens etiquetados correctamente por cada sistema entre el número total de tokens del corpus de test. La evaluación se llevó a cabo teniendo en cuenta las dos primeras letras de las etiquetas (o tres para algunas categorías), que fueron las únicas utilizadas en el proceso de entrenamiento. La tabla 1 muestra los resultados obtenidos por los tres sistemas. FreeLing alcanza la precisión más alta superando en 0,4 puntos a nuestro sistema, que a su vez, supera en 0,93 a Tree-Tagger. Estos resultados están muy próximos a otros experimentos descritos en (Gamallo y Garcia, 2013) para el portugués, donde FreeLing supera en 1,6 puntos a Tree-Tagger, usando también ambos sistemas el mismo corpus de entrenamiento y el mismo léxico.

Cabe destacar que hemos realizado otra prueba con la versión de Tree-Tagger original para español[7], adaptando las etiquetas de salida a nuestro corpus de test. Los resultados de esta versión son significativamente más bajos, alcanzando apenas el 92,13 % de precisión.

Sistemas	**Precisión (%)**
CitiusTagger	96,45
FreeLing	96,85
Tree-Tagger	95,52

Tabla 1: Precisión de tres sistemas de análisis morfosintáctico con el mismo corpus de entrenamiento y léxico

[7]http://www.cis.uni-muenchen.de/-~schmid/tools/TreeTagger/

4 Clasificación de Entidades (NEC)

Otra de las herramientas de nuestra suite lingüística es un módulo de clasificación de entidades nombradas (CitiusNEC). En la versión actual, este módulo está configurado para clasificar cuatro tipo de entidades: personas, lugares, organizaciones y otras (o misceláneo). El sistema se basa principalmente en consultas a ficheros de recursos extraídos de fuentes enciclopédicas y en un conjunto de reglas de desambiguación semántica. Los ficheros de recursos utilizados por el módulo fueron construidos utilizando como principales fuentes las grandes bases de información enciclopédica FreeBase[8] y DBpedia[9]. Las reglas de desambiguación se aplican a las entidades previamente identificadas por el reconocedor de entidades (NER), con el fin de resolver problemas de entidades ambiguas (a las que se asocian varias clases semánticas) o desconocidas (que no existen en los recursos). Para el caso de entidades conocidas no ambiguas, la clasificación es trivial, pues la clase asignada es la que se encuentra en el recurso enciclopédico correspondiente.

4.1 Reglas de desambiguación

La entrada del sistema NEC es texto etiquetado con entidades previamente identificadas por el sistema NER. Además, se requieren dos tipos de recursos: *gazetteers* y *triggers*. Los primeros son entidades clasificadas que fueron extraídas de FreeBase y DBpedia. Se construyeron tres ficheros de gazetteers (para personas, lugares y organizaciones), dejando fuera la categoría más heterogénea: "misceláneo". Los triggers son sustantivos que pueden subclasificar cada una de las tres categorías utilizadas. Por ejemplo, para la categoría "organización", seleccionamos como triggers, en base al árbol de categorías de la Wikipedia, lemas tales como: *institución, partido, asociación, federación, sindicato, club, entidad, empresa, cooperativa*, etc.

Dada una Entidad Nombrada (EN), el algoritmo de desambiguación procede de esta manera:

Consulta a los gazetteers: Si la EN se encuentra solo en una clase de gazetteers,

[8]http://www.freebase.com/

[9]http://dbpedia.org

entonces se considera que no es ambigua y se le asigna la clase encontrada.

Búsqueda de triggers: Si la EN aparece en varios gazetteers (ambigua) o si es desconocida (no se encuentra en los gazetteers), entonces se busca en el contexto lingüístico la aparición de triggers relevantes. El contexto se define como una ventana de N lemas a la izquierda y a la derecha de la EN a desambiguar, siendo la instanciación $N = 3$ la que mejores resultados devuelve en expermientos preliminares.

Ordenamiento de clases: Si la EN es ambigua y no puede desambiguarse por medio de la búsqueda contextual (etapa anterior), se selecciona la clase más probable (*prior probability*). Calculamos la probabilidad calculando la distribución de los gazetteers en la Wikipedia.

Verificación interna: Si la EN es desconocida y no puede desambiguarse por medio de la búsqueda contextual, entonces se verifica si la primera expresión constituyente de la EN coincide con la primera expresión de una EN en los gazetteers, o si es un nombre común que se encuentra en alguna de las listas de triggers. En caso de que se den varias opciones, se da preferencia a los gazetteers sobre los triggers y, cuando hay ambiguedad, utilizamos el ordenamiento de clases, tal y como se ha descrito arriba.

Else: Si ninguna regla se aplica, la EN se clasifica como "misceláneo".

Cabe destacar que las reglas en sí mismas son independientes de la lengua. Lo que es dependiente de una lengua concreta son los recursos utilizados. En (Gamallo y Garcia, 2011; Garcia, González, y del Río, 2012) se describe un sistema similar para el portugués y gallego, respectivamente, basado en reglas y dependiente de recursos externos.

4.2 Experimentos y evaluación

A continuación, comparamos el módulo CitiusNEC descrito arriba con dos sistemas NEC de aprendizaje supervisado:

- El módulo NEC de FreeLing, el cual obtuvo los mejores resultados en la competición *CoNLL-2002 shared task* (Tjong y Erik, 2002). Este sistema se basa en el algoritmo AdaBoost que consiste en combinar varios clasificadores básicos (Carreras et al., 2002). También utiliza recursos externos (gazetteers y triggers) para definir atributos específicos. El modelo utilizado en los experimentos descritos en esta sección es el que se encuentra en la última versión estable del paquete FreeLing (versión 3.1).
- Apache OpenNLP (Apache Software Foundation, 2014), cuyo módulo NEC permite entrenar modelos con dos algoritmos de aprendizaje: uno basado en redes neuronales (perceptrón) y otro basado en el principio de máxima entropía. Hemos entrenado un modelo para NEC con la siguiente configuración: algoritmo de máxima entropía y $Cutoff = 1$, que fue la que nos proporcionó mejores resultados.

Es importante reseñar que los modelos estadísticos de estos dos sistemas fueron entrenados con el corpus de entrenamiento de *CoNLL-2002 shared task*. En el caso de OpenNLP hemos añadido el corpus de desarrollo de dicha competición. Nuestro sistema, en cambio, no depende de ningún corpus de entrenamiento anotado, ya que se basa en recursos externos (supervisión distante).

Para llevar a cabo la evaluación cualitativa, nos servimos de dos corpus de test: el utilizado en *CoNLL-2002 shared task* para evaluar los sistemas en competición, y otro que llamamos *Hetero*, construido por nosotros a partir de diferentes fuentes: la Wikipedia y noticias de periódicos online (Gamallo y Garcia, 2011). Las tablas 2 y 3 muestran la precisión, *cobertura* y *f-score* obtenidos por los tres sistemas en ambos corpus de test.

Los resultados muestran que los dos sistemas entrenados con el corpus de *CoNLL-2002 shared task*, FreeLing y OpenNLP, consiguen mejores resultados que CitiusNEC cuando se evalúan con el corpus de test de *CoNLL-2002 shared task*, y por lo tanto, de características semejantes al corpus de entrenamiento. La precisión de estos dos sistemas baja significativamente cuando se evalúan con un corpus de naturaleza distinta a la del corpus de entrenamiento, como es el corpus de test *Hetero*. Nuestro módulo, CitiusNEC, mantiene resultados estables independientemente del tipo de corpus utilizado en la evaluación e, in-

Sistemas	Precisión (%)	Cobertura (%)	F-score (%)
CitiusNEC	67,47	66,33	66,89
FreeLing	75,08	76,90	75,98
OpenNLP	**78,96**	**79,09**	**79,02**

Tabla 2: Resultados de tres sistemas NEC utilizando el corpus de test *CoNLL-2002 shared task*

Sistemas	Precisión (%)	Cobertura (%)	F-score (%)
CitiusNEC	**67,47**	65,37	**66,40**
FreeLing	65,67	65,44	65,56
OpenNLP	64,50	**66,84**	65,65

Tabla 3: Resultados de tres sistemas NEC utilizando el corpus de test *Hetero*

cluso, supera ligeramente en f-score a estos dos sistemas con el corpus *Hetero*. Cabe destacar que en el trabajo descrito en (Gamallo y Garcia, 2011), se observó la misma tendencia al evaluar FreeLing y una versión anterior a CitiusNEC para el portugués.

5 *Paralelización en la ejecución de los módulos*

Una tubería de módulos como la descrita en este artículo puede adaptarse para su procesamiento en paralelo mediante el modelo de programación MapReduce, y usando Apache Hadoop. Desde la publicación por parte de Google del modelo de programación MapReduce (Dean y Ghemawat, 2004), surgieron un conjunto de herramientas, muchas de ellas de código abierto, que utilizan este algoritmo distribuido. MapReduce divide el procesamiento de un algoritmo en etapas paralelizables que se ejecutan en muchos nodos (*mappers*), así como en etapas de agregación donde los datos obtenidos en la fase previa son procesados en un único nodo (*reducers*). De esta manera se facilita el almacenamiento y procesamiento del llamado *Big Data*. La herramienta más extendida basada en el modelo MapReduce es Hadoop, desarrollada inicialmente por Yahoo y después publicada como proyecto Apache. Hadoop permite almacenar y procesar de manera distribuida enormes cantidades de datos no estructurados haciendo uso de clusters de computadores de bajo coste (*commodity hardware*), proporcionando al mismo tiempo facilidad de programación, escalabilidad y tolerancia a fallos. De hecho, su modelo de programación es simple y oculta al desarrollador detalles de bajo nivel. Según estimaciones de Shaun Connolly lanzadas en el Hadoop Summit 2012, en el año 2015 alrededor del 50 % de todos los datos mundiales serán almacenados y procesados en clusters basados en Apache Hadoop[10].

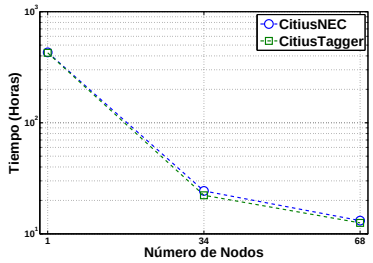
La integración de nuestra suite lingüística en Hadoop permite que los textos de entrada se particionen en subconjuntos, y que cada uno de ellos se procese en un nodo del cluster en paralelo con el resto de nodos. Con el objetivo de evitar cualquier tipo de modificación sobre los módulos PLN originales aprovechamos también un servicio que ofrece Hadoop, llamado "*Hadoop streaming*", que permite el uso de cualquier ejecutable como *mapper* o *reducer*, independientemente del lenguaje de programación en el que esté escrito.

5.1 Datos sobre eficiencia

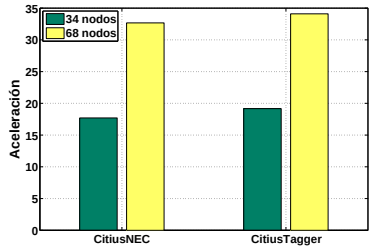
En la figura 2 mostramos los resultados de rendimiento de los módulos CitiusNEC y CitiusTagger una vez integrados en la infraestructura de Hadoop. Los experimentos se han llevado a cabo en un clúster del Centro de Supercomputación de Galicia (CESGA) compuesto por 68 nodos, cada uno de los cuales a su vez consiste en un procesador Intel Xeon E5520. Como texto de entrada de ambos módulos para los tests se ha utilizado la Wikipedia en español, cuyo tamaño en texto plano es de 2,1 GB.

En la figura 2(a) mostramos una comparación entre los tiempos de ejecución de los módulos originales cuya ejecución es secuencial (es decir, usando un único procesador), y los módulos una vez integrados con Hadoop usando en la ejecución paralela 34 y 68 nodos. Los módulos en su ejecución secuencial tardan en torno a unas 425 horas para procesar la Wikipedia en español, lo que supone algo más de 2 semanas. Estos tiempos de ejecución hacen inviable la aplicación de estos

[10] http://www.biganalytics2012.com/resources/Shaun_Connolly-HadoopNowNextBeyond-v10.pdf



(a)



(b)

Figura 2: Rendimiento de los módulos PLN después de su integración con Hadoop: tiempos de ejecución (a) y aceleración (b).

módulos PLN a problemas de gran tamaño como el que abordamos aquí. Sin embargo, cuando usamos la versión paralela con Hadoop los tiempos se reducen de forma notable. Así, por ejemplo, usando 68 nodos, los módulos CitiusNEC y CitiusTagger necesitarían 13 y 12,1 horas respectivamente. Aunque este tiempo sigue siendo elevado debemos tener en cuenta que si usásemos un mayor número de nodos en nuestro clúster Hadoop, el tiempo seguiría escalando y reduciéndose.

Para comprobar la escalabilidad de nuestros módulos PLN mostramos en la figura 2(b) los resultados de aceleración con respecto al caso secuencial. La aceleración se calcula como el cociente entre el tiempo de ejecución secuencial y el tiempo de ejecución paralelo. En el caso ideal, usando un clúster de N procesadores, obtendríamos una aceleración de N. Sin embargo, esta situación es difícilmente alcanzable, puesto que siempre hay tiempos necesarios para la inicialización, además de partes no paralelizables del código (Ley de Amdhal). En nuestro caso obtenemos aceleraciones superiores a 30 usando 68 nodos. Esto indica que los módulos integrados en la infraestructura Hadoop obtienen los resultados 30 veces más rápido que los módulos en su versión secuencial.

5.2 Trabajo relacionado

A diferencia de los algoritmos de minería de datos donde existen herramientas específicas que explotan las capacidades analíticas de Hadoop (p.ej. Apache Mahout para clasificadores, recomendadores y algoritmos de clustering y Apache Giraph para el procesamiento de grafos), no conocemos a día de hoy ninguna herramienta que emplee de forma integrada soluciones de PLN en *Big Data*.

Recientemente, el paradigma MapReduce se ha comenzado a aplicar a algunas tareas de PLN, como por ejemplo la tradución estatística (Ahmad et al., 2011; Dyer, Cordora, y Lin, 2008), la construción de matrices de coocurrencias (Lin, 2008), la minería de textos (Balkir, Foster, y Rzhetsky, 2011), la computación de similaridades semánticas (Pantel et al., 2009), o la adquisición de paráfrasis (Metzel y Hovy, 2011).

6 Conclusiones

Hemos presentado una suite de módulos lingüísticos escritos en Perl y organizados en una simple arquitectura en *tuberías*, cuyo desempeño está próximo al estado del arte. Los datos de entrada pueden particionarse en varios tubos ejecutados en paralelo en un clúster con la ayuda de *Hadoop Streaming*.

Cabe destacar que el módulo de clasificación de entidades nombradas puede extenderse con facilidad a un conjunto mayor de clases semánticas, puesto que no depende de la construcción de un corpus anotado. Al basarse en la técnica de la supervisión distante, solo depende de las fuentes externas enciclopédicas. En estos momentos, estamos extendiendo los recursos para dar cuenta de nuevas categorías semánticas específicas ligadas al dominio de la informática.

Actualmente, hemos desarrollado una versión de los módulos para portugués y estamos también desarrollando nuevas versiones para gallego e inglés, con el propósito de crear una suite multilingüe adaptada para su uso con tecnologías *Big Data*.

Bibliografía

Agerri, R., J. Bermudez, y G. Rigau. 2014. Efficient and easy nlp processing with ixa pipeline. En *Demo Sessions of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden.

Ahmad, R., P. Kumar, B. Rambabu, P. Sajja anda M.K. Sinha, y P. Sangal. 2011. Enhancing throughout of a machine translation system using mapreduce framework: An engineering approach,. En *9th International Conference on Natural Language Processing ICON-2011*, Hyderabad, India.

The Apache Software Foundation, 2014. *Apache OpenNLP*.

Balkir, A.S., I. Foster, y A. Rzhetsky. 2011. A distributed look-up architecture for text mining applications using mapreduce. En *International Conference for High Performance Computing, Networking, Storage and Analysis*.

Banko, Michele y Robert Moore. 2004. Part of speech tagging in context. En *COLING'04, 20th international conference on Computational Linguistics*.

Brants, Throrsten. 2000. Tnt: A statistical part-of-speech tagger. En *6th Conference on Applied Natural Language Processing. ANLP, ACL-2000*.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. An Open-Source Suite of Language Analyzers. En *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Carreras, X., L. Marquez, L. Padró, y M. Padró. 2002. Named entity extraction using adaboost. En *COLING-02 proceedings of the 6th Conference on Natural Language Learning*.

Dean, J. y S. Ghemawat. 2004. Mapreduce: Simplified data processing on large clusters OSDI-04. En *Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, EE.UU.

Dyer, C., A. Cordora, y J. Lin. 2008. Fast, easy and cheap: Construction of statistical machine translation model with mapreduce. En *3rd Workshop on Statistical Machine Translation*, Columns, Ohio.

Gamallo, Pablo y Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. *LNCS*, 7026:610–623.

Gamallo, Pablo y Marcos Garcia. 2013. Freeling e treetagger: um estudo comparativo no âmbito do português. En *ProLNat Technical Report, vol. 01, URL: http://gramatica.usc.es/∼gamallo/artigos-web/PROLNAT_Report_01.pdf.*

Garcia, M., I. González, y I. del Río. 2012. Identificação e classificação de entidades mencionadas em galego. *Estudos de Linguística Galega*, 4:13–25.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics*, 31(1):147–151.

Lin, J. 2008. Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with mapreduce. En *2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, USA.

Metzel, D. y E. Hovy. 2011. Mavuno: a scalable and effective hadoop-based paraphrase acquisition system. En *LDMTA-11, Third Workshop on Large Scale Data Mining: Theory and Applications*.

Padró, Lluís. y Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Pantel, P., E. Crestan, A. Borkovsky, A.M. Popescu, y V. Vyas. 2009. Web-scale distributional similarity and entity set expansion. En *Conference on Empirical Methods in Natural Language Processing*, Singapur.

Schimd, Helmut. 1995. Improvements in part-of-speech tagging with an application to german. En *ACL SIGDAT Workshop*, Dublin, Ireland.

Tablan, V., I. Roberts, H. Cunningham, y K. Bontcheva. 2013. Gatecloud. net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371.

Taulé, M., M.A. Martí, y M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. En *The 6th International Conference on Language Resources and Evaluation (LREC).*, Marrakesh, Morocco.

Tjong, Kim Sang y F. Erik. 2002. Introduction ot the CoNLL-2002 shared task: Language independent named entity recognition. En *Conference on Natural Language Learning*.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 25-32 recibido 15-04-14 revisado 06-06-14 aceptado 10-06-14

Entity-Centric Coreference Resolution of Person Entities for Open Information Extraction *

Resolución de Correferencia Centrada en Entidades Persona para Extracción de Información Abierta

Marcos Garcia and **Pablo Gamallo**
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
{marcos.garcia.gonzalez, pablo.gamallo}@usc.es

Resumen: Este trabajo presenta un sistema de resolución de correferencia de entidades *persona* cuya arquitectura se basa en la aplicación secuencial de módulos de resolución independientes y en una estrategia centrada en las entidades. Diversas evaluaciones indican que el sistema obtiene resultados prometedores en varios escenarios (≈ 71% y ≈ 81% de F1 CoNLL). Con el fin de analizar la influencia de la resolución de correferencia en la extracción de información, un sistema de extracción de información abierta se ha aplicado sobre textos con anotación correferencial. Los resultados de este experimento indican que la extracción de información mejora tanto en cobertura como en precisión. Las evaluaciones han sido realizadas en español, portugués y gallego, y todas las herramientas y recursos son distribuidos libremente.
Palabras clave: correferencia, anáfora, extracción de información abierta

Abstract: This work presents a coreference resolution system of person entities based on a multi-pass architecture which sequentially applies a set of independent modules, using an entity-centric approach. Several evaluations show that the system obtains promising results in different scenarios (≈ 71% and ≈ 81% F1 CoNLL). Furthermore, the impact of coreference resolution in information extraction was analyzed, by applying an open information extraction system after the coreference resolution tool. The results of this test indicate that information extraction gives better both recall and precision results. The evaluations were carried out in Spanish, Portuguese and Galician, and all the resources and tools are freely distributed.
Keywords: coreference, anaphora, open information extraction

1 Introduction

Relation Extraction (RE) systems automatically obtain different kinds of knowledge from unstructured texts, used for instance to build databases or to populate ontologies.

While RE systems usually depend on a set of predefined semantic relations for obtaining the knowledge (e.g., `hasHeadquartersAt`{*Organization, Location*}), Open Information Extraction (OIE) approaches perform unsupervised extractions of all types of verb-based relations (Banko et al., 2007).

In this respect, one of the main kind of extractions is related to person entities, and its objective is to obtain information about concrete people. For example, from a sentence such as "Obikwelu won the 100m gold medal at the 2009 Lusophony Games", an OIE system may obtain the following structured knowledge (with two arguments and a verb-based relation in each extraction):

OIE$_1$: *Obikwelu*$_{\text{Arg1}}$ `won` *the 100m gold medal*$_{\text{Arg2}}$

OIE$_2$: *Obikwelu*$_{\text{Arg1}}$ `won_the_100m_gold medal_at` *the 2009 Lusophony Games*$_{\text{Arg2}}$

However, many of the mentions of each person entity in a text are different from the others, so the final extraction may not be semantically complete. An OIE system could extract, from the same text than the previous example, relations like the following:

OIE$_3$: *Francis Obiorah Obikwelu*$_{\text{Arg1}}$ `is_a` *sprint athlete*$_{\text{Arg2}}$

OIE$_4$: *who*$_{\text{Arg1}}$ `is_based_in` *Lisbon*$_{\text{Arg2}}$

OIE$_5$: *He*$_{\text{Arg1}}$ `was_5th_in` *the 200m*$_{\text{Arg2}}$

* This work has been supported by the HPCPLN project – Ref: EM13/041 (Galician Government) and by the Celtic – Ref: 2012-CE138 and Plastic – Ref: 2013-CE298 projects (Feder-Interconnecta).

ISSN 1135-5948

On the one hand, these extractions do not include the referents of the pronouns (*who*, *He*), so the knowledge could not be semantically useful. On the other hand, they do not report that *Obikwelu*, *Francis Obiorah Obikwelu* and *He* refer to the same entity, while *who* refers to other person.[1]

Related to that, Coreference Resolution (CR) systems use different techniques for clustering the various mentions of an entity into the same group. So, applying a coreference resolution tool before an OIE (or RE) system should improve the extraction in two ways: (1) increasing the recall by disambiguating the pronouns and (2) adding semantic knowledge by clustering both nominal and pronominal mentions of each entity.

This paper presents an open-source CR system for person entities which uses a multi-pass architecture. The approach, inspired by the Stanford Coreference Resolution System (Raghunathan et al., 2010), consists of a battery of modules applied from high-precision to high-recall. The system is also applied before a state-of-the-art OIE tool, in order to evaluate the impact of CR when performing information extraction.

The individual evaluations of the CR system show that the multi-pass architecture achieves promising performance when analyzing person entities ($\approx$ 71%/81% F1 CoNLL). Moreover, the OIE experiments prove that applying CR before an OIE system allows it to increase both the precision and the recall of the extraction.

Section 2 contains some related work and Section 3 presents the coreference resolution system. Then, Section 4 shows the results of both CR and OIE experiments while Section 5 points out the conclusions of this work.

2 Related Work

The different strategies for CR can be organized using two dichotomies: On the one hand, mention-pair *vs* entity-centric approaches. On the other hand, rule-based *vs* machine learning systems.

Mention-pair strategies decide if two mentions corefer using the features of these specific mentions, while entity-centric approaches take advantage of features obtained from other mentions of the same entities.

Rule-based strategies make use of sets of rules and heuristics for finding the best element to link each mention to (Lappin and Leass, 1994; Baldwin, 1997; Mitkov, 1998).

Machine learning systems rely on annotated data for learning preferences and constraints in order to classify pairs of mentions or entities (Soon, Ng, and Lim, 2001; Sapena, Padró, and Turmo, 2013). Some unsupervised models apply clustering approaches for solving coreference (Haghighi and Klein, 2007). Even though complex machine learning models obtain good results in this task, Raghunathan et al. (2010) presented a rule-based system that outperforms previous approaches. This system is based on a multi-pass strategy which first solves the *easy* cases, then increasing recall with further rules (Lee et al., 2013). Inspired by this method, *EasyFirst* uses annotated corpora in order to know whether coreference links are easier or harder (Stoyanov and Eisner, 2012).

For Spanish, Palomar et al. (2001) described a set of constraints and preferences for pronominal anaphora resolution, while Recasens and Hovy (2009) analyzed the impact of several features for CR. The availability of a large annotated corpus for Spanish (Recasens and Martí, 2010) also allowed other supervised systems being adapted for this language (Recasens et al., 2010).

Concerning OIE, several strategies were also applied since the first system, TextRunner (Banko et al., 2007). This tool (and further versions of it, such as ReVerb (Fader, Soderland, and Etzioni, 2011)) uses shallow syntax and labeled data for extracting triples (*argument_1, relation, argument_2*) which describe basic propositions.

Other OIE systems take advantage of dependency parsing for extracting the relations, such as WOE (Wu and Weld, 2010), which uses a learning-based model, or *DepOE* (Gamallo, Garcia, and Fernández-Lanza, 2012), a multi-lingual rule-based approach.

3 LinkPeople

This section describes LinkPeople, an entity-centric system which sequentially applies a battery of CR modules (Garcia and Gamallo, 2014a). Its architecture is inspired by Raghunathan et al. (2010), but it adds new modules for both cataphoric and elliptical pronouns as well as a set of syntactic constraints which

[1] In this paper, a *mention* is every instance of reference to a person, while an *entity* is all the mentions referring to the same person in the text (Recasens and Martí, 2010).

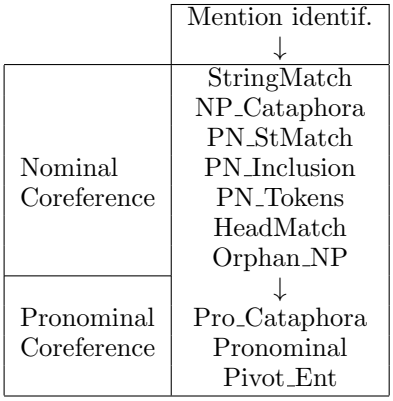


Figure 1: Architecture of the system.

Who was $_1$[the singer of the Beatles]$_1$. $_2$[The musician John Winston Ono Lennon]$_1$ was one of the founders of the Beatles. With $_3$[Paul McCartney]$_2$, $_4$[he]$_1$ formed a songwriting partnership. $_5$[Lennon]$_1$ was born at Liverpool Hospital to $_6$[Julia]$_3$ and $_7$[Alfred Lennon]$_4$. $_{8/9}$[$_{10}$[His]$_1$ parents]$_{3/4}$ named $_{11}$[him]$_1$ $_{12}$[John Winston Lennon]$_1$. $_{13}$[Lennon]$_1$ revealed a rebellious nature and acerbic wit. $_{14}$[The musician]$_1$ was murdered in 1980.

Figure 2: Example of a coreference annotation of person entities. Mentions appear inside brackets. Numbers at the left are mention ids, while entity ids appear at the right.

restrict some links between pronominal and nominal mentions.

3.1 Architecture

The different modules of LinkPeople are applied starting from the most precise ones. Thus, *easy* links between mentions are done first. The entity-centric approach allows the system to use, in further modules, features that had been extracted in previous passes.

Figure 1 summarizes the architecture of LinkPeople. It starts with a mention identification module which selects the markables in a text. After that, a battery of nominal CR modules is executed. Finally, the pronominal solving passes —including a set of syntactic constraints— are applied.

Figure 2 contains a text with coreference annotation of person entities, used for exemplifying the architecture of LinkPeople.

In the first stage, a specific pass identifies the mentions referring to a person entity, using the information provided by the PoS-tagger and the NER as well as applying basic approaches for NP and elliptical pronoun identification: First, personal names (and noun phrases including personal names) are identified. Then, it seeks for NPs whose head may refer to a person (e.g., "the singer"). Finally, this module selects singular possessives and applies basic rules for identifying relative, personal and elliptical pronouns (Ferrández and Peral, 2000). At this step, each mention belongs to a different entity. Each entity contains the gender, number, head of a noun phrase, head of a Proper Noun (PN) and full proper noun as features. Once the mentions are identified, the CR modules are sequentially executed.

Each module applies the following strategy (except for some exceptional rules, explained below): mentions are traversed from the beginning of the text and each one is *selected* if (i) it is not the first mention of the text and (ii) it is the first mention of its entity. Once a mention is selected, it looks backwards for *candidates* in order to find an appropriate antecedent. If an antecedent is found, mentions are merged together in the same entity. Then, the next selected mention is evaluated. Apart from the mention identification pass, the current version of LinkPeople contains the following modules:

StringMatch: performs strict matching of the whole string of both mentions (the selected one and the candidate). In the example (Figure 2), mentions 13 and 5 are linked in this step.

NP_Cataphora: verifies if the first mention is a NP without a personal name. If so, it is considered a cataphoric mention, and the system checks if the next sentence contains a personal name as a Subject. In this case, these mentions are linked if they agree in gender and number. In the example, mentions 1 and 2 merge. Note that, at the end of this pass, this entity has as NP heads the words 'singer' and 'musician', and 'John Winston Ono Lennon' as the PN.

PN_StMatch: looks for mentions which share the whole PN, even if their heads are different (or if one of them does not have head). "The musician John Lennon" and "John Lennon" (not in Figure 2) would be an example.

PN_Inclusion: verifies if the full PN of the selected mention (in the entity) includes the

PN of the candidate mention (also in the entity), or vice-versa. In the example, this rule links mentions 5 and 2.

PN_Tokens: splits the full PN of a partial entity in its tokens, and verifies if the full PN of the candidate contains all the tokens in the same order, or vice-versa (except for some stop-words, such as "Sr.", "Jr.", etc.). As the pair "John Winston Ono Lennon" - "John Winston Lennon" are compatible, mentions 12 and 5 are merged.

HeadMatch: checks if the selected mention and the candidate one share the heads (or the heads of their entities). In Figure 2, mention 14 is linked to mention 13.

Orphan_NP: applies a pronominal-based rule to orphan NPs. A definite NP is marked as orphan if it is still a singleton and it does not contain a personal name. Thus, an orphan NP is linked to the previous PN with gender and number agreement. In the example, the mentions 8/9 are linked to 7 and 6.

Pro_Cataphora: verifies if a text starts with a personal (or elliptical) pronoun. If so, it seeks in the following sentence if there is a compatible PN.

Pronominal: this is the standard module for pronominal CR. For each selected pronoun, it verifies if the candidate nominal mentions satisfy the syntactic (and morphosyntactic) constraints. They include a set of constraints for each type of pronoun, which remove a candidate if any of them is violated. Some of them are: an object pronoun (direct or indirect) cannot corefer with its subject (mention 11 *vs* mentions 8/9); a personal pronoun does not corefer with a mention inside a prepositional phrase (mention 4 *vs* mention 3), a possessive cannot corefer with the NP it belongs to (mention 10 *vs* mentions 8/9) or a pronoun prefers a subject NP as its antecedent (mentions 10 and 11 *vs* mentions 6 and 7). This way, in Figure 2 the pronominal mention 4 is linked to mention 2, and mentions 10 and 11 to mention 5. This module has as a parameter the number of previous sentences for looking for candidates.

Pivot_Ent: this module is only applied if there are orphan pronouns (not linked to any proper noun/noun phrase) at this step. First, it verifies if the text has a pivot entity, which is the most frequent personal name in a text whose frequency is at least 33% higher than the second person with more occurrences. Then, if there is a pivot entity, all the orphan pronouns are linked to its mention. If not, each orphan pronoun is linked to the previous PN/NP (with no constraint).

4 Experiments

This section contains the performed evaluations. First, several experiments on CR are described. Then, a test of an OIE system is carried out, analyzing how LinkPeople influences the results of the extraction. All the experiments were performed in three languages: Spanish, Portuguese and Galician.[2]

4.1 Coreference Resolution

This section performs several tests with LinkPeople, comparing its results in different scenarios. First, the performance of LinkPeople using corpora with the mentions already identified (*gold-mentions*). Then, the basic mention identification module described in Section 3.1 is applied (*system-mentions*).[3] Both gold-mentions and system-mentions results were obtained with predicted information (*regular setting*): lemmas and PoS-tags (Padró and Stanilovsky, 2012; Garcia and Gamallo, 2010), NER (Padró and Stanilovsky, 2012; Garcia, Gayo, and González López, 2012; Gamallo and Garcia, 2011), and dependency annotation (Gamallo and González López, 2011), and with any kind of external knowledge (*closed setting*).

The experiments were performed with a Spanish corpus (46k tokens and $\approx 4,500$ mentions), a Portuguese one (51k tokens and $\approx 4,000$ mentions) and a Galician dataset (42k tokens and $\approx 3,500$ mentions) (Garcia and Gamallo, 2014b).

Three baselines were used: (i) *Singletons*, where every mention belongs to a different entity. (ii) *All_in_One*, where all the mentions belong to the same entity and (iii) *Head-Match_Pro*, which clusters in the same entity those mentions sharing the head, and links each pronoun to the previous nominal mention with gender and number agreement.[4]

[2] All the tools and resources are freely available at `http://gramatica.usc.es/~marcos/LinkP.tbz2`

[3] Except for elliptical pronouns, where the gold-mentions were used for preserving the alignment, required for computing the results. Experiments in Section 4.2 simulate a real scenario.

[4] Due to language and format differences, other CR systems could not be used for comparison (Lee et al., 2013; Sapena, Padró, and Turmo, 2013).

Language	Model	MUC			B^3			$CEAF_e$			CoNLL
		R	P	F1	R	P	F1	R	P	F1	F1
Spanish	HeadMatch_Pro	78.2	90.7	84.0	35.3	81.2	49.2	72.9	51.5	60.4	64.5
	LinkPeople	84.1	94.1	**88.8**	62.9	84.8	**72.2**	83.4	71.0	**76.7**	**79.2**
Portuguese	HeadMatch_Pro	76.0	91.2	82.9	46.0	85.8	59.9	76.7	49.2	59.9	67.6
	LinkPeople	82.7	92.7	**87.4**	65.8	84.5	**74.0**	84.4	67.9	**75.2**	**78.9**
Galician	HeadMatch_Pro	81.9	89.8	85.7	44.1	83.6	57.7	70.0	53.5	60.6	68.0
	LinkPeople	89.0	94.6	**91.7**	72.9	88.4	**79.9**	87.6	76.6	**81.7**	**84.4**

Table 1: Results of LinkPeople (gold-mentions) compared to the best baseline (HeadMatch_Pro).

Language	Model	MUC			B^3			$CEAF_e$			CoNLL
		R	P	F1	R	P	F1	R	P	F1	F1
Spanish	Singletons	-	-	-	9.2	90.2	16.7	63.3	8.3	14.7	10.5
	All_In_One	77.5	78.8	78.1	69.0	44.2	53.9	5.7	73.5	10.5	47.5
	HeadMatch_Pro	68.2	81.5	74.2	31.4	68.1	43.0	62.6	52.0	56.8	58.0
	StringMatch	65.1	81.7	72.5	26.5	70.4	38.5	63.3	42.1	50.6	53.8
	PN_StMatch	66.3	81.8	73.2	27.2	69.4	39.0	63.1	45.0	52.6	54.9
	PN_Inclusion	69.6	82.8	75.6	34.2	68.2	45.5	64.6	55.0	59.4	60.2
	PN_Tokens	69.7	82.8	75.7	35.0	68.2	46.2	64.7	55.5	59.8	60.6
	Pronominal	70.2	82.3	75.8	35.9	67.3	46.8	64.6	59.4	61.9	61.5
	LinkPeople	72.8	85.3	**78.5**	50.4	70.7	**58.9**	73.5	68.3	**70.8**	**69.4**
Portuguese	Singletons	-	-	-	12.9	88.2	22.5	62.5	11.1	18.8	13.8
	All_In_One	75.6	73.1	74.3	67.1	37.7	48.3	9.5	61.3	16.4	46.3
	HeadMatch_Pro	65.4	79.4	71.7	41.5	68.9	51.8	69.2	54.2	60.8	61.4
	StringMatch	61.4	79.4	69.2	35.0	71.6	47.0	70.8	46.0	55.8	57.3
	PN_StMatch	63.7	79.9	70.9	36.6	70.8	48.3	71.6	50.7	59.4	59.5
	PN_Inclusion	67.9	81.1	73.9	45.5	69.7	55.1	74.0	61.9	67.4	65.5
	PN_Tokens	68.1	81.1	74.0	45.8	69.7	55.3	74.0	62.4	67.7	65.7
	Pronominal	69.0	81.2	74.6	47.9	69.2	56.6	74.0	65.4	69.4	66.9
	LinkPeople	69.9	82.0	**75.5**	55.8	69.2	**61.8**	76.6	68.8	**72.5**	**69.9**
Galician	Singletons	-	-	-	12.7	84.0	22.0	67.5	10.0	17.4	13.1
	All_In_One	83.1	71.1	76.6	75.9	40.9	53.2	7.4	67.1	13.3	47.7
	HeadMatch_Pro	72.4	73.7	73.0	38.0	62.3	47.2	61.4	52.5	56.6	59.0
	StringMatch	68.7	73.8	71.2	31.5	65.1	42.4	66.4	45.3	53.8	55.8
	PN_StMatch	70.7	74.0	72.3	34.4	64.1	44.8	66.6	50.4	57.4	58.1
	PN_Inclusion	74.6	75.2	74.9	42.8	63.0	50.9	66.7	60.1	63.2	63.0
	PN_Tokens	74.9	75.2	75.1	43.5	63.0	51.5	66.9	61.0	63.8	63.4
	Pronominal	75.0	75.0	75.0	44.2	62.7	51.8	67.0	62.4	64.6	63.8
	LinkPeople	78.5	78.5	**78.5**	65.0	67.0	**66.0**	78.6	73.4	**75.9**	**73.5**

Table 2: Results of LinkPeople (system-mentions) compared to the baselines.

The results were obtained using four metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_{entity}$ (Luo, 2005) and CoNLL (Pradhan et al., 2011). They were computed with the CoNLL 2011 scorer.

Table 1 contains the results of the best baseline and of LinkPeople using gold-mentions (for the full results of this scenario see Garcia and Gamallo (2014a)).

Table 2 includes the results of the three baselines and the performance values of LinkPeople using different modules added incrementally.[5] Central rows of Table 2 (StringMatch > Pronominal) include two baseline rules which classify mentions not covered by the active modules: (1) nominal mentions not analyzed are singletons and (2) pronouns are linked to the previous mention with number and gender agreement.

In every language and scenario, *HeadMatch_Pro* obtains good results, (as Recasens and Hovy (2010) shown), with $\approx$ 10% (F1 CoNLL) more than *All_in_One*.

The first module of LinkPeople (StringMatch) obtains lower results than the *HeadMatch_Pro* baseline, but with better pre-

[5]For spatial reasons, results of the modules with less performance improvements are omitted.

cision (except with the $CEAF_e$ metric). After including more matching modules (NP_Cataphora and PN_StMatch), the results are closer to the best baseline, while the addition of PN_Inclusion and PN_Tokens modules allows the system to surpass it.

Then, HeadMatch, Orphan_NP and PRO_Cataphora slightly improve the performance of the system, while the pronominal resolution module notoriously increases the results in every evaluation and language. At this stage, LinkPeople obtains $\approx$ 76% and $\approx$ 64% (F1 CoNLL) in the gold-mentions and system-mentions scenarios, respectively.

Finally, one of the main contributions to the performance of LinkPeople is the combination of the Pronominal module with the Pivot_Ent one. This combination reduces the scope of the pronominal module, thus strengthening the impact of the syntactic constraints. Furthermore, Pivot_Ent looks for a prominent person entity in each text, and links the orphan pronouns to this entity.

The results of LinkPeople ($\approx$ 81% —gold-mentions— and $\approx$ 71% —system-mentions) show that this approach performs well for solving the coreference of person entities in different languages and text typologies.

4.2 Open Information Extraction

In order to measure the impact of LinkPeople in OIE, the most recent version of *DepOE*, was executed on the output of the CR tool.

LinkPeople was applied using a system-mentions approach, and without external resources. Apart from that, a basic elliptical pronoun module was included, which looks for elliptical pronouns in sentence-initial position, after adverbial phrases and after prepositional phrases. All the linguistic information was predicted by the same NLP tools referred in Section 4.1.

One corpus for each of the three languages was collected for performing the experiments. Each corpus contains 5 articles from Wikipedia and 5 from online journals.

DepOE was applied two times: First, using as input the plain text of the selected corpora (DepOE). Then, applied on the output of LinkPeople (DepOE+).

For computing precision of *DepOE*, 300 randomly selected triples containing at least a mention of a person entity as one of its arguments were manually revised (100 per language: 50 from Wikipedia and 50 from the journals). In the first run (without CR), extractions with pronouns as arguments were not computed, since they were considered as semantically underspecified. Thus, the larger number of extractions in the second run (DepOE+) is due to the identification of personal (including elliptical) pronouns. The central column of Table 3 contains an example of a new extraction obtained by virtue of CR.

LinkPeople also linked nominal mentions with different forms (right column in Table 3), thus enriching the extraction by allowing the OIE system to group various information of the same entity. An estimation of this improvement was computed as follows: from all the correct (revised) triples, it was verified if the personal mention in the argument had been correctly solved by LinkPeople. These cases were divided by the total number of correct triples, being these results considered as the *enrichment* value.

Table 4 contains the results of both DepOE and DepOE+ runs. DepOE+ was capable of extracting 22.7% more triples than the simple model, and its precision increased in about 10.6%. These results show that the improvement was higher in Wikipedia. This is due to the fact that the largest (person) entity in encyclopedic texts is larger than those in journal articles. Besides, Wikipedia pages contain more anaphoric pronouns referring to person entities (Garcia and Gamallo, 2014b). Finally, last column of Table 4 includes the percentage of enrichment of the extraction after the use of LinkPeople. Even tough these values are not a direct evaluation of OIE, they suggest that the information extracted by an OIE system is about 79% better when obtained after the use of a CR tool.

5 Conclusions

This paper presented LinkPeople, an entity-centric coreference resolution system for person entities which uses a multi-pass architecture and a set of linguistically motivated modules. It was evaluated in three languages, using different scenarios and evaluation metrics, achieving promising results.

The performance of the system was also evaluated in a real-case scenario, by analyzing the impact of coreference solving for open information extraction. The results show that using LinkPeople before the application of an OIE system allows to increase the performance of the extraction.

Sentence	"Debutó en la Tercera división"	"Anderson viajou por Europa"
DepOE	∅	*Anderson* `viajou_por` *Europa*
DepOE+	*Ander Herrera* `debutó_en` *la Tercera división*	*Wes Anderson* `viajou_por` *Europa*

Table 3: Extraction examples of DepOE and DepOE+ in Spanish (left) and Portuguese (right). The DepOE+ extraction in Spanish extracts a new triple —not obtained by DepOE— from a sentence with elliptical subject, while the first argument of the Portuguese example is enriched with the full proper name (and linked to other mentions in the same text).

Lg	**DepOE**			**DepOE+**			**E**
	W	**J**	**P**	**W**	**J**	**P**	
Sp.	47	82	49%	80	86	58%	84%
Pt.	82	133	39%	111	155	56%	75%
Gl.	168	114	49%	221	115	54%	77%

Table 4: Results of the two runs of *DepOE*. *W* and *J* include the number of extractions from Wikipedia and journalistic articles, respectively. *P* is the precision of the extraction, and *E* refers to the quality enrichment provided by LinkPeople.

In further work, the implementation of rules for handling plural mentions is planned, together with the improvement of nominal and pronominal constraints.

References

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Workshop on Linguistic Coreference at the 1st International Conference on Language Resources and Evaluation*, volume 1, pages 563–566.

Baldwin, Breck. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45.

Banko, Michele, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the International Joint Conference on Artifical Intelligence*, pages 2670–2676.

Fader, Anthony, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Ferrández, Antonio and Jesús Peral. 2000. A computational approach to zero-pronouns in spanish. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 166–172.

Gamallo, Pablo and Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. In *Progress in Artificial Intelligence (LNCS/LNAI)*, volume 7026/2011, pages 610–623.

Gamallo, Pablo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based Open Information Extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18.

Gamallo, Pablo and Isaac González López. 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

Garcia, Marcos and Pablo Gamallo. 2010. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.

Garcia, Marcos and Pablo Gamallo. 2014a. An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information. In *Proceedings of the International Conference on Computational Linguistics*.

Garcia, Marcos and Pablo Gamallo. 2014b. Multilingual corpora with coreference annotation of person entities. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3229–3233.

Garcia, Marcos, Iria Gayo, and Isaac González López. 2012. Identificação e Classificação de Entidades Mencionadas em Galego. *Estudos de Lingüística Galega*, 4:13–25.

Haghighi, Aria and Dan Klein. 2007. Unsupervised coreference resolution in a non-

parametric bayesian model. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 45, pages 848–855.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Lee, Heeyoung, Angel Chang, Yves Peirsman, N. Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Luo, Xiaoqiang. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, volume 2, pages 869–875.

Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference*.

Palomar, Manuel, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, and Rafael Muñoz. 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Raghunathan, Kathik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Recasens, Marta and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*. pages 29–42.

Recasens, Marta and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1423–1432.

Recasens, Marta and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44.4:315–345.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 1–8.

Sapena, Emili, Lluís Padró, and Jordi Turmo. 2013. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, 39(4).

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Stoyanov, Veselin and Jason Eisner. 2012. Easy-first coreference resolution. In *Proceedings of the International Conference on Computational Linguistics*, pages 2519–2534.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference 6*, pages 45–52.

Wu, Fei and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 118–127.

Desambiguación Léxica y Análisis de Corpus

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 35-42 recibido 17-03-14 revisado 15-07-14 aceptado 15-07-14

Etiquetado de metáforas lingüísticas en un conjunto de documentos en español

Linguistic metaphor labelling for a Spanish document dataset

Fernando Martínez Santiago, Miguel A. García Cumbreras, Manuel C. Díaz Galiano, Arturo Montejo Ráez
Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{dofer,magc,amontejo,mcdiaz}@ujaen.es

Resumen: En este trabajo se han etiquetado manualmente las metáforas lingüísticas presentes en la colección de documentos utilizados en SemEval 2013 en la tarea correspondiente a desambiguación léxica en español. El objetivo del trabajo es doble: por un lado realizar una primera aproximación de las dificultades específicas que presenta la identificación de metáforas en el idioma español, y por otro crear un nuevo recurso lingüístico conformado por una colección documental en español que tiene etiquetadas ciertas palabras tanto con su sentido literal como metafórico.
Palabras clave: metáfora lingüística, metáfora cognitiva, recursos lingüísticos, ambigüedad léxica

Abstract: This paper introduces the work performed to manually label linguistic metaphors in the document collection of SemEval 2013, in the Spanish lexical disambiguation task. The objectives of this work are two: first, to make a prior identification of the difficulties inherent in metaphor detection in Spanish, and second, to generate a new linguistic resource as a collection of Spanish documents with certain terms label with both, the literal and the metaphoric sense.
Keywords: linguistic metaphor, cognitive metaphor, linguistic resource, lexical ambiguity

1 Introducción

La metáfora tradicionalmente es entendida como un recurso lingüístico con una finalidad principalmente artística o retórica. Sin embargo, la metáfora también es una herramienta de uso cotidiano que permite explicar un dominio conceptual en términos de otro dominio (Lakoff y Johnson, 2003). Por dominio conceptual entendemos cualquier organización coherente de experiencia. Por ejemplo, en la frase "Alicia no ve el problema" se utiliza una experiencia física (ver) para explicar una experiencia más abstracta (conocimiento). Se establece así un dominio origen y uno destino.

Una metáfora *lingüística* es la expresión lingüística de una metáfora conceptual. En el ejemplo anterior la metáfora lingüística es el verbo "ver".

Otros ejemplos de metáforas son:

(i) "Resolveremos el problema más *adelante"* (*tiempo* es *movimiento*)

En una primera clasificación podemos hablar de dos tipos de metáforas: conceptual o cognitiva y lingüística.

El proceso de aproximarnos a un dominio conceptual a partir de otro dominio mejor conocido es lo que denominamos metáfora conceptual o cognitiva, como en el ejemplo anterior. Una metáfora conceptual o cognitiva se refiere a la comprensión de una idea, o dominio conceptual, en términos de otro, denominados dominio origen y destino. Siguiendo con el ejemplo anterior, se está asimilando el proceso de ver, que es el dominio destino, con la adquisición de conocimiento, que es el dominio fuente u origen.

(ii) "Él va sin *dirección* por la vida " (*vida* es *viaje*)
(iii) "Tus declaraciones son *indefendibles"* (argumento es guerra)
(iv) "Tengo la *mente* puesta en otro lugar" (*mente* es *objeto físico*)
(v) "¿Cómo puedo *matar* un proceso?" (*proceso* es *ser vivo*)

ISSN 1135-5948

(vi) " *Parad* el *mundo*, que me *bajo*" (*mundo* es *transporte*)

En (Kövecses, 2010) se proponen otros criterios de clasificación. Quizás el más usual atiende al grado de convencionalidad de la metáfora, bien en su expresión lingüística (se expresa una metáfora conceptual ya conocida de un modo novedoso) o en los dominios involucrados (se involucran dos dominios conceptuales que es novedoso que participen en la misma metáfora). La frase (vi) es un caso de metáfora no convencional a nivel lingüisitico: es una metáfora conceptual frecuente, *vida* es un *viaje*, pero utilizada de un modo original.

Un tercer criterio de clasificación atiende a la riqueza del trasvase de información del dominio fuente al dominio destino. En los casos (i) (ii) (iii) y (v) un dominio fuente presta no solo el concepto si no su estructura al dominio destino. Por ejemplo, al asimilar *tiempo* con *movimiento* inferimos que el tiempo es un ente con una posición y movimientos determinados. Este tipo de metáforas son denominadas *estructurales*, en contraposicion a las denominadas metáforas *ontológicas*, más vagas e imprecisas. Por ejemplo, en la frase (iv) nos referimos a la *mente* como un *objeto físico*, es un modo de aproximarnos a la idea de mente. En el presente trabajo:

- se realiza un estudio de metáforas lingüísticas en español,
- se propone un método de etiquetado manual, derivado del propuesto en (Steen, y otros, 2010) y adaptado al caso del español y
- se genera un recurso a partir de la colección de documentos utilizados en SemEval 2013 (Lefever y Hoste, 2013) en la tarea correspondiente a desambiguación léxica en español. Estos documentos han sido etiquetados manualmente con los sentidos metafóricos de ciertas palabras.

Las palabras etiquetadas son los sintagmas nominales desambiguados en la colección, verbos, adjetivos y adverbios. Se trata de un recurso que cuenta con nombres desambiguados y además etiquetados según sean metafóricos o no. Los verbos, que no están desambiguados en la colección original, se han clasificado según sean metafóricos o no debido a la elevada frecuencia con que estos suelen presentar un sentido metafórico: (Cameron, 2003) estimó que el 50% de las metáforas presentes en libros de texto se corresponde con verbos.

El resto del artículo está organizado como sigue: la sección 2 presenta una breve motivación sobre el interés de la metáfora desde el punto de vista del procesamiento del lenguaje natural. En la sección 3 se describe con detalle el procedimiento de etiquetado que se ha seguido. En la sección 4 se presenta la colección documental y el grado de acuerdo entre los anotadores. En la sección 5 indicamos algunas consideraciones sobre el corpus anotado: metáforas encontradas y se discuten algunos casos particulares. El artículo concluye esbozando las líneas de trabajo futuras.

2 La metáfora en el procesamiento del lenguaje natural

La metáfora es un concepto ubicuo, cotidiano, y no siempre es un proceso creativo o novedoso. Así pues, ¿es necesario interpretarlas de un modo diferente al de expresiones no metafóricas?

En lo relativo a las metáforas convencionales, existen estudios que demuestran que incluso la comprensión de las metáforas convencionales conlleva la activación en el cerebro de las zonas que se asocian tanto con el dominio fuente como con el dominio destino (Boroditsky, 2001) (Gibbs, Bogdanovich, Sykes y Barr, 1997) (Shutova E. , 2010). Por lo tanto se puede esperar que un sistema que requiera interpretar un texto debe tener acceso a los mismos dominios conceptuales que una persona utilizaría para interpretar ese mismo texto.

Son muchas las tareas donde esta detección e interpretación de las metáforas pueden ayudar.

Por ejemplo, (Shutova, Teufel y Korhonen, 2012) encontraron que un 44% de las traducciones realizadas por Google Translate de inglés al ruso contenían algún error debido a la mala traducción de al menos un término usado metafóricamente, convencional o no. Esto mismo es fácil encontrarlo en traducciones entre inglés y español. Así en la frase "Juan soltó todo lo que sabía", la interpretación del verbo *soltar* es metafórica, pero convencional. Sin embargo se traduce erróneamente por *give* (dar).

La implicación textual (TE, del inglés *textual entailment*) o los sistemas de búsquedas de respuestas son otras tareas que podrían beneficiarse de un sistema capaz de manejar metáforas. Por ejemplo, en TE la frase "Juan va

a *explotar* si no se calma" implica a "Juan está *enfadado*". Este ejemplo parte de la metáfora conceptual "*enfado* es un *líquido caliente en un contenedor cerrado*".

Por lo tanto se requiere la aplicación del conocimiento sobre el dominio fuente para inferir que el grado de enfado es muy elevado. Este tipo de razonamientos donde se requiere metáforas estructurales con un rico traspaso de conocimiento de un dominio a otro es conocido como *implicación metafórica* (Kövecses, 2010), lo cual es posible cuando el nivel de paralelismo estructural entre el dominio fuente y destino es muy elevado.

Tal como se ha descrito previamente, en este trabajo se ha procedido a etiquetar un corpus utilizado en tareas de desambiguación del sentido de las palabras (*Word Sense Disambiguation*, WSD). La relación entre WSD y metáforas es clara en determinados sentidos de algunas palabras, los cuales tienen un origen metafórico. Tomemos como ejemplo las dos siguientes frases:

(i) "El reclamo podría obligar a los jueces de apelación a *lidiar* con la pregunta"

(ii) "Pero los abogados del señor Hayes han presentado escritos en la corte que *delinean* los argumentos…"

El primer ejemplo *lidiar* se interpreta a partir de un dominio fuente más específico para explicar un dominio más genérico, y ambos dominios están presentes en el diccionario[1]: 1. "Luchar con el toro incitándolo y esquivando sus acometidas hasta darle muerte" y 2. "Batallar, pelear". Claramente el sentido 2 es metafórico a partir del sentido 1, y es posible que un sistema WSD pueda detectar correctamente ambos sentidos. Sin embargo, en otros casos el sentido metafórico de la palabra no se corresponde con un sentido reconocido en un diccionario, por lo que deberían generarse dinámicamente nuevos sentidos acorde al contexto de la palabra. Tal es el caso del verbo "*delinear*" en la frase (ii) donde se está realizando la metáfora conceptual "*argumento es construcción*"[2]. Sin embargo, el único significado de "*delinear*" es que está recogido en el diccionario de la RAE "trazar las líneas de una figura" por lo que no sería posible establecer tal metáfora conceptual con tan solo la ayuda de diccionarios.

Finalmente, existen estudios que cuantifican que en promedio se encuentra al menos un término usado metafóricamente cada tres frases (Shutova y Teufel, 2010). Se trata pues de un fenómeno que además de requerir su identificación e interpretación en multitud de escenarios, es muy frecuente.

En resumen, la detección e interpretación de la metáfora permite establecer un paralelismo entre el dominio origen y destino. En algunos casos, esto permite inferir nuevo conocimiento en el dominio destino al aplicar en este el conocimiento que se tiene del dominio origen, incluso en el caso de metáforas convencionales. En el caso de las metáforas creativas, éstas implican significados novedosos de palabras, lo que imposibilita asignar correctamente un sentido a tales palabras con la sola ayuda de un diccionario. Finalmente, no se trata de un fenómeno lingüístico poco frecuente, sino que se utilizan en el lenguaje cotidiano, y aun en tales casos es sabido que accedemos tanto al dominio fuente como destino (Kövecses, 2010). Todo ello justifica el interés de la metáfora en el ámbito del procesamiento del lenguaje natural.

En el caso del español podemos encontrar al menos dos corpus desambiguados que han estudiado e incorporado sentidos metafóricos. Sensem (Castellón, 2012) es un corpus constituido por 100 oraciones para cada uno de los 250 verbos más frecuentes del español, etiquetado sintáctica y semánticamente. AnCora (Taulé, Martí y Recasens, 2008) es un corpus del catalán y del español con distintos niveles de anotación, desde el lema y la categoría morfológica hasta las clases semánticas de los verbos

3 Descripción del procedimiento de etiquetado de metáforas

El procedimiento de etiquetado de metáforas que proponemos está basado en MIP (*Metaphor Identification Procedure*), que fue desarrollado por (Pragglejaz Group, 2007) para el etiquetado de metáforas lingüísticas formadas por una única unidad léxica, generalmente una palabra. En el resto de esta sección se describe este procedimiento, que se ha seguido para etiquetar

[1] Sentidos extraídos del diccionario de la Real Academia Española

[2] En este ejemplo, hemos considerado construcción como dominio conceptual fuente ya que el dibujo del plano es una de las primeras fases de la construcción, de modo análogo a como la documentación presentada se entiende que constituye una de fase preliminar en elaboración del argumento de la defensa.

un corpus y se destacan algunas diferencias significativas respecto de MIP.

3.1 Procedimiento de etiquetado propuesto

A continuación se muestran los pasos definidos como procedimiento para el etiquetado de metáforas:

1. Lea todo el texto para establecer una comprensión general del significado
2. Determine las unidades léxicas en el texto
3. Establezca el significado de cada unidad léxica en el texto, teniendo en cuenta su contexto, la situación evocada por el texto (significado contextual). Tome en cuenta lo que viene antes y después de la unidad léxica
4. Para cada unidad léxica, determine si tal unidad puede tener, en algún otro contexto, un significado diferente y más *básico*. Para nuestros propósitos, significados *básicos* tienden a ser:
    a. Más concreto (lo que evocan es fácil de imaginar, ver, oír , sentir, oler y saborear)
    b. Relativo al movimiento, o a acciones corporales
    c. Más preciso (en oposición a vago)
    d. Históricamente más antiguo

    Los significados *básicos* no son necesariamente los significados más frecuentes de la unidad léxica

    En caso que efectivamente la unidad léxica pueda tener en otro contexto un significado más básico: decida si el significado encontrado en el texto puede ser interpretado a partir del significado básico.
5. En caso afirmativo, marque la unidad léxica como metafórica

Directrices adicionales:

En caso de duda en el paso 3, se recomienda buscar la unidad léxica en el diccionario de la Real Academia Española. Si la duda persiste, consultar en otros recursos como BabelNet e intentar identificar el significado contextual y el significado básico entre los provistos en el diccionario

En caso de duda en el paso 4, trate de identificar el dominio fuente y el dominio destino.

Casos particulares que debe considerar:

- La personalización es metafórica si se asigna a una cualidad/funcionalidad a parte que no le corresponde ("el corazón *rugió* enfurecido" es metáfora, "sus ojos *buscaban* los ojos del niño" es sinécdoque, no metáfora)
- Son frecuentes las expresiones multi-palabras donde una de las palabras es usada metafóricamente. Por ejemplo "efecto *invernadero*". Se deben considerar las unidades léxicas que la conforman separadamente si tal multi-palabra no se encuentra registrada en el diccionario y además el anotador manual no marca la expresión multi-palabra como unidad léxica. En otro caso, se trata de una única unidad léxica constituida por más de una palabra

Las palabras que aparecen en frases o expresiones hechas (una frase o expresión que tiene forma fija, tiene sentido figurado y es de uso común por la mayoría de hablantes de una comunidad lingüística) tal como "se marchó con las *manos vacías*" no deben, en general, considerarse metafóricas.

3.2 Diferencias entre MIP y el procedimiento de etiquetado propuesto

El método descrito difiere de MIP en algunos aspectos. En primer lugar MIP se basa sistemáticamente en diccionarios para que el anotador decida el significado base y los posibles significados metafóricos. Sin embargo, (Lönneker y Eilts, 2004) encontraron que el modo en que se incluyen significados metafóricos en diccionarios es poco sistemático, algunos sentidos aparecen otros no. Por este motivo (Shutova, Devereux y Korhonen, 2013) propone no utilizar diccionario alguno, sino confiar únicamente en el conocimiento del anotador. Posiblemente esta estrategia posibilita etiquetar significados metafóricos que no se encuentran registrados en el diccionario, pero a costa de una menor cohesión entre los anotadores. Nosotros hemos tomado una posición intermedia: al etiquetador no se le obliga a revisar cada término en un diccionario pero, para aquellos términos que le resulten dudosos, sí se le facilita un procedimiento basado en un diccionario, entre otros recursos.

Recursos utilizados. Si bien MIP no es dependiente de un idioma concreto, en su formulación original utiliza diccionarios

específicos del idioma inglés. En nuestro caso, estos recursos se han sustituido por el diccionario de la Real Academia del Español y BabelNet.

Uso de dominios cognitivos. Aunque en el presente trabajo sólo se etiquetan metáforas lingüísticas, detectar un significado base y un significado metafórico conlleva, implícitamente, reconocer que existen dos dominios cognitivos bien diferenciados aun sin llegar a identificarlos. En casos dudosos hemos encontrado útil tratar de identificar explícitamente los dominios fuente y destino. Para ello se facilita a los anotadores dos listas con los dominios fuente y destino más usuales, como las mostradas previamente. Estas listas se han adaptado de las propuestas en (Kövecses, 2010).

Finalmente, el procedimiento propuesto va acompañado de un conjunto de reglas que deben aplicarse en algunos casos particulares, pero relativamente frecuentes. Se ha seguido el criterio que se describe para cada uno de esos casos en (Steen, y otros, 2010). Solo en el caso de las expresiones multi-palabra hemos tenido que seguir un criterio ligeramente diferente. Mientras que para MIP una expresión multi-palabra es una única unidad léxica solo si esta se encuentra en el diccionario, nosotros hemos ampliado este supuesto al caso que el analizador sintáctico que han utilizado marque la expresión multi-palabra como tal. En definitiva, el criterio que se sigue con las expresiones multi-palabra y las expresiones hechas es el mismo: solo se marca metafóricamente sus constituyentes si la expresión completa es utilizada metafóricamente.

4 Descripción y etiquetado de la colección documental

4.1 La colección documental: metáfora y desambiguación

La colección documental etiquetada se corresponde con la colección de textos suministrada para la evaluación de sistemas en la tarea *Multilingual Word Sense Disambiguation* correspondiente al *SemEval 2013* (Lefever y Hoste, 2013). Es una colección de 13 textos relativos a noticias con un total de 306 oraciones conformados por 8779 palabras, de las cuales 1198 son verbos y 1481 están desambiguadas léxicamente. Estos textos provienen del corpus paralelo Europarl, y fueron desambiguados y lematizados manualmente por los organizadores de la tarea.

Se ha seleccionado esta colección porque existe una clara relación entre ambigüedad léxica y metáfora. Por ejemplo, MIP utiliza los sentidos registrados en diccionarios para diferenciar el sentido metafórico y sentido básico. En consecuencia MIP se puede interpretar como un procedimiento de desambiguación manual con especial énfasis en metáfora (Shutova y Teufel, 2010). La diferencia clave entre ambas tareas es que, mientras en el caso de la ambigüedad léxica se parte de un conjunto predefinido de sentidos, en el caso de la metáfora estos significados deben ser inducidos automáticamente lo que impide utilizar un enfoque tradicional de desambiguación léxica en el ámbito de la metáfora lingüística (Shutova, Teufel y Korhonen, 2012), (Pustejovsky, 1995). Sin embargo, a pesar de la estrecha relación entre ambigüedad léxica y metáfora, hasta donde nuestro conocimiento llega no existe ningún recurso, en español u otro idioma, que etiquete los términos en ambas vertientes.

4.2 Esquema de anotación

El corpus ha sido anotado por tres personas que tienen el español como idioma nativo, pero sin conocimientos específicos relativos a metáforas conceptuales. Por ello previamente se les entrenó a través de algunos ejemplos. Luego se les facilitaron las instrucciones para el procedimiento de etiquetado descrito en la sección 3.1. Con la finalidad de comprobar que se había asimilado correctamente la tarea, se les solicitó que etiquetaran cinco frases de ejemplo siguiendo tal procedimiento y se revisó el resultado de forma conjunta. Finalmente se facilitó la colección documental con las palabras resaltadas a etiquetar como metafóricas o no. Del total de 8779 palabras del corpus, sólo se han anotado nombres y verbos, de tal modo que cada anotador ha etiquetado 2679 unidades léxicas.

Dado que la tarea propuesta es inherentemente subjetiva, es esencial cuantificar el acuerdo existente entre los anotadores, esto es, la similitud de las anotaciones producidas por diferentes anotadores. Se evaluó la confiabilidad del sistema de anotación propuesto en términos del valor estadístico κ (Siegel y Castellan, 1988) obteniendo un valor $\kappa = 0{,}62$ ($n = 2$; $N = 2679$;

k = 3), donde n representa el número de categorías (metafórico/no metafórico), N para el número de casos anotados y k el número de anotadores. Este nivel de acuerdo es considerado sustancial (Landis y Koch, 1977).

5 Análisis de los resultados

En el presente trabajo se ha seguido el criterio de que para que un término sea considerado metafórico debe haber consenso de al menos dos anotadores. De este modo, se ha encontrado que 113 de las 306 oraciones contienen al menos un término metafórico, lo que supone que se encuentra al menos una metáfora cada tres oraciones aproximadamente. Este resultado es consistente con (Shutova y Teufel, 2010).

	Metafóricos	Literales	Total
Verbos	110	1088	1198
Nombres	176	1305	1481
Total	286	1393	2679

Tabla 1. Metáforas anotadas

En cuanto al número de verbos y nombres etiquetados metafóricamente (Tabla 1), se han contado 110 verbos utilizados metafóricamente, lo que supone un 10,9% del total. En el caso de los nombres, en 176 casos han sido utilizados metafóricamente (8,4%). Esto supone que un 39% del total de las metáforas tiene origen verbal. Este porcentaje es sensiblemente inferior al 50% que apunta (Cameron, 2003), si bien el motivo puede ser que mientras el trabajo de Cameron está centrado en textos educacionales, el presente trabajo parte de un colección de textos cuyo origen son noticias publicadas en diversos medios de comunicación. Además, la restricción impuesta de consenso entre dos anotadores resta necesariamente el número total de metáforas anotadas finalmente.

5.1 Consideraciones adicionales

5.1.1 Metáforas novedosas o creativas

A los anotadores también se les solicitó que indicaran aquellas metáforas que son nóveles como tales esto es, metáforas cuyo significado no es posible encontrar en un diccionario. Se etiquetaron tan sólo 13 metáforas de este tipo, menos del 5% del total. Aun siendo un número bajo, una cuestión que surge con las metáforas novedosas es cómo actualmente se están marcando tales palabras en las colecciones textuales desambiguadas manualmente. Esta cuestión es relevante puesto que guía el modo en que los sistemas de desambiguación son evaluados. Por ejemplo, en la colección utilizada en el presente trabajo, encontramos el siguiente ejemplo:

(i) "Al principio de su mandato, Nicolas_Sarkozy quería ir a buscar los puntos de crecimiento que le faltaban con los *dientes*"

En la frase precedente, "*dientes*" es marcado como metáfora creativa, ya que no es posible encontrar en un diccionario el significado que se le atribuye en esta frase. Además, dientes no es metonimia ya que aunque estén representando a la persona, no se usan del modo que en la frase se recoge. Es por lo tanto un significado novedoso el que se le está atribuyendo a la palabra y sin embargo tal palabra está etiquetada manualmente con el sentido más usual de "*dientes*", lo cual desde nuestro punto de vista es erróneo: un sistema que interpretara la palabra con el significado literal nunca podría interpretar correctamente el significado real de la frase.

Una situación similar se encuentra en la metonimia y la sinécdoque, ya que las utilizamos para evocar un concepto a partir de otro concepto, si bien se diferencian en el mecanismo que activa el concepto destino. Es frecuente encontrar ejemplos de metonimias en textos informativos. De nuevo es posible encontrar en el diccionario significados que tienen su origen en la metonimia, pero otras son novedosas. Por ejemplo:

(ii) "…manifestó el *portavoz* del Ministerio de Petróleo Assam Jihad"

El significado de "*portavoz*" tiene origen metonímico, pero tal significado está reflejado en el diccionario por lo que cabe la desambiguación correcta del término. Sin embargo, dada la frase:

(iii) "…el asesinado deja cuatro *bocas* que alimentar "

El término "*bocas*" es una sinécdoque de los hijos, y tal significado no se encuentra en el diccionario. Entonces ¿cómo debería desambiguarse tal término?

5.1.2 Significados base inusuales

En ocasiones es posible encontrar significados muy precisos y concretos, pero muy poco conocidos, o bien muy específicos de determinados ámbitos. En tales casos esos

significados no se han considerado como significados base: aquellas cuyo significado base está ya en desuso y por lo tanto no se concluye que la palabra sea utilizada metafóricamente. Ejemplos son:

(iv) "La *expulsión* presuntamente urdida por dos jugadores del Real_Madrid"

(v) "El texto, que podría proporcionar la base para un *acuerdo* político final"

En la frase (iv) el término *expulsión* no es metafórico a pesar de que *expulsión* también significa "Golpe que da el diestro sacudiendo violentamente con la fuerza de su espada la flaqueza de la del contrario, para desarmarlo", lo cual es un significado más preciso, pero generalmente desconocido. En el caso de la frase (v), *acuerdo* tampoco es metáfora. *Acuerdo* en el sentido de "*templar instrumentos musicales*" es el significado base porque es un significado más concreto y además, posiblemente más antiguo. Pero aun así no se le da sentido metafórico al *acuerdo* del ejemplo (frase v) porque el significado más concreto y/o antiguo de la palabra es prácticamente desconocido. Lo que subyace en estos dos ejemplos es que, para que haya metáfora, deben ser conocidos los dos sentidos de la palabra (en el caso de la metáfora creativa uno de los dos sentidos no se conoce de antemano, pero se infiere a partir del sentido base y del contexto).

5.1.3 Multi-palabras y unidades léxicas

Algunos términos utilizados metafóricamente no han podido etiquetarse como tales porque en la colección documental aparecían como parte de una expresión multi-palabra. Por ejemplo la expresión *economía emergente* es marcada como una expresión multi-palabra en la colección documental, y por lo tanto debe evaluarse como una única unidad léxica. Sin embargo, *economía emergente* no se encuentra en el diccionario de la RAE ni en BabelNet, lo que habría permitido evaluar *emergent*" como una unidad léxica y, consecuentemente, como una metáfora. La solución que se ha adoptado es ampliar la definición de unidad léxica en el caso de multi-palabras: una multi-palabra es una única unidad léxica si se encuentra recogida en el diccionario o bien es marcada como tal por el analizador sintáctico. Este mismo criterio se ha seguido en el caso de las expresiones hechas.

Por ejemplo en la frase (vi) la unidad léxica *mano* no es marcada como metafórica, ya que está siendo utilizada del modo usual en una expresión bien conocida (*ir de la mano*)

(vi) "Con esto los institutos pretenden liberarse de las restricciones que habían *ido de la mano* con la aceptación del dinero"

5.1.4 Personalización

Siguiendo el enfoque propuesto en MIP, la personalización no se ha considerado en general metáfora, salvo que la metáfora sea en relación a la persona o personas referidas mediante la personalización. Esto es, primero se interpreta la personalización, y luego se evalúa si hay metáfora. Por ejemplo:

(vii) "La *ONU esboza* un plan para reducir las emisiones"

(viii) "La_*Unión_Europea* también *estimuló* las conversaciones"

En (vii) el término *ONU* no se refiere a la ONU, sino a algunas personas que pertenecen a esa organización, por lo tanto ni "*ONU*" ni *esbozar* constituye metáfora. En (viii) *Unión_Europea* es un caso similar, pero una vez asumido que se refiere a las personas y no a la organización, sí cabe interpretar *estimuló* de un modo metafórico ya que este término tiene un significado base más específico ("aguijonear, picar, punzar" en comparación con "incitar, excitar con viveza la ejecución de algo").

6 Conclusiones y trabajo futuro

La metáfora es un concepto resbaladizo y ubicuo; quizás por ello es difícil encontrar trabajos relativos al etiquetado, identificación e interpretación de la metáfora en el caso del idioma español. Sin embargo, es un fenómeno lingüístico que debe ser abordado en aquellas tareas que requieran acceder al significado del texto. En este trabajo se presenta un método de etiquetado manual de metáforas para el caso del español, basado en MIP. Tal método ha sido validado sobre una colección documental con verbos y nombres etiquetados según sean usados con un sentido metafórico o no. El acuerdo alcanzado entre los anotadores es considerado sustancial y demuestra la validez del método propuesto. Además, los nombres etiquetados también están desambiguados, lo cual permite en un futuro evaluar el desempeño de los sistemas de desambiguación automática en aquellos términos que son usados con un

sentido metafórico. Como trabajo futuro también se propone aplicar el método de etiquetado manual propuesto en otras colecciones documentales, con la finalidad de estudiar la incidencia de la metáfora en tareas donde el acceso al significado sea relevante, tales como la implicación textual o el análisis de sentimientos.

Agradecimientos

Esta investigación ha sido subvencionada por el proyecto ATTOS (TIN2012-38536-C03-01) financiado por el Gobierno de España y el proyecto europeo FIRST (FP7-287607).

Bibliografía

Boroditsky, L. 2001. Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology, 43*, 1-22.

Cameron, L. 2003. *Metaphor in Educational Discourse.* London: Continuum.

Castellón, I. 2012. Constitución de un corpus de semántica verbal del español: Metodología de anotación de núcleos argumentales. *RLA*, 13-38.

Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R. y Barr, D. J. 1997. Metaphor in Idiom Comprehension. *Journal of memory and language, 37*, 141-154.

Kövecses, Z. 2010. *Metaphor.* Oxford University Press; 2nd edition.

Lakoff, G. y Johnson, M. 2003. *Metaphors We Live By.* University Of Chicago Press; 2nd edition.

Landis, J. y Koch, G. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lefever, E. y Hoste, V. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (págs. 158-166). Atlanta, Georgia, USA: Association for Computational Linguistics.

Lönneker, B. y Eilts, C. 2004. A current resource and future perspectives for enriching wordnets with. *Proceedings of the second international wordnet conference (GWC 2004)*, (págs. 157-162). Brno.

Martin, J. 1990. *A Computational Model of Metaphor Interpretation.* San Diego, CA, USA: Academic Press.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*, 1-39.

Pustejovsky, J. 1995. *The Generative Lexicon.* Cambridge, MA: MIT Press.

Shutova, E. 2010. Models of Metaphor in NLP. *Proceedings of ACL 2010*.

Shutova, E. y Teufel, S. 2010. Metaphor corpus annotated for source-target domain mappings. *Proceedings of LREC 2010*, (págs. 3255–3261). Malta.

Shutova, E., Devereux, B. y Korhonen, A. 2013. Conceptual metaphor theory meets the data: a corpus-based human annotation study. *Language Resources & Evaluation, 47*(4), 1261-1284.

Shutova, E., Teufel, S. y Korhonen, A. 2012. Statistical Metaphor Processing. *Computational Linguistics, 39*(2).

Siegel, S. y Castellan, N. J. 1988. *Nonparametric statistics for the behavioral sciences.* New York, USA: McGraw-Hill Book Company.

Steen, G., Dorst, A., Hermann, J., Kaal, A., Krennmayr, T. y Pasma, T. 2010. *A Method for Linguistic Metaphor Identificaction.* John Benjamins Publishing Company.

Taulé, M., Martí, M. y Recasens, M. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of the 6th conference on International Language Resources and Evaluation*, 96-101.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 43-50 recibido 08-04-14 revisado 06-06-14 aceptado 13-06-14

Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit*

Metodología y evaluación de la expansión del WordNet del gallego con WN-Toolkit

Xavier Gómez Guinovart
Universidade de Vigo
Vigo, Galiza
xgg@uvigo.es

Antoni Oliver
Universitat Oberta de Catalunya
Barcelona, Catalunya
aoliverg@uoc.edu

Resumen: En este artículo se presenta la metodología utilizada en la expansión del WordNet del gallego mediante el WN-Toolkit, así como una evaluación detallada de los resultados obtenidos. El conjunto de herramientas incluido en el WN-Toolkit permite la creación o expansión de wordnets siguiendo la estrategia de expansión. En los experimentos presentados en este artículo se han utilizado estrategias basadas en diccionarios y en corpus paralelos. La evaluación de los resultados se ha realizado de manera tanto automática como manual, permitiendo así la comparación de los valores de precisión obtenidos. La evaluación manual también detalla la fuente de los errores, lo que ha sido de utilidad tanto para mejorar el propio WN-Toolkit, como para corregir los errores del WordNet de referencia para el gallego.
Palabras clave: WordNet, adquisición de información léxica, corpus paralelos, recursos plurilingües

Abstract: In this paper the methodology and a detailed evaluation of the results of the expansion of the Galician WordNet using the WN-Toolkit are presented. This toolkit allows the creation and expansion of wordnets using the *expand* model. In our experiments we have used methodologies based on dictionaries and parallel corpora. The evaluation of the results has been performed both in an automatic and in a manual way, allowing a comparison of the precision values obtained with both evaluation procedures. The manual evaluation provides details about the source of the errors. This information has been very useful for the improvement of the toolkit and for the correction of some errors in the reference WordNet for Galician.
Keywords: WordNet, lexical acquisition, parallel corpora, multilingual resources

1 Introduction

WordNet (Fellbaum, 1998) is a lexical resource where nouns, verbs, adjectives and adverbs are organised in sets of synonyms called *synsets*. In this resource, synsets are connected by semantic relations as hiponymy, antonomy, meronomy, troponomy, etc. The original WordNet was created for English in the Princeton University and nowadays there are WordNet versions for several languages. In the website of the Global WordNet Association[1] a list of existing wordnets are available. Some of these wordnets hold a free license and in the Open Multilingual WordNet project (Bond and Kyonghee, 2012) they are published under a common format.

Two general methodologies are available for WordNet creation (Vossen, 1998):

- The *merge model*, that implies the creation of a new ontology for the target language.
- The *expand model*, where the variants associated with the Princeton WordNet synsets are translated using different strategies.

In our experiments we are using a set of tools based on the expand model, that is, we are translating English variants using several strategies.

* This research has been carried out thanks to the Project SKATeR (TIN2012-38584-C06-01 and TIN2012-38584-C06-04) supported by the Ministry of Economy and Competitiveness of the Spanish Government

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

ISSN 1135-5948

Several other wordnets have been developed using the expand model. The Spanish WordNet (Atserias et al., 1997) and the Catalan WordNet (Benítez et al., 1998) were constructed using this model and they have also been expanded using the WN-Toolkit, which is described below in section 3. The expand model has been also used in the MultiWordNet project for Italian (Pianta, Bentivogli, and Girardi, 2002), the Indonesian WordNet (Putra, Arfan, and Manurung, 2008), the Hungarian WordNet (Miháltz et al., 2008), the Croatian WordNet (Raffaeli et al., 2014), the French WOLF WordNet (Sagot and Fišer, 2008) and more recently for the French WoNeF WordNet (Pradet, de Chalendar, and Desormeaux, 2014) and the KurdNet for Kurdish (Aliabadi, Ahmadi, and Salavati, 2014).

2 The Galnet project

The aim of the Galnet project (Gómez Guinovart et al., 2011) is building a WordNet for Galician aligned with the ILI (the inter-lingual index –namely, a list of meanings that allows a mapping of concepts of different languages) (Vossen, 1998) generated from the English WordNet 3.0 and with a lexical coverage similar to the English WordNet. The development of Galnet is integrated in the framework of the Multilingual Central Repository[2] (MCR) (González Agirre and Rigau, 2013). The MCR integrates in the same EuroWordNet framework wordnets from five languages (English, Spanish, Catalan, Basque and Galician) interlingually connected by the ILI and semantically categorized with several ontologies and taxonomies –IRST-Domains (Bentivogli et al., 2004), Suggested Upper Model Ontology (Pease, Niles, and Li, 2002), and Top Concept Ontology (Alvez et al., 2008). Thus, the MCR is a multilingual semantic resource of broad range suitable for use in language processing tasks that require large amounts of multilingual knowledge.

Galnet is distributed under a Creative Commons license CC BY 3.0[3] as part of the MCR. The version of Galnet included in that distribution, reaches a lexical coverage of about one fifth of the English WordNet 3.0 (WN30) variants, as shown in detail in Table 1.

[2] http://adimen.si.ehu.es/web/MCR/

[3] http://creativecommons.org/

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	117798	82115	18949	14285
V	11529	13767	1416	612
Adj	21479	18156	6773	4415
Adv	4481	3621	0	0
Total	155287	117659	27138	19312

Table 1: Galnet current distribution

This early version of Galnet includes the Galician translation of the nominal and verbal synsets belonging to a set of basic concepts defined for WordNet, the Basic Level Concepts (BLC) (Izquierdo, Suárez, and Rigau, 2007), namely, 649 nominal synsets and 616 verbal synsets grouped in the freqmin20/all[4] folder in the official distribution of the BLC for WordNet 3.0.[5] This version of Galnet also includes the Galician entries for the WordNet lexicographer files (Fellbaum, 1998) corresponding to the names denoting body parts (noun.body) and substances (noun.substance), and the Galician equivalents for the adjectives of general type (adj.all).[6] Finally, we extended the lexical coverage of this early version of Galnet using the WN-Toolkit (Oliver, 2012) to expand Galnet from two existing bilingual English–Galician resources, the Wikipedia and the English–Galician CLUVI Dictionary,[7] reaching the final coverage shown in Table 1 (Gómez Guinovart et al., 2013). From this first distribution of Galnet we apply new expansions by means of lexical extraction from a Galician thesaurus (Gómez Guinovart and Simões, 2013) and using the WN-Toolkit, which is the aim of this paper.

Both the current distribution of Galnet and its development version can be explored through a specific web query interface for Galnet,[8] or together with other lexical and textual resources for Galician through the RILG (Integrated Language Resources for Galician) platform.[9]

[4] The BLC in that set represent at least a number of synsets equal than 20, and have been obtained getting into account all relations of the synsets.

[5] http://adimen.si.ehu.es/web/BLC/

[6] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html

[7] http://sli.uvigo.es/dicionario/

[8] http://sli.uvigo.es/galnet/

[9] http://sli.uvigo.es/RILG/

3 The WN-Toolkit

The WN-Toolkit (Oliver, 2014) is a set of programs written in Python for the creation of wordnets following the expand model. At the moment no user interface is provided so all programs must be run in a command line. The toolkit also provides some free language resources. This language resources are pre-processed so they can be easily used with the toolkit.

The toolkit is divided in the following parts:

- Dictionary-based strategies
- Babelnet-based strategies (Navigli and Ponzetto, 2012)
- Parallel corpus-based strategies
- Resources, such as freely available lexical resources, pre-processed corpora, etc.

The toolkit is distributed under the GNU-GPL license version 3.0 and can be freely downloaded from `http://lpg.uoc.edu/wn-toolkit`.

This toolkit has been developed under the SKATeR project and it has been previously successfully used for the expansion of Catalan and Spanish wordnets.

4 Experimental settings and automatic evaluation

4.1 Experimental settings

In the experiments with Galnet presented in this paper we have used the following strategies and resources:

- Dictionary-based strategies
  - Apertium English–Galician dictionary[10]
  - English–Galician Wiktionary[11]
- Babelnet-based strategies
  - Using Babelnet 2.0[12]
- Parallel corpus-based strategies
  - Machine translation of sense-tagged corpora (Oliver and Climent, 2012)
    - English–Galician Semcor Corpus.[13] The translation has been performed using Google Translate.[14]
  - Automatic sense-tagging of parallel corpora (Oliver and Climent, 2014). Only the English part of the parallel corpora has been sense-tagged using Freeling with the UKB word sense disambiguator (Padró et al., 2010).
    - Unesco CLUVI[15] Corpus of Spanish–Galician scientific-technical texts
    - Lega CLUVI Corpus of Galician–Spanish legal texts
    - Consumer Eroski CLUVI Corpus of Spanish-Galician texts
    - Tectra CLUVI Corpus of English–Galician literary texts

4.2 Automatic evaluation

In Table 2 we can observe the precision and number of new variants obtained with each method. The evaluation has been performed in an automatic way, comparing the obtained variants with the existing variants in the current distribution of Galnet. If the variant obtained for a given synset is one of the variants in the same synset of the existing Galnet, the result is evaluated as correct. If we do not have any Galician variant for a given synset in the reference Galnet, this result is not evaluated. The automatically obtained precision values tend to be lower than real values. The reason is that sometimes we have one or more for a given synset in the reference Galnet, but the obtained variant is not present. If the obtained variant turns out to be correct, it will be evaluated as incorrect anyway.

4.3 Getting variants from several sources

After this first analysis of the results we have evaluated the precision of the extracted variants taking into account the number the number of sources contributing the same entry. In Table 3 we can observe the precision of the results (from automatic evaluation) for the variants obtained in several experiments. We can also observe the precision for those variants obtained in a single experiment.

[10] `http://sourceforge.net/projects/apertium/`
[11] `http://www.wiktionary.org`
[12] `http://babelnet.org`
[13] `http://www.gabormelli.com/RKB/SemCor_Corpus/`
[14] Thanks to the University Research Program for Google Translate.
[15] `http://sli.uvigo.es/CLUVI/`

	Precision	New variants
Apertium	78,83	1.230
Wiktionary	80,91	744
Babelnet	83,29	4.794
Semcor	78,13	2.053
Unesco	80,84	2.150
Lega	77,42	1.172
Eroski	80,28	1.777
Tectra	82,74	948

Table 2: Precision values and number of new variants obtained in the experiments

	Precision	New variants
7	100	1
6	97,22	5
5	96,26	41
4	97,71	178
3	83,68	646
2	81,44	1.159
1	77,02	9.650

Table 3: Precision values and number of new variants obtained in relation with the number of experiments leading to the same variant

As expected, with very few exceptions, the higher the number of experiment leading to a variant the higher the value of precision. It is also important to keep in mind that the value of precision here is calculated in an automatic way, comparing the results with the reference WordNet for Galician.

5 Errors analysis and revision

Human revision of the results of the experiment have been done not only for the new variants found for synsets without previous Galician variants in the reference Galnet (candidates automatically not evaluated), but also for the variants found for synsets which already have some Galician variant in the Galnet (candidates automatically evaluated as incorrect). The next two subsections show the detais in both these cases.

5.1 New variants for empty synsets

For the human revision of the data we have a text file (noeva-1, noeva-2, noeva-3, etc.) for each set of candidate variants, grouped by the number of experiments which lead to them, as shown in Table 3. Each line contains a proposal of new candidate variant as a sequence of offset, Galician proposed variant, English variants, and gloss, as in the following example:

02936714-n gaiola —cage, coop —an enclosure made or wire or metal bars in which birds or animals can be kept

Lexicographers have reviewed the files line by line from the noeva-1 to the noeva-7. When there were doubts about the correctness of the proposed Galician variant, the review method involved (1) commenting with hashes the line in the noeva file and (2) generating a report containing (a) the wrong line commented, (b) another commented line with the hypothetical cause of the error, and (c) the line with the error corrected. In this way we can also profit the suggestion to extend the Galnet, and also we justify and explain the cause of the error. The reports are saved in a separate file (noeva-0).

For instance, there is a wrong proposal in noeva-1 which states:

02522399-n brincadeira —cod, codfish —major food fish of Arctic and cold-temperate waters

But in Galician a *brincadeira* is a jibe or a joke, and never a cod, as is the meaning of this synset. So we comment this line in noeva-1 and create three new lines in the report file noeva-0, including as the third line the manually corrected proposal:

#02522399-n brincadeira —cod, codfish —major food fish of Arctic and cold-temperate waters

##Polysemy: "cod" in English can be translated in Galician in another sense by "brincadeira". The wrong sense was chosen.

02522399-n bacallau —cod, codfish —major food fish of Arctic and cold-temperate waters

In this way, we document the error of the extraction, remark its cause (English lexical polisemy), and manually create a right proposal for file export to Galnet.

While typology of errors is very varied, the errors in extraction implying capitalization are very frequent. From 173 errors identified, 89 imply an error in the use of capital letters, for instance:

#09034967-n dar_es_salaam —Dar_es_Salaam, capital_of_Tanzania —the capital and largest port city of Tanzania on the Indian Ocean

##Letter case

09034967-n Dar_es_Salaam —Dar_es_Salaam,

capital_of_Tanzania —the capital and largest port city of Tanzania on the Indian Ocean

The errors related to the Spanish source of the current distribution of MCR are another frequent cause of errors in extraction, accounting for 40 error cases, for instance:

#07410207-n cinto —knock, bash, bang, smash, belt —a vigorous blow

##Spanish bad or dubious equivalent

07410207-n golpe —knock, bash, bang, smash, belt —a vigorous blowy

In this proposal, the Galician variant *cinto* has been extracted by the WN Toolkit from its alignment in the processed Spanish-Galician resources with Spanish variant *cinturón*, which is present in the synset by error (in fact, it is an alternative translation of English *belt*).

The third most frequent cause of errors in extraction (16 cases) is the bad selection of meaning from polysemous variants in English (as in the previous example of *cod*) or Spanish, for instance:

#03365592-n solo —floor, flooring —the inside lower horizontal surface (as of a room, hallway, tent, or other structure)

##Polysemy: "suelo" in Spanish can be translated in Galician in another sense by "solo". The wrong sense was chosen.

03365592-n chan —floor, flooring —the inside lower horizontal surface (as of a room, hallway, tent, or other structure)

That being said, the results of human evaluation of the new variants extracted for synsets without previous Galician variants in the reference Galnet, in comparison with their automatic evaluation, are shown in Table 4. Due to time limitations, human review of noeva-1 has been limited by now to the 200 first lines.

5.2 New variants for not empty synsets

The new variants extracted for synsets which already have some Galician variant in the Galnet –and automatically evaluated as *incorrect* candidates– are stored in a set of text files (incorrect-1, incorrect-2, incorrect-3...) grouped by the number of experiments which lead to them, as shown in Table 5.

Each line contains a sequence of offset, Galician proposed variant, Galician variants already in the synset, English variants, and gloss, as in the following example:

	AP	CV	WC	RP
7	100	1	0	100
6	97,22	5	0	100
5	96,26	41	1	97,56
4	97,71	178	0	100
3	83,68	646	18	97,21
2	81,44	1.159	94	91,89
1	77,02	9.650	60	70

Table 4: Human evaluation of candidate new variants for empty synsets (AP = automatic precision, CV = candidate variants, WC = wrong candidates, RP = real precision)

	Variants
6	1
5	4
4	7
3	80
2	187
1	2.053

Table 5: Number of candidate variants for non-empty synsets obtained in relation with the number of experiments leading to the same variant

14541852-n risco —perigo —hazard, jeopardy, peril, risk, endangerment —a source of danger; a possibility of incurring loss or misfortune

Lexicographers have also reviewed the files from the incorrect-1 to the incorrect-6. When a wrong proposed Galician variant is found, the review method involved (1) commenting with hashes the line in the incorrect file, (2) generating a report in a separate file (incorrect-0) containing (a) the wrong line commented, (b) another commented line with the hypothetical cause of the error, and (c) if possible, the line with the error corrected, and (3) generating a report in a separate file (modify-0) with the corrections needed in the existing variants (or examples) of the synset.

For instance, there is a wrong proposal in incorrect-3 which states:

08586825-n sede —sé —see —the seat within a bishop's diocese where his cathedral is located

But, differently from Spanish *sede* –a polysemous word which means see (as in 'the Holy See'), venue or headquarters–, Galician

sede means only venue or headquarters, not see, as is the meaning of this synset. In fact, the Galician word for *see* is the yet existing Galician variant *sé*. So we comment this line in incorrect-3 and create two new lines in the report file incorrect-0:

#08586825-n sede —sé —see —the seat within a bishop's diocese where his cathedral is located

##Polysemy: "sede" in Spanish can be translated in Galician in another sense by "sede". The wrong sense was chosen.

While typology of errors is varied, the three causes most frequent are again the polysemy of the English or Spanish lexical source, the bad or dubious Spanish source, and the bad choice of letters case. Nevertheless, a characteristic feature of these errors in the new variants extracted for not empty synsets is that often they indicate an error in the existing variants of the distribution version of Galnet.

For instance, there is a correct proposal in incorrect-3 which states:

05320899-n oído —orella —ear —the sense organ for hearing and equilibrium

With the revision of this proposal, lexicographers can discover that there is a bad existing Galician variant *orella* for this synset in the reference Galnet (*orella* doesn't mean 'the sense organ for hearing and equilibrium' but 'the externally visible cartilaginous structure of the external ear'). In that case, the review protocol implies (1) not commenting the line in incorrect-3, which implies including the new Galician variant *oído* in the extended version of Galnet; and (2) copying the line, such as it is, in the file modify-0, which implies deleting the existing Galician variant *orella* from this synset in the extended version of Galnet.

All in all, the results of human evaluation of the new variants extracted for synsets with previous Galician variants in the reference Galnet, are shown in Table 6. Due to time limitations, human review of incorrect-1 has been limited by now to the 100 first lines.

5.3 Galnet expansion results

After the revision of errors, we have incorporated the new variants and the required modifications of synsets to Galnet, obtaining the results shown in Table 7.

	CV	WC	RP	SM
6	1	0	100	0
5	4	0	100	0
4	7	0	100	3
3	80	14	85,10	20
2	187	22	88,23	22
1	2.053	47	53	11

Table 6: Human evaluation of candidate new variants for not empty synsets (CV = candidate variants, WC = wrong candidates, RP = real precision, SM= suggested modifications of synsets in reference Galnet)

	Syns	Vars	Unique vars
Reference GN	19.312	27.138	23.125
Extended GN	21.509	29.687	24.661
Δ	2.197	2.549	1.536

Table 7: Galnet expansion results

Both the reference Galnet (its current distribution in the MCR) and the extended Galnet (the work in process) can be explored through the interface at `http://sli.uvigo.es/galnet/`, where the results of the expansion of Galnet with the WN-Toolkit can be viewed selecting "wnt7" as experiment in the query of the development version.

6 Conclusions and future work

In this paper we have presented a practical application of multilingual resources exploitation for lexical acquisition. We have discussed the efficiency of a tool for lexical extraction as the WN-Toolkit in extending Galician WordNet coverage from bilingual resources of Galician in combination with English and Spanish, including lexical resources such as the dictionaries of Apertium, Wiktionary and Babelnet, and textual resources such as the CLUVI and SemCor corpora.

The precision of the extraction has reached high levels of efficiency in obtaining new variants coming from two or more bilingual resources: 91,89% for candidates to new variants for empty synsets and 88,23% for candidates to new variants for synsets not empty. Parallel corpora based strategies have the problem of a very low recall, mainly due to the use of a very simple alignment algorithm based on the most frequent translation. In future experiments, we will try to widen the coverage of the results im-

proving the WN-Toolkit with new alignment algorithms yielding greater coverage such as those used by Giza++ (Och and Ney, 2003).

References

Aliabadi, Purya, Mohamed Sina Ahmadi, and Kyumars Sheykh Salavati, Shahin adn Esmaili. 2014. Towards building kurdnet, the kurdish wordnet. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia.

Alvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Antoni Oliver, and German Rigau. 2008. Consistent annotation of eurowordnet with the top concept ontology. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Atserias, Jordi, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingual WordNets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, pages 101–108, Ginebra.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. 1998. Methods and Tools for Building the Catalan WordNet. In *In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages.*

Bond, Francis and Paik Kyonghee. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th International Global WordNet Conference*, pages 64–71, Matsue, Japan.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database.* The MIT press.

Gómez Guinovart, Xavier, Xosé María Gómez Clemente, Andrea González Pereira, and Verónica Taboada Lorenzo. 2011. Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1):61–67.

Gómez Guinovart, Xavier, Xosé María Gómez Clemente, Andrea González Pereira, and Verónica Taboada Lorenzo. 2013. Sinonimia e rexistros na construción do WordNet do galego. *Estudos de lingüística galega*, 5:27–42.

Gómez Guinovart, Xavier and Alberto Simões. 2013. Retreading dictionaries for the 21st century. In José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors, *2nd Symposium on Languages, Applications and Technologies*, pages 115–126, Saarbrücken. Dagstuhl Publishing.

González Agirre, Antoni and German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.

Izquierdo, Rubén, Armando Suárez, and German Rigau. 2007. Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07)*, Borovetz, Bulgaria.

Miháltz, M., C. Hatvani, J. Kuti, G. Szarvas, J. Csirik, G. Prószéky, and T. Váradi. 2008. Methods and results of the Hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference. GWC*, pages 387–405, Szeged, Hungary.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Oliver, A. and S. Climent. 2012. Building wordnets by machine translation of sense tagged corpora. In *Proceedings of the Global WordNet Conference, Matsue, Japan.*

Oliver, Antoni. 2012. WN-Toolkit: un toolkit per a la creació de wordnets a partir de diccionaris bilingües. *Linguamática*, 4(2):93–101.

Oliver, Antoni. 2014. Wn-toolkit: Automatic generation of wordnets following the

expand model. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia.

Oliver, Antoni and Salvador Climent. 2014. Automatic creation of wordnets from parallel corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Padró, L., S. Reese, E. Agirre, and A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and UKB. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Pease, Adam, Ian Niles, and John Li. 2002. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, Edmonton.

Pianta, E., L. Bentivogli, and C. Girardi. 2002. MultiWordNet. developing an aligned multilingual database. In *1st International WordNet Conference*, pages 293–302, Mysore, India.

Pradet, Quentin, Gaël de Chalendar, and Jaume Baguenier Desormeaux. 2014. Wonef, an improved, expanded and evaluated automatic french translation of wordnet. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia.

Putra, D. D, A. Arfan, and R. Manurung. 2008. Building an Indonesian WordNet. In *Proceedings of the 2nd International MALINDO Workshop*.

Raffaeli, Ida, Bekavac Božo, Željko Agić, and Marko Tadić. 2014. Building croatian wordnet. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Sagot, Benoît and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex*.

Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Springer.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 51-58 recibido 11-04-14 revisado 30-06-14 aceptado 02-07-14

An Unsupervised Algorithm for Person Name Disambiguation in the Web*

Algoritmo no Supervisado para Desambiguación de Nombres de Personas en la Web

Agustín D. Delgado, Raquel Martínez, Víctor Fresno
Universidad Nacional de Educación a Distancia (UNED)
Juan del Rosal, 16, 28040 - Madrid
{agustin.delgado, raquel, vfresno}@lsi.uned.es

Soto Montalvo
Universidad Rey Juan Carlos (URJC)
Tulipán, S/N, 28933 - Móstoles
soto.montalvo@urjc.es

Resumen: En este trabajo presentamos un sistema no supervisado para agrupar los resultados proporcionados por un motor de búsqueda cuando la consulta corresponde a un nombre de persona compartido por diferentes individuos. Las páginas web se representan mediante n-gramas de diferente información y tamaño. Además, proponemos un algoritmo de clustering capaz de calcular el número de clusters y devolver grupos de páginas web correspondientes a cada uno de los individuos, sin necesidad de entrenamiento ni umbrales predefinidos, como hacen los mejores sistemas del estado del arte en esta tarea. Hemos evaluado nuestra propuesta con tres colecciones de evaluación propuestas en diferentes campañas de evaluación para la tarea de Desambiguación de Personas en la Web. Los resultados obtenidos son competitivos y comparables a aquellos obtenidos por los mejores sistemas del estado del arte que utilizan algún tipo de supervisión.
Palabras clave: aprendizaje no supervisado, clustering, n-gramas, búsqueda de personas en la web

Abstract: In this paper we present an unsupervised approach for clustering the results of a search engine when the query is a person name shared by different individuals. We represent the web pages using n-grams, comparing different kind of information and different length of n-grams. Moreover, we propose a new clustering algorithm that calculates the number of clusters and establishes the groups of web pages according to the different individuals, without the need of any training data or predefined thresholds, as the successful state of the art systems do. Our approach is compared with three gold standard collections compiled by different evaluation campaigns for the task of Web People Search. We obtain really competitive results, comparable to those obtained by the best approaches that use annotated data.
Keywords: unsupervised learning, clustering, n-grams, web people search

1 Introduction

Resolving the ambiguity of person names in web search results is a challenging problem becoming an area of interest for Natural Language Processing (NLP) and Information Retrieval (IR) communities. This task can be defined informally as follows: given a query of a person name in addition to the results of a search engine for that query, the goal is to cluster the resultant web pages according to the different individuals they refer to. Thus, the challenge of this task is estimating the number of different individuals and grouping the pages of the same individual in the same cluster.

The difficulty of this task resides in the fact that a single person name can be shared by many people. This problem has had an impact in the Internet and that is why several vertical search engines specialized in web people search have appeared in the last years, e.g. `spokeo.com`, `123people.com` or `zoominfo.com`. This

* The authors would like to thank the financial support for this research to the Spanish research project Holopedia funded by the Ministerio de Ciencia e Innovación under grant TIN2010-21128-C02 and by UNED Project (2012V/PUNED/0004).

task should not be mixed up with *entity linking* (EL). The goal of EL is to link name mentions of entities in a document collection to entities in a reference knowledge base (typically Wikipedia), or to detect new entities.

The main difficulties of clustering web pages referring to the same individual come from their possible heterogeneous nature. For example, some pages may be professional sites, while others may be blogs containing personal information. To overcome these difficulties the users have to refine the queries with additional terms. This task gets harder when the person name is shared by a celebrity or by a historical figure, because the results of the search engines are dominated by that individual, making the search of information about other individuals more difficult.

WePS[1] (Web People Search) evaluation campaigns proposed this task in a web searching scenario providing several corpora for evaluating the results of their participants, particularly WePS-1, WePS-2 and WePS-3 campaigns. This framework allows to compare our approach with the state of the art systems.

The most successful state of the art systems have addressed this problem with some kind of supervision. This work proposes a data-driven method for this task with the aim of eliminating the elements of human annotation involvement in the process as much as possible. The main contribution of this work is a new unsupervised approach for resolving person name ambiguity of web search results. It is based on the use of capitalized n-grams to represent the pages that share the same person name, and also in an algorithm that decides if two web pages have to be grouped using a threshold that only depends on the information of both pages.

The paper is organized as follows: in Section 2 we discuss related work; Section 3 details the way we represent the web pages and our algorithm; in Section 4 we describe the collections used for evaluating our approach and we show our results making a comparison with other systems; the paper ends with some conclusions and future work in Section 5.

2 Related Work

Several approaches have been proposed for clustering search results for a person name query. The main differences among all of them are the features they use to represent the web pages and the clustering algorithm. However, the most successful of them have in common that they use some kind of supervision: learning thresholds and/or fixing manually the value of some parameters according to training data.

Regarding the way of representing a web page, the most popular features used by the most successful state of the art approaches are Name Entities (NE) and Bag of Words (BoW) weighted by TF-IDF function. In addition to such features, the systems usually use other kind of information. Top systems from WePS-1 and WePS-2 campaigns, CU_COMSEM (Chen and Martin, 2007) and PolyUHK (Chen, Yat Mei Lee, and Huang, 2009), distinguish several kind of tokens according to different schemes (URL tokens, title tokens, ...) and build a feature vector for each sort of tokens, using also information based on the noun phrases appearing in the documents. PolyUHK also represents the web pages with n-grams and adds pattern techniques, attribute extraction and detection when a web page is written in a formal way. A more recent system, HAC_Topic (Liu, Lu, and Xu, 2011), also uses BoW of local and global terms weighted by TF-IDF. It adds a topic capturing method to create a Hit List of shared high weighted tokens for each cluster obtaining better results than WePS-1 participants. IRST-BP system (Popescu and Magnani, 2007), the third in WePS-1 participant ranking, proposes a method based in the hypothesis that appropriated n-grams characterize a person and makes extensive use of NE and other features as temporal expressions. PSNUS system (Elmacioglu et al., 2007) use a large number of different features including tokens, NE, hostnames and domains, and n-gram representation of the URL links of each web page. (Artiles, Amigó, and Gonzalo, 2009a) studies, using also the collections WePS-1 and WePS-2, the role of several features as NE, n-grams or noun phrases for this task reformulating this problem as a classification task. In their conclusions, they claim that using NE does not provide substantial improvement than using other combination of features that do not require linguistic processing (snippet tokens, n-grams, ...). They also present results applying only n-grams of different length, claiming that n-grams longer than 2 are not effective, but bigrams improves the results of tokens. On the other hand, the WePS-3 best system, YHBJ (Chong and Shi, 2010), uses information extracted manually from Wikipedia adding

[1] http://nlp.uned.es/weps/

to BoW and NE weighted by TF-IDF.

Regarding the clustering algorithms, looking at WePS campaigns results, the top ranked systems have in common the use of the Hierarchical Agglomerative Clustering algorithm (HAC) described in (Manning, Raghavan, and Schütze, 2008). Different versions of this algorithm were used by (Chen and Martin, 2007; Chen, Yat Mei Lee, and Huang, 2009; Elmacioglu et al., 2007; Liu, Lu, and Xu, 2011; Balog et al., 2009; Chong and Shi, 2010).

The only system that does not use training data, DAEDALUS (Lana-Serrano, Villena-Román, and González-Cristóbal, 2010), which uses k-Medoids, got poor results in WePS-3 campaign. In short, the successful state of the art systems need some kind of supervised learning using training data or fixing parameters manually. In this paper we explore and propose an approach to address this problem by means of data-driven techniques without the use of any kind of supervision.

3 Proposed Approach

We distinguish two main phases in this clustering task: web page representation (Sections 3.1 and 3.2) and web page grouping (Sections 3.3 and 3.4).

3.1 Feature Selection

The aim of this phase is to extract relevant information that could identify an individual. Several of the state of the art systems use word n-grams to represent the whole or part of the information of a web page. Our main assumption is that co-occurrences of word n-grams, particularly of capitalized words, could be an effective representation in this task. We assume the main following hypotheses:

(i) Capitalized n-grams co-occurrence could be a reliable way for deciding when two web pages refer the same individual. Capitalized n-grams usually are NE (organizations and company names, locations or other person names related with the individual) or information not detected by some NE recognizers as for example, the title of books, films, TV shows, and so on. In a previous study with WePS-1 training corpus using the Stanford NER[2] to annotate NE, we detected that only 55.78 % of the capitalized tokens were annotated as NE or components of a NE by the NER tool. So the use of capitalized tokens allows increase the number of features in connection to the use of only NE. We also compared the n-gram representation with capitalized tokens and with NE. We found that 30.97 % of the 3-grams composed by capitalized tokens were also NE 3-grams, and 25.64 % of the 4-grams composed by capitalized tokens were also NE 4-grams. So also in the case of n-grams the use of capitalized tokens increases the number of features compared to the use of only NE.

(ii) If two web pages share capitalized n-grams, the higher is the value of n, the more probable the two web pages refer to the same individual. We define "long enough n-grams" as those compose by at least 3 capitalized tokens.

Thus, a web page W is initially represented as the sequence of tokens starting in uppercase, in the order as they appear in the web page. Notice that some web pages could not be represented with this proposal because all their content was written in lowercase. In the case of the collections that we describe in Section 4.1, 0.66 % of the web pages are not represented for this reason.

3.2 Weighting Functions

We test the well known TF and TF-IDF functions, and z-score (Andrade and Medina, 1998). The z-score of an n-gram a in a web page W_i is defined as follows:

$$z\text{-score}(a, W_i) = \frac{TF(a,W_i)-\mu}{\sigma}$$

where $TF(a, W_i)$ is the frequency of the n-gram a in W_i; μ is the mean frequency of the n-gram a in the background set; and σ is the deviation of the n-gram a in the background set. In this context the background set is the set of web pages that share the person name. This score gives an idea of the distance of the frequency of an n-gram in a web page from the general distribution of this n-gram in the background set.

3.3 Similarity Functions

To determine the similarity between two web pages we try the cosine distance, a widely measure used in clustering, and the weighted Jaccard coefficient between two bags of n-grams defined as:

$$W.Jaccard(W_i^n, W_j^n) = \frac{\sum_k min(m(t_{k_i}^n, i), m(t_{k_j}^n, j))}{\sum_k max(m(t_{k_i}^n, i), m(t_{k_j}^n, j))}$$

where the meaning of $m(t_{k_i}^n, i)$ is explained in Section 3.4. Since weighted Jaccard coefficient

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

needs non-negative entries and we want the cosine similarity of two documents to range from 0 to 1, we translate the values of the z-score so that they are always non-negative.

3.4 Algorithm

The algorithm $UPND$ (Unsupervised Person Name Disambiguator) can be seen in Algorithm 1.

$UPND$ algorithm receives as input a set of web documents with a mention to the same person name, let be $\mathcal{W} = \{W_1, W_2, \ldots, W_N\}$, and starts assigning a cluster C_i for each document W_i. $UPND$ also receives as input a pair of positive integer values r_1 and r_2, such that $r_2 \geq r_1$, specifying the range of values of n in the n-grams extracted from each web document. In each step of the algorithm we assign to each web page W_i a bag of n-grams $W_i^n = \{(t_1^n, m(t_1^n, i)), (t_2^n, m(t_2^n, i)), \ldots, (t_{k_i}^n, m(t_{k_i}^n, i))\}$, where each t_r^n is a n-gram extracted from W_i and $m(t_r^n, i)$ is the corresponding weight of the n-gram t_r^n in the web page W_i, being $r \in \{1, 2, \ldots, k_i\}$. In Algorithm 1 the function $setNGrams(n, \mathcal{W})$ in line 6 calculates for each web page in the set $\mathcal{W}$ its bag of n-grams representation. $Sim(W_i^n, W_j^n)$ in line 9 refers to the similarity between web pages W_i and W_j.

To decide when two web pages refer the same individual we propose a threshold γ. This threshold takes into account two factors: the number of n-grams shared by the web pages and the size of both web pages. For each pair of web pages represented as bag of n-grams, let be W_i^n and W_j^n, we define the following thresholds:

$$\gamma_{max}(W_i^n, W_j^n) = \frac{min(k_i, k_j) - shared(W_i^n, W_j^n)}{max(k_i, k_j)}$$

$$\gamma_{min}(W_i^n, W_j^n) = \frac{min(k_i, k_j) - shared(W_i^n, W_j^n)}{min(k_i, k_j)}$$

where k_i and k_j are the number of n-grams of W_i and W_j respectively, and $shared(W_i^n, W_j^n)$ is the number of n-grams shared by those web pages, i.e. $shared(W_i^n, W_j^n) = |W_i^n \cap W_j^n|$. Notice that $shared(W_i^n, W_j^n)$ is superiorly limited by $min(k_i, k_j)$.

These thresholds hold two desirable properties: (i) The more n-grams are shared by W_i and W_j, the lower the threshold is, so the clustering condition of the algorithm is less strict. (ii) It avoids the penalization due to big differences between the size of the web pages.

γ_{min} benefits the grouping of those web pages that are subsets of other bigger web pages. However, this can lead to a mistake when a small web page is similar to part of other bigger one, but that belongs to different persons. Then, we try to balance this effect by including also γ_{max}. The final threshold is the arithmetic mean of the previous functions:

$$\gamma_{avg}(W_i^n, W_j^n) = \frac{\gamma_{max}(W_i^n, W_j^n) + \gamma_{min}(W_i^n, W_j^n)}{2}$$

what avoids giving advantage to web pages according to their size. We tested these three threshold and γ_{avg} shows a behavior more independent of the size of the n-grams, the similarity functions and the weighting functions.

Thus, two web pages W_i and W_j refer to the same person if $Sim(W_i^n, W_j^n) \geq \gamma_{avg}(W_i^n, W_j^n)$, so $C_i = C_i \cup C_j$ (lines 9, 10 and 11).

The algorithm has three input parameters: $\mathcal{W}$, the set of web pages with the same person name, and r_1 and r_2 that allows the algorithm to iterate this process for r_1-grams to r_2-grams.

This algorithm is polynomial and has a computational cost in $\mathcal{O}(N^2)$, where N is the number of web pages.

Algorithm 1 $UPND(\mathcal{W}, r_1, r_2)$

Require: Set of web pages that shared a person name $\mathcal{W} = \{W_1, W_2, ..., W_N\}$, $r_1, r_2 \geq 1$ such that $r_2 \geq r_1$
Ensure: Set of clusters $\mathcal{C} = \{C_1, C_2, ..., C_l\}$

```
1:  for n = 1 to N do
2:      C_i = {W_i}
3:  end for
4:  C = {C_1, C_2, ..., C_N}
5:  for n = r_1 to r_2 do
6:      setNGrams(n, W)
7:      for i = 1 to N do
8:          for j = i + 1 to N do
9:              if Sim(W_i^n, W_j^n) ≥ γ_avg(W_i^n, W_j^n) then
10:                 C_i = C_i ∪ C_j
11:                 C = C \ {C_j}
12:             end if
13:         end for
14:     end for
15: end for
16: return C
```

4 Experiments

In this section we present the corpora of web pages, the experiments carried out and the results.

4.1 Web People Search Collections

WePS is a competitive evaluation campaign that proposes several tasks including resolution of disambiguation on the Web data. In particular, WePS-1, WePS-2 and WePS-3 campaigns provide an evaluation framework consisting in several annotated data sets composed of English person names.

In these experiments we use WePS-1 (Artiles, Gonzalo, and Sekine, 2007) test corpus composed by 30 English person names and the top 100 search results from Yahoo! search engine; WePS-2 (Artiles, Gonzalo, and Sekine, 2009b) containing 30 person names and the top 150 search results from Yahoo! search engine; and WePS-3 (Artiles et al., 2010) with 300 person names and the top 200 search results from Yahoo!

4.2 Results and Discussion

We present our results for all the corpora comparing them with the state of the art systems. The figures in the tables are macro-averaged, i.e., they are calculated for each person name and then averaged over all test cases. The metrics used in this section are the BCubed metrics defined in (Bagga and Baldwin, 1998): BCubed precision (BP), BCubed recall (BR) and their harmonic mean $F_{0,5}(BP/BR)$. (Artiles, 2009) showed that these metrics are accurate for clustering tasks, particularly for person name disambiguation in the Web.

We use the Wilcoxon test (Wilcoxon, 1945) to detect statistical significance in the differences of the results considering a confidence level of 95 %. In order to compare our algorithm with the WePS better results using the Wilcoxon test, the samples consist in the pairs of values $F_{\alpha=0,5}(BP/BR)$ of each system for each person name.

In order to evaluate our representation approach we first run our algorithm representing the web pages with the n-grams considering all the tokens. Table 1 shows the results of $UPND$ algorithm representing the web pages with 4-grams ($UPND(\mathcal{W}, 4, 4)$) and 3-grams ($UPND(\mathcal{W}, 3, 3)$). Previous experiments using bigrams showed that they are less suitable for this approach. For the representation of $\mathcal{W}$ we discard those n-grams that only appear in one document. The figures shows that, in general, the results obtained with 4-grams outperform those with 3-grams. Weighted Jaccard similarity seems to be more independent of the weighting functions than Cosine. On the other hand, most of the times Cosine gets its best scores when it is applied with z-score. Notice that Jaccard obtains an improvement of the Recall results, whereas Cosine gets better Precision results. The significance test comparing the best scores for Jaccard and Cosine (TF with Jaccard, z-score with Cosine) shows that there are not significant differences. In this case the representation with all 4-grams obtains high Precision scores, whereas the representation with 3-grams increase Recall but with too low Precision scores.

Then we carried out the same experiments but representing the web pages with capitalized n-grams. Table 2 shows these results. In this case, the figures shows that, in general and contrary to the previous experiments, it is not obvious which size of n works the best. The significance test comparing the best scores for each size of n: 4-grams with z-score and Jaccard, and 3-grams with z-score and Cosine shows that there are not significant differences. Thus, given than the representation with 3-grams is less expensive that the one with 4-grams we selected the former. Focussing on 3-grams, the significance test comparing the best scores for Jaccard and Cosine (TF with Jaccard, z-score with Cosine) shows that only with the WePS-3 data set there is a significant difference in favor of z-score+Cosine.

Since we consider that in this task is more relevant Precision than Recall, as we want to have groups of mostly true positives (web pages of the same individual), we select the combination of z-score as weighting function and cosine as similarity function as the most suitable combination for our algorithm. Therefore we use it in the following experiments.

Finally, comparing the selected representation with all the n-grams (4-grams, z-score, cosine) with the selected one for capitalized n-grams (3-grams, z-score, cosine) the significance test shows that only there is a significance difference with WePS-1 data set in favor of the representation with all the n-grams. Thus, we consider that the representation only with capitalized n-grams is competitive, since it obtains comparable results to those obtained with all the n-grams, with the advantage of being more efficient both in space and time.

Table 3 shows the results of $UPND$ with WePS-1 test, WePS-2 and WePS-3 corpora in addition to the top ranking systems of the campaigns, and also the results obtained by

		WePS-1			WePS-2			WePS-3		
		BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)
4-grams										
W. Jaccard	TF	0.86	0.75	**0.79**	0.90	0.72	**0.79**	0.62	0.57	**0.54**
	z-score	0.85	0.75	**0.79**	0.9	0.73	**0.79**	0.61	0.58	**0.54**
	TF-IDF	0.86	0.75	**0.79**	0.90	0.72	**0.79**	0.62	0.57	**0.54**
Cosine	TF	0.90	0.70	0.78	0.95	0.63	0.74	0.70	0.47	0.52
	z-score	0.89	0.71	0.78	0.95	0.67	0.77	0.69	0.50	0.53
	TF-IDF	0.90	0.69	0.77	0.95	0.57	0.7	0.72	0.44	0.51
3-grams										
W. Jaccard	TF	0.58	0.87	0.68	0.68	0.89	0.76	0.36	0.81	0.45
	z-score	0.57	0.88	0.67	0.68	0.89	0.75	0.35	0.82	0.45
	TF-IDF	0.58	0.87	0.68	0.68	0.89	0.76	0.36	0.81	0.45
Cosine	TF	0.69	0.8	**0.73**	0.78	0.81	0.78	0.46	0.66	**0.49**
	z-score	0.66	0.83	0.72	0.78	0.84	**0.8**	0.44	0.71	**0.49**
	TF-IDF	0.7	0.79	**0.73**	0.78	0.76	0.75	0.48	0.63	**0.49**

Table 1: Results of $UPND$ algorithm for WePS test data sets using all the n-grams.

		WePS-1			WePS-2			WePS-3		
		BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)	BP	BR	$F_{0,5}$(BP/BR)
4-grams										
W. Jaccard	TF	0.89	0.67	**0.76**	0.95	0.69	**0.79**	0.68	0.51	0.53
	z-score	0.89	0.67	**0.76**	0.93	0.69	**0.79**	0.67	0.52	**0.54**
	TF-IDF	0.89	0.67	**0.76**	0.95	0.69	**0.79**	0.68	0.51	0.53
Cosine	TF	0.93	0.63	0.75	0.96	0.60	0.72	0.74	0.44	0.51
	z-score	0.92	0.65	**0.76**	0.96	0.63	0.75	0.73	0.46	0.52
	TF-IDF	0.93	0.63	0.74	0.96	0.59	0.71	0.74	0.44	0.51
3-grams										
W. Jaccard	TF	0.72	0.78	0.73	0.81	0.83	**0.81**	0.46	0.70	0.50
	z-score	0.70	0.79	0.73	0.8	0.84	**0.81**	0.45	0.72	0.50
	TF-IDF	0.72	0.78	0.73	0.81	0.83	**0.81**	0.46	0.70	0.50
Cosine	TF	0.78	0.73	0.74	0.85	0.76	0.79	0.56	0.59	**0.52**
	z-score	0.76	0.76	**0.75**	0.85	0.79	**0.81**	0.54	0.62	**0.52**
	TF-IDF	0.78	0.75	**0.75**	0.86	0.75	0.79	0.57	0.57	0.52

Table 2: Results of $UPND$ algorithm for WePS test data sets using capitalized n-grams.

HAC_Topic system in the case of WePS-1. We include the results obtained by three unsupervised baselines called ALL_IN_ONE, ONE_IN_ONE and Fast AP. ALL_IN_ONE provides a clustering solution where all the documents are assigned to a single cluster, ONE_IN_ONE returns a clustering solution where every document is assigned to a different cluster, and Fast AP applies a fast version of Affinity Propagation described in (Fujiwara, Irie, and Kitahara, 2011) using the function TF-IDF to weight the tokens of each web page, and the cosine distance to compute the similarity.

Our algorithm $UPND$ outperforms WePS-1 participants and all the unsupervised baselines described before. HAC_Topic also outperforms the WePS-1 top participant systems and our algorithm. This system uses several parameters obtained by training with the WePS-2 data set: token weight according to the kind of token (terms from URL, title, snippets, ...) and thresholds used in the clustering process. Note that WePS-1 participants used the training corpus provided to the campaign, the WePS-1 training data, so in this case the best performance of HAC_Topic could be not only due to the different approach, but also because of the different training data set.

$UPND$ obtains significative better results than the WePS-1 top participant results, and HAC_Topic obtains significative better results than it according to the Wilcoxon test. $UPND$ obtains significative better results than IRST-BP system (the third in the WePS-1 ranking), also based on the co-ocurrence of n-grams.

Regarding WePS-2 we add in Table 3 two oracle systems provided by the organizers. The oracle systems use BoW of tokens (ORACLE_1) or bigrams (ORACLE_2) weighted by TF-IDF, deleting previously stop words, and later applying HAC with single linkage with the best thresholds for each person name. We do not include the results of the HAC_Topic system since it uses this

	System	BP	BR	$F_{0.5}$(BP/BR)
WePS-1	(+) HAC_Topic	0.79	0.85	**0.81** †
	(-) ***UPND (all-4g)***	0.89	0.71	0.78 •
	(-) ***UPND (cap-3g)***	0.76	0.76	0.75 •
	(+)(*) CU_COMSEM	0.61	0.83	0.70 †
	(+)(*) PSNUS	0.68	0.73	0.70 †
	(+)(*) IRST-BP	0.68	0.71	0.69 †
	(+)(*) UVA	0.79	0.50	0.61 †
	(+)(*) SHEF	0.54	0.74	0.62 †
	(-) ONE_IN_ONE	**1.00**	0.43	0.57 •
	(-) Fast AP	0.69	0.55	0.56 †
	(-) ALL_IN_ONE	0.18	**0.98**	0.25 •
WePS-2	(+) ORACLE_1	0.89	0.83	**0.85** •
	(+) ORACLE_2	0.91	0.81	**0.85** •
	(+)(*) PolyUHK	0.87	0.79	0.82
	(+)(*) ITC-UT_1	0.93	0.73	0.81
	(-) ***UPND (cap-3g)***	0.85	0.79	0.81 •
	(+)(*) UVA_1	0.85	0.80	0.81
	(-) ***UPND (all-4g)***	0.95	0.67	0.77 •
	(+)(*) XMEDIA_3	0.82	0.66	0.72 †
	(+)(*) UCI_2	0.66	0.84	0.71 †
	(-) ALL_IN_ONE	0.43	**1.00**	0.53 •
	(-) Fast AP	0.80	0.33	0.41 †
	(-) ONE_IN_ONE	**1.00**	0.24	0.34 •
WePS-3	(+)(*) YHBJ_2	0.61	0.60	**0.55**
	(-) ***UPND (cap-3g)***	0.54	0.62	0.52 •
	(+)(*) AXIS_2	0.69	0.46	0.50 †
	(-) ***UPND (all-4g)***	0.44	0.71	0.49 •
	(+)(*) TALP_5	0.40	0.66	0.44 †
	(+)(*) RGAI_AE_1	0.38	0.61	0.40 †
	(+)(*) WOLVES_1	0.31	0.80	0.40 †
	(-)(*) DAEDALUS_3	0.29	0.84	0.39 †
	(-) Fast AP	0.73	0.30	0.38 †
	(-) ONE_IN_ONE	**1.00**	0.23	0.35 •
	(-) ALL_IN_ONE	0.22	**1.00**	0.32 •

Table 3: Result of $UPND$ and the top state of the art systems with WePS corpora: (+) means system with supervision; (-) without supervision and (*) campaign participant. Significant differences between $UPND$ and other systems are denoted by (†); (•) means that in this case the statistical significance is not evaluated.

data set for training their algorithm.

The significance test shows that the top WePS-2 systems PolyUHK, UVA_1 and ITC-UT_1 obtain similar results than $UPND(cap-3g)$, however they use some kind of supervision. The results of all these systems are the closest to the oracle systems, which know the best thresholds for each person name.

In the case of WePS-3, the organizers did not consider for evaluation the whole clustering solution provided by the systems like in previous editions, but only checks the accuracy of the clusters corresponding to two selected individuals per person name. In this case, the first two systems YHBJ_2 and $UPND(cap-3g)$ do not have significant differences in their results. Notice that YHBJ_2 system makes use of concepts extracted manually from Wikipedia. $UPND$ also obtains significative better results than DAEDALUS_3, the only one participant that does not use training data.

(Artiles, Amigó, and Gonzalo, 2009a) applied HAC algoritm over n-grams of length 2 to 5 getting similar results of precision than $UPND$ but very low recall scores. This means that applying HAC only over n-grams is not a good choice and $UPND$ takes more advantage of these features.

After all these experiments, we can conclude that our approach gets the best results of all the completely unsupervised approaches. Moreover, the precision scores for all collections are very high and confirm that our approach is accurate to get relevant information for characterizing an individual. We also obtain competitive recall results, what lead to a competitive system that carries out person name disambiguation in web search results without any kind of supervision.

5 *Conclusions and Future Work*

We present a new approach for person name disambiguation of web search results. Our method does not need training data to calculate thresholds to determine the number of different individuals sharing the same name, or whether two web pages refer to the same individual or not. Although supervised approaches have been successful in many NLP and IR tasks, they require enough and representative training data to guaranty consistent results for different data collections, which requires a huge human effort.

The proposed algorithm provides a clustering solution for this task by means of data-driven methods that do not need learning from training data. Our approach obtains very competitive results in all the data sets compared with the best state of the art systems. It is based on getting reliable information for disambiguating, particularly long n-grams composed by uppercase tokens. According to our results, this hypothesis has shown successful, getting high precision values and acceptable recall scores. Anyway, we would like to improve recall results without losing of precision, filter out noisy capitalized n-grams, and build an alternative representation for web pages containing all their tokens in lowercase.

Person name disambiguation has been mainly addressed in a monolingual scenario, e.g. WePS corpora are English data sets. We would like to address this task in a multilingual scenario. Although search engines return their results taking into account the country of the user, with some queries we can get results written in several languages. This scenario has not been considered by the state of the art systems so far.

References

Andrade, M.A. and A. Valencia. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14:600-607.

Artiles, J. 2009. Web People Search. PhD Thesis, UNED University.

Artiles, J., J. Gonzalo, and S. Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64-69. ACL.

Artiles, J., E. Amigó, and J. Gonzalo. 2009a. The Role of Named Entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Artiles, J., J. Gonzalo, and S. Sekine. 2009b. Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Artiles, J., A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Bagga, A. and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pages 79-85. ACL.

Balog, K., J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. de Rijke. 2009. The University of Amsterdam at WePS-2. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Chen, Y. and J. Martin. 2007. CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Named Disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 125-128. ACL.

Chen, Y., S. Yat Mei Lee, and C. Huang. 2009. PolyUHK: A Robust Information Extraction System for Web Personal Names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Elmacioglu, E., Y. Fan Tan, S. Yan, M. Kan, and D. Lee. 2007. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 268-271. ACL.

Fujiwara, Y., G. Irie, and T. Kitahara. 2011. Fast Algorithm for Affinity Propagation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence(IJCAI)- Volume Three*, pages 2238-2243.

Lana-Serrano, S., J. Villena-Román, and J.C. González-Cristóbal. 2010. Daedalus at WebPS-3 2010: k-Medoids Clustering using a Cost Function Minimization. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Liu, Z., Q. Lu, and J. Xu. 2011. High Performance Clustering for Web Person Name Disambiguation using Topic Capturing. In *International Workshop on Entity-Oriented Search (EOS)*.

Long, C. and L. Shi. 2010. Web Person Name Disambiguation by Relevance Weighting of Extended Feature Sets. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

Mann, G.S. 2006. Multi-Document Statistical Fact Extraction and Fusion. PhD thesis, Johns Hopkins University, Baltimore, MD, USA. AAI3213760.

Manning, C.D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, USA.

Popescu, O. and B. Magnini. 2007. IRST-BP: Web People Search Using Name Entities In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 195-198. ACL.

Wilcoxon, F. 1945. *Individual Comparisons by Ranking Methods*, 1(6). Biometrics Bulletin.

Aprendizaje Automático en Procesamiento del Lenguaje Natural

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 61-68 recibido 17-04-14 revisado 06-06-14 aceptado 10-06-14

Translating sentences from 'original' to 'simplified' Spanish

Traducción de frases del español 'original' al español 'simplificado'

Sanja Štajner
Research Group in Computational Linguistics
Research Institute in Information and Language Processing
University of Wolverhampton, UK
sanjastajner@wlv.ac.uk

Resumen: La Simplificación de Textos (ST) tiene como objetivo la conversión de oraciones complejas en variantes más sencillas, que serían más accesibles para un público más amplio. Algunos estudios recientes han abordado este problema como un problema de traducción automática (TA) monolingüe (traducir de lengua 'original' a 'simplificada' en lugar de traducir de un idioma a otro), utilizando el modelo estándar de traducción automática basado en frases. En este estudio, investigamos si el mismo enfoque tendría el mismo éxito independientemente del tipo de simplificación que se quiera estudiar, dado que cada público meta require diferentes niveles de simplificación. Nuestros resultados preliminares indican que el modelo estándar podría no ser capaz de aprender las fuertes simplificaciones que se necesitan para algunos usuarios, e.g. gente con el síndrome de Down. Además, mostramos que las tablas de traducción obtenidas durante el proceso de traducción parecen ser capaces de capturar algunas simplificaciones léxicas adecuadas.
Palabras clave: simplificación de textos, traducción automática estadística

Abstract: Text Simplification (TS) aims to convert complex sentences into their simpler variants, which are more accessible to wider audiences. Several recent studies addressed this problem as a monolingual machine translation (MT) problem (translating from 'original' to 'simplified' language instead of translating from one language into another) using the standard phrase-based statistical machine translation (PB-SMT) model. We investigate whether the same approach would be equally successful regardless of the type of simplification we wish to learn (given that different target audiences require different levels of simplification). Our preliminary results indicate that the standard PB-SMT model might not be able to learn the strong simplifications which are needed for certain users, e.g. people with Down's syndrome. Additionally, we show that the phrase-tables obtained during the translation process seem to be able to capture some adequate lexical simplifications.
Keywords: text simplification, phrase-based statistical machine translation

1 Introduction

Since the late nineties, several initiatives raised awareness of the complexity of the vast majority of written documents and the difficulties they pose to people with any kind of reading or learning impairments. These initiatives proposed various guidelines for writing in a simple and easy-to-read language which would be equally accessible to everyone. However, manual adaptation of existing documents could not keep up with the everyday production of material written in a 'complex' language. This motivated the need for automatic Text Simplification (TS), which aims to convert complex sentences into their simpler variants, while preserving the original meaning.

The first TS systems were traditionally rule-based, e.g. (Devlin, 1999; Canning et al., 2000), requiring a great number of hand-crafted simplification rules produced by highly specialised people. They consisted of

ISSN 1135-5948

two main simplification modules – lexical and syntactic. The lexical simplification module replaces long and uncommon words with their shorter and more commonly used synonyms. The syntactic simplification module recursively applies a set of handcrafted rules to each sentence as long as there are any rules which can be applied. The main drawbacks of those systems are that such rules cannot be easily adapted to different languages or genres, and that they lead to TS systems with high precision and low recall. With the emergence of Simple English Wikipedia (SEW)[1], which together with the 'original' English Wikipedia (EW)[2] provided a large parallel corpus for TS, some new machine learning oriented approaches have appeared. Several recent studies addressed text simplification as a monolingual machine translation (MT) problem. Instead of translating from one language to another, they tried to translate from the 'original' to the 'simplified' language.

In this paper, we explore the influence of the level of simplification in the training dataset on the performance of a phrase-based statistical machine translation (PB-SMT) model which tries to translate from 'original' to 'simplified' Spanish. Our preliminary results indicate that PB-SMT systems might not be appropriate when the training set contains a great number of 'strong' simplifications (which are needed for some target populations such as people with Down's syndrome for example), while they might work reasonably well when trained on the datasets which contain only the 'weak' simplifications (which are sufficient for some other target populations such as non-native speakers or people with low literacy levels). Additionally, we show that the phrase-based tables produced during the translation process contain a great number of adequate lexical paraphrases which could be used to build a separate lexical simplification module if necessary.

The remainder of the paper is structured as follows: Section 2 presents the related work on text simplification with a special emphasis on previous uses of PB-SMT systems in TS; Section 3 describes the corpora which were used and the experiments conducted; Section 4 presents the performances of the two PB-SMT systems trained on different corpora, and discusses the possibilities for using the phrase-based tables produced during the translation process; Section 5 lists the main findings and gives directions for future work.

2 Related Work

Due to the lack of large parallel corpora of original and simplified texts, many of the recent TS systems are still rule-based, e.g. (Saggion et al., 2011; Drndarević et al., 2013; Orasan, Evans, and Dornescu, 2013). However, the number of machine learning (ML) approaches to TS has increased in the last few years. This increase is especially pronounced in English TS, due to the large and freely available parallel corpus of original and simplified texts – English Wikipedia (EW) and Simple English Wikipedia (SEW). Napoles and Drezde (2010) built a statistical classification system that can distinguish which version of English Wikipedia a text belongs to, thus confirming the possibility of using those corpora in TS. Yatskar et al. (2010) used edit histories in SEW to extract lexical simplifications, and Biran et al. (2011) applied an unsupervised method for learning pairs of complex and simple synonyms from the EW and SEW. Zhu et al. (2010) proposed a tree-based simplification model, while Woodsend and Lapata (2011) used quasi-synchronous grammar to learn a wide range of rewriting transformations for TS.

Several recent studies addressed the TS as a monolingual MT problem. Instead of translating from one language to another, they tried to translate from the 'original' to the 'simplified' language. Coster and Kauchak (2011) applied the standard PB-SMT model implemented in Moses toolkit to 137,000 sentence pairs from the EW and SEW. They also suggested an extension of that model, which adds phrasal deletion to the probabilistic translation model in order to better cover deletion, which is a frequent phenomenon in TS. The obtained results (BLEU = 59.87 on the standard model without phrasal deletion, and BLEU = 60.46 on the extended model) were promising, although not far from the baseline (no translation performed), thus suggesting that the system is overcautious in performing simplifications. In order to overcome this issue, Wubben et al. (2012) performed post-hoc re-ranking on the output

[1] http://simple.wikipedia.org/wiki/Main_Page
[2] http://en.wikipedia.org/wiki/Main_Page

Version	Example
Original	Ahora se amplía, aunque siempre según el parecer del juez, a conducir con un exceso de velocidad superior en 60 kilómetros por hora en vía urbana o en 80 kilómetros por hora en vía interurbana, o conducir bajo la influencia de las drogas o con una tasa de alcohol superior a 1,2 gramos por litro en sangre.
Weak	*Esta medida* se amplía, *dependiendo del juez*, a conducir con un exceso de velocidad mayor de 60 kilómetros por hora en vía urbana o de 80 kilómetros por hora en vía interurbana, o conducir drogado o con una tasa de alcohol mayor a 1,2 gramos por litro en sangre.
Strong	Ahora *los jueces también podrán quitar el coche a las personas condenadas por otras causas. Algunas causas son conducir muy rápido dentro de las ciudades o beber alcohol o tomar drogas antes de conducir.*
Original	El fallo definitivo con la ciudad ganadora del concurso se conocerá el próximo 3 de diciembre de 2010, fecha en la que se celebra el Día Internacional y Europeo de las Personas con Discapacidad.
Weak	*La decisión definitiva* con la ciudad ganadora del concursó se *sabrá* el próximo 3 de diciembre de 2010. *El 3 de Diciembre es* el Día Internacional y Europeo de las Personas con Discapacidad.
Strong	*El premio se entregará el* 3 de diciembre de 2010. *El 3 de diciembre es* el Día Internacional y Europeo de las Personas con Discapacidad.

Table 1: Weak vs strong simplification (deviations from the original sentence are shown in italics)

(simplification hypotheses) based on their dissimilarity to the input (original sentences), i.e. they selected the output that is as different as possible from the original sentence.

Specia (2010) used the standard PB-SMT model implemented in Moses toolkit to try to learn how to simplify sentences in Brazilian Portuguese. She used 4,483 original sentences and their corresponding 'natural' simplifications obtained under the PorSimples project (Gasperin et al., 2009). The project was aimed at people with low literacy levels and the newswire texts were simplified manually by a trained human editor, offering two levels of simplification: 'natural' and 'strong'. Specia (2010) used only the sentence pairs obtained by 'natural' simplification (where the most common simplification operation was lexical substitution), which would correspond to our 'weak' simplification in Spanish. The performance of the translation model was reasonably good – BLEU score of 60.75 – especially taking into account the relatively small size of the corpora (4,483 sentence pairs).

3 Methodology

The main goal of this study was to investigate how far the level of simplification present in the training dataset influences the performance of a PB-SMT system which tries to learn how to translate from 'original' to 'simplified' language. Therefore, we trained the standard PB-SMT system on two TS corpora in Spanish, which contained different levels of simplification (were targeted to different users and were thus compiled using different simplification guidelines). The corpora and the experimental settings are described in the next two sub-sections.

3.1 Corpora

The first corpus (*Strong* simplification) was compiled under the Simplext project[3], following detailed easy-to-read guidelines prepared especially for simplifying texts for readers with Down's syndrome. The 200 newswire texts were simplified manually by the trained human editors. Many sentences required a very high level of simplification (given the specific needs of the target population), as can be observed in Table 1.

The second corpus (*Weak* simplification) was created by three native speakers of Spanish, following the given guidelines with no concrete target population in mind. The guidelines consisted of the same main simplification rules (e.g. use simple sentences, use common words, remove redundant words, use a simpler paraphrase if applicable) as those present in the Simplext guidelines. This time, the editors were explicitly instructed not to use strong paraphrases, i.e. to limit the use of the 'use simpler paraphrase, if applicable' rule to a minimum and not to apply it to the whole sentence but rather only to a specific (short) part of the sentence.

[3] www.simplext.es

The differences in the simplifications obtained by the aforementioned two simplification strategies (*strong* and *weak*) are presented in Table 1. The corpora characteristics: the average number of words per sentence in both original and simplified corpora, and the average sentence-wise BLEU score (S-BLEU) of the sentence pairs (original sentence and its corresponding manually simplified version) for each corpus are presented in Table 2.

Corpus	ASL-O	ASL-S	S-BLEU
Strong	31.82	14.30	0.17
Weak	25.98	16.91	0.60

Table 2: Corpora characteristics: the average number of words per sentence in the original (*ASL-O*) and the simplified corpora (*ASL-S*), and the average sentence-wise BLEU score (*S-BLEU*)

BLEU (Papineni et al., 2002) evaluates MT output by using exact n-gram matching between the hypothesis and the reference translation. Additionally, it applies the brevity penalty which penalises the hypothesis (automatically simplified sentences, in our case) which are shorter than the reference translations (original sentences, in our case). As BLEU is designed to evaluate output on a document level, it is not ideal for sentence-level scoring. Instead, we use S-BLEU (sentence-level BLEU) to evaluate the sentence pairs. Unlike BLEU, S-BLEU will still positively score segments that do not have higher n-gram matching. The low average S-BLEU score on the training dataset (Table 2) suggests that there are many string transformations and strong paraphrases to be learnt, and thus the standard phrase-based translation model might not be the most suitable for the task.

3.2 Experiments

For the translation experiments, we used the standard PB-SMT system implemented in the Moses toolkit (Koehn et al., 2007), the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), and the refinement and phrase-extraction heuristics described further in (Koehn, Och, and Marcu, 2003). The systems were tuned using minimum error rate training (MERT) (Och, 2003). The Spanish Europarl corpus[4] (portion of 500,000 sentences) was used to build the 3-gram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002). The stack size was limited to 500 hypotheses during decoding.

Both experiments were conducted on exactly the same amount of data: 700 sentence pairs for training and 100 sentence pairs for development. The obtained translation models were evaluated on three test sets: (1) 50 sentence pairs randomly selected from the corpora with strong simplifications (*Test-s*), (2) 50 sentence pairs randomly selected from the corpora with weak simplifications (*Test-w*), and (3) the mixed dataset which contained 100 sentence pairs from the previous two test sets (*Test-m*). In all cases, the sentence pairs used for testing were different from those used for training and development.

4 Results and Discussion

The results of the two translation experiments are presented in Table 3.

Corpus	Test-s	Test-w	Test-m
Strong	0.0937	0.3944	0.2609
Weak	0.0930	0.4638	0.2996
Original	0.0939	0.5282	0.3187

Table 3: Results of the translation experiments (BLEU scores) on the 50 sentences with the strong simplification (*Test-s*), on the 50 sentences with weak simplification (*Test-w*), and on the 100 sentences mixed test set (*Test-m*); the *Original* corpus corresponds to the baseline when no simplification/translation is performed

The BLEU score results on the *Test-s* presented in Table 3 should be taken with caution, given the specificities of the strong simplifications present in that corpus (*Strong*). The 'gold standard' manually simplified versions of the sentences in this corpus are already very different from their corresponding originals (See the BLEU score for the *Original* corpus in Table 3, which corresponds to the baseline when no simplification/translation is performed). Therefore, the same (and very low) BLEU scores obtained by both experiments (*Strong* and *Weak*) on the test set from the corpus with

[4] http://www.statmg.org/europarl/

Ex.	Type	Sentence
1-s	Original	"Amnistía denuncia que el Gobierno egipcio ha detenido a 150 **opositores** antes de las elecciones."
	Automatic-weak	"Amnistía denuncia que el Gobierno egipcio ha detenido a 150 **contrarios** antes de las elecciones."
	Automatic-strong	"Amnistía *Internacional* que el Gobierno egipcio ha detenido a 150 opositores antes de las elecciones."
	Manual-test	"El Gobierno de Egipto detiene a personas de el partido de la oposición antes de las elecciones. La organización Amnistía Internacional ha denunciado que el Gobierno egipcio ha detenido a 150 personas de la oposición."
2-w	Original	"ÁVILA Y BARCELONA, ENTRE LAS CUATRO FINALISTAS DEL PREMIO CIUDAD EUROPEA ACCESIBLE."
	Automatic-weak	"ÁVILA Y BARCELONA **ESTÁN** ENTRE LAS CUATRO FINALISTAS DEL PREMIO CIUDAD EUROPEA ACCESIBLE."
	Automatic-strong	"ÁVILA y BARCELONA, ENTRE las CUATRO FINALISTAS *de el* CIUDAD EUROPEA ACCESIBLE."
	Manual-test	"ÁVILA Y BARCELONA, ENTRE LAS CUATRO ELEGIDAS PARA GANAR EL PREMIO CIUDAD EUROPEA ACCESIBLE."
3-w	Original	"Ahora se amplía, **aunque siempre según el parecer del juez**, a conducir con un exceso de velocidad **superior** en 60 kilómetros por hora en vía urbana o en 80 kilómetros por hora en vía interurbana, o conducir bajo la influencia de las drogas o con una tasa de alcohol superior a 1,2 gramos por litro en sangre."
	Automatic-weak	"Ahora se amplía, **dependiendo del juez** a conducir con un exceso de velocidad **mayor** de 60 kilómetros por hora en vía urbana o en 80 kilómetros por hora en vía interurbana o conducir bajo la influencia de las drogas o con una tasa de alcohol superior a 1,2 gramos por litro en sangre."
	Automatic-strong	"Ahora se amplía, aunque siempre *en* el parecer del juez, a conducir con un exceso de velocidad superior en 60 por *su helicóptero* en vía urbana, en un 80 por *helicóptero en, por vía* interurbana, conducir bajo la influencia de las drogas*,* con *un tipo* de alcohol superior a 1,2 gramos por litro en sangre."
	Manual-test	"Con la reforma del Código Penal la pérdida del vehículo se amplía a conducir con un exceso de velocidad superior en 60 kilómetros por hora en vía urbana o en 80 kilómetros por hora en vía interurbana, o conducir bajo la influencia de las drogas o con una tasa de alcohol superior a 1,2 gramos por litro en sangre."
4-w	Original	"Ana Juan **fue galardonada** con el Premio Nacional de Ilustración correspondiente a 2010, por el conjunto de su obra."
	Automatic-weak	"Ana Juan **recibió** el Premio Nacional de Ilustración correspondiente a 2010, por el conjunto de su obra."
	Automatic-strong	"Ana Juan fue galardonada con el Premio Nacional de Ilustración correspondiente a 2010, por el *que el leído*."
	Manual-test	"Ana Juan ganó el Premio Nacional de Ilustración de 2010 por el conjunto de la obra de Ana Juan."

Table 4: Automatic simplification obtained by training the PB-SMT system on two different datasets – the one containing strong simplifications (*Automatic-strong)*, and the other containing weak simplifications (*Automatic-weak*). Differences to the original sentence are shown in italics and bold, where the good replacements are shown in bold and the bad ones in italics. *Manual-test* contains the 'gold standard' manual simplification from the test set.

Ex.	Source phrase	Target phrase	p
1	educar bien resulta cansado	enseñar bien es mucho trabajo	0.50
2	educar bien resulta cansado	educar bien es cansado	0.50
3	sublevación	rebelión	0.50
4	sublevación	sublevación	0.50
5	subrayaron que la edad media de inicio	dijeron que la edad media de inicio	0.67
6	subrayaron que la edad media de inicio	indicaron que la edad media de inicio	0.33
7	sufrieron síndrome de inmersión	Las personas tenían síndrome de inmersión	1.00
8	través del cine	mediante el cine	0.75
9	través del cine	través del cine	0.25
10	través del cine y exponer su heterogeneidad	mediante el cine y mostrar su diversidad	0.50
11	través del cine y exponer su heterogeneidad	través del cine y exponer su heterogeneidad	0.50

Table 5: Examples of *source* and *target* phrases and their 'target given source' probabilities (p) in the phrase-tables produced from the training dataset with weak simplifications

strong simplifications (*Test-s*) does not necessarily mean that both systems are equally unsuccessful. Those results only indicate that the obtained automatic simplifications are very different from the 'gold standard' (which was expected given that no automatic simplification could propose such strong paraphrases as those present in that corpus), but not necessarily bad. However, the manual inspection of the automatically simplified sentences revealed that the output of the system trained on the corpus with *strong* simplifications is barely readable and is not able to learn any adequate simplifications. On the contrary, it only worsens the original sentences by making them ungrammatical and/or changing their meaning (see examples in Table 4). On the other hand, the output of the system trained on the corpus with *weak* simplifications was grammatical and in most of the cases it contained at least one adequate lexical simplification (see examples in Table 4). However, it seems that the system was overcautious in applying any transformations, and thus the output of the system did not differ much from the original sentences. Nevertheless, the automatically simplified sentences obtained by this system were as grammatical and usually less complex than the originals.

4.1 Additional experiment

Given the notable similarity of our 'weak' simplifications with the 'natural' simplifications used in (Specia, 2010), we performed an additional experiment. We randomly selected only a portion of the corpus used in (Specia, 2010) – 741 sentence pairs for training, 94 for development and 90 for testing – and performed the same translation experiment as for our two Spanish corpora (using the same setup in the Moses toolkit, but this time using the Lácio-Web corpus[5] for the LM). The average S-BLEU score in this portion of Brazilian Portuguese corpora was 0.58, thus very similar to the one obtained on our Spanish corpus with weak simplifications (Table 2). The obtained BLEU score on the test set for Brazilian Portuguese was 0.5143, while the baseline (no simplification) was 0.5747. These results are again comparable to those obtained on our Spanish corpus with weak simplifications (Table 3).

4.2 Phrase-tables

We additionally examined the phrase-tables produced from the training dataset with *weak* simplifications. We observed many examples of identical source and target phrases with high probabilities. However, the phrase-tables contained a great number of adequate lexical simplifications and simple rewritings (Table 5). While the phrase-tables also provided many examples of bad lexical substitutions, most of them had a very low probabilities. These substitutions were thus discarded in the later stages by either the translation model or the language model.

In many cases, the probability score of the phrases which remain unchanged in the source and target was equal to or higher than the probability of the target phrase which is an adequate simplification of the source phrase (see examples 3 and 4, and 10 and 11 in Table 5). This might be one of the main reasons for the system being overcautious in applying any transformations. If this is the

[5] http://www.nilc.icmc.usp.bf/lacioweb/

case, the translation model could be modified in the way that it forces the system to pick the target phrase which is different from the source phrase whenever the probability of such a translation is higher than some carefully selected threshold.

Alternatively, the phrase-tables obtained during the translation process could be used to build an independent lexical simplification module. Such a module would go beyond the one word substitution level, offering lexical simplification for any phrase which consists of up to seven words (the default configuration in the Moses toolkit builds phrases with up to seven tokens). However, given the small size of the training sets, this approach would suffer from the sparseness problem. It would, therefore, need to be combined with a traditional lexical simplification module which would be used in cases when the 'complex' phrase cannot be found in the phrase-table.

5 Conclusions and Future Work

Text simplification has recently been treated as a statistical machine translation problem and addressed by using the standard phrase-based SMT models. Motivated by the fact that different target populations need different types of simplification, we investigated how much the level of simplification present in the training datasets influences the success of such a TS system.

It appears that a PB-SMT model works reasonably well only when the training dataset does not contain a great number of strong simplifications. Our results indicate that such translation models should not be used when we wish to learn strong simplifications which are needed for some specific audiences, e.g. people with Down's syndrome. Given the very small size of the training datasets used in this study, the reported results should only be regarded as preliminary. To the best of our knowledge, there are no other parallel corpora consisting of original and manually simplified texts in Spanish which could be used to enlarge our training datasets. Therefore, we cannot completely rule out the possibility that the PB-SMT systems would not reach some reasonably good performance if trained on much larger datasets.

The phrase-tables produced during the translation process open two possible avenues for future research. First, it would be interesting to explore how much we could improve the performance of the PB-SMT system if we force it to use the target phrases which are different from the source ones whenever the probability of such a translation is higher than some carefully selected threshold. Second, we could build an independent lexical simplification module based on the information contained in the phrase-tables. Such a lexical simplification module would go beyond performing the substitutions on the word level, offering lexical simplifications for phrases which consists of up to seven words.

Acknowledgements

I would like to express my gratitude to Profs. Ruslan Mitkov and Horacio Saggion, my director of studies and co-supervisor, for their input and help with the resources, as well as to the reviewers for their valuable comments and suggestions.

References

Biran, O., S. Brody, and N. Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Canning, Y., J. Tait, J. Archibald, and R. Crawley. 2000. Cohesive generation of syntactically simplified newspaper text. In *Proceedings of the Third International Workshop on Text, Speech and Dialogue*, TDS '00, pages 145–150, London, UK, UK. Springer-Verlag.

Coster, W. and D. Kauchak. 2011. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1–9.

Devlin, S. 1999. *Simplifying natural language text for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.

Drndarević, B., S. Štajner, S. Bott, S. Bautista, and H. Saggion. 2013. Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of the*

12th International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science. Samos, Greece, 24-30 March, 2013., pages 488–500.

Gasperin, C., L. Specia, T. Pereira, and S.M. Aluísio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA-2009), Bento Gonçalves, Brazil.*, pages 809–818.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Napoles, C. and M. Dredze. 2010. Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 160–167.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Orasan, C., R. Evans, and I. Dornescu, 2013. *Towards Multilingual Europe 2020: A Romanian Perspective*, chapter Text Simplification for People with Autistic Spectrum Disorders, pages 287–312. Romanian Academy Publishing House, Bucharest.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Saggion, H., E. Gómez Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Specia, L. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg.

Stolcke, A. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Woodsend, K. and M. Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wubben, S., A. van den Bosch, and E. Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1015–1024, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhu, Z., D. Berndard, and I. Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 69-76 recibido 15-04-14 revisado 01-07-14 aceptado 02-07-14

Descripción y Evaluación de un Sistema de Extracción de Definiciones para el Catalán

Description and Evaluation of a Definition Extraction System for Catalan

Luis Espinosa-Anke, Horacio Saggion
Universitat Pompeu Fabra
C/ Tànger, 122-134, 4a Planta
{luis.espinosa, horacio.saggion}@upf.edu

Resumen: La extracción automática de definiciones (ED) es una tarea que consiste en identificar definiciones en texto. Este artículo presenta un método para la identificación de definiciones para el catalán en el dominio enciclopédico, tomando como corpora para entrenamiento y evaluación una colección de documentos de la Wikipedia en catalán (Viquipèdia). El corpus de evaluación ha sido validado manualmente. El sistema consiste en un algoritmo de clasificación supervisado basado en Conditional Random Fields. Además de los habituales rasgos lingüísticos, se introducen rasgos que explotan la frecuencia de palabras en dominios generales y específicos, en definiciones y oraciones no definitorias, y en posición de definiendum (el término que se define) y de definiens (el clúster de palabras que define el definiendum). Los resultados obtenidos son prometedores, y sugieren que la combinación de rasgos lingüísticos y estadísticos juegan un papel importante en el desarrollo de sistemas ED para lenguas minoritarias.
Palabras clave: Extracción de Definiciones, Extracción de Información, Lexicografía Computacional

Abstract: Automatic Definition Extraction (DE) consists of identifying definitions in naturally-occurring text. This paper presents a method for the identification of definitions in Catalan in the encyclopedic domain. The train and test corpora come from the Catalan Wikipedia (Viquipèdia). The test set has been manually validated. We approach the task as a supervised classification problem, using the Conditional Random Fields algorithm. In addition to the common linguistic features, we introduce features that exploit the frequency of a word in general and specific domains, in definitional and non-definitional sentences, and in definiendum (term to be defined) and definiens (cluster of words that defines the definiendum) position. We obtain promising results that suggest that combining linguistic and statistical features can prove useful for developing DE systems for under-resourced languages.
Keywords: Extracción de Definiciones, Extracción de Información, Conditional Random Fields, Wikipedia

1 Introducción

Las enciclopedias y bases de datos terminológicas son bases de conocimiento de gran importancia para establecer relaciones semánticas entre distintos conceptos. No obstante, el desarrollo manual de estos recursos es habitualmente costoso y lento (Bontas y Mochol, 2005). La Extracción automática de Definiciones (ED), entendida como la tarea de identificar automáticamente definiciones en una producción lingüística natural, puede jugar un papel importante en este contexto. Existe un creciente interés en la ED en los campos del Procesamiento del Lenguaje Natural, la Lingüística Computacional y la Lexicografía Computacional. Además, trabajo previo ha demostrado su potencial para el desarrollo automático de glosarios (Muresan y Klavans, 2002; Park, Byrd y Boguraev, 2002), bases de datos léxicas (Nakamura y Nagao, 1988) herramientas de búsqueda de respuestas (Saggion y Gaizauskas, 2004; Cui, Kan y Chua, 2005), como apoyo para aplicaciones terminológicas (Meyer, 2001; Sierra

ISSN 1135-5948

et al., 2006), o el desarrollo de aplicaciones de aprendizaje en línea (Westerhout y Monachesi, 2007; Espinosa-Anke, 2013).

Este artículo presenta un trabajo orientado a la ED para el catalán basado en aprendizaje automático. En primer lugar, se describe la compilación y creación de un corpus extraído de Viquipèdia[1] (la Wikipedia catalana). Partimos de la idea de que un concepto monosémico con una entrada en Viquipèdia será definido en la primera oración de su artículo. Además, consideramos como no definiciones otras oraciones del mismo artículo en el que el término también aparece, pues si bien aportan información relevante para el término definido, ésta no es factual e independiente del contexto en el que se enuncia. De hecho, en muchos casos son "definiciones falsas sintácticamente plausibles" (Navigli, Velardi y Ruiz-Martínez, 2010), es decir, oraciones que muestran un comportamiento sintáctico muy similar al de una definición.

Nuestro corpus de entrenamiento consiste en un subcorpus de la rama en catalán de Wikicorpus (Reese et al., 2010), mientras que el corpus de test consiste en un subconjunto de una versión en catalán del WCL dataset (Navigli, Velardi y Ruiz-Martínez, 2010). El corpus de entrenamiento consiste en 10375 definiciones y 8010 no definiciones. Por otra parte, el corpus de test incluye 1407 no definiciones y 2796 definiciones.

El proceso de aprendizaje automático se basa en el uso del algoritmo Conditional Random Fields (CRF) (Lafferty, McCallum, y Pereira, 2001), específicamente el *toolkit* CRFsuite (Okazaki, 2007). CRF es un algoritmo de etiquetado secuencial que permite la incorporación de rasgos de observaciones y transiciones no sólo adyacentes, sino también de larga distancia, previos y posteriores a la observación actual. Es apropiado para nuestra tarea ya que consideramos la ED como una tarea de etiquetado secuencial, en la que cada palabra puede estar al principio, dentro o fuera de una definición.

El resto del artículo se estructura de la siguiente manera: La Sección 2 repasa trabajo anterior en el campo de la clasificación de definiciones, además de la tarea específica de ED. A continuación, la Sección 3 describe el proceso de compilación de los corpora utilizados en este estudio, además de los rasgos aplicados en el modelado de datos y su justificación. La Sección 4 muestra los resultados obtenidos con distintas configuraciones. Finalmente, las secciones 5 y 6 ofrecen un a visión general y resumida del trabajo descrito en este artículo, y señalan futuras líneas de investigación en el ámbito de ED para lenguas minoritarias, respectivamente.

2 Estado de la Cuestión

La Extracción automática de Definiciones (ED) es una tarea que consiste en identificar oraciones que incluyen información definitoria (Navigli y Velardi, 2010). El procesado automático de textos con fines terminográficos puede constituir una herramienta que facilite la elaboración de glosarios o diccionarios especializados, bases de datos de conocimiento léxico o bien para la elaboración de ontologías (Alarcón, 2009).

El estudio de la relación entre un concepto unívoco y monosémico y su definición se remonta al modelo aristotélico de definición, el modelo *genus et differentia*, en el que un término es definido mencionando su género próximo además de un conjunto de características particulares. Generalmente, se conoce al término definido como *definiendum*, y al clúster de palabras que le define, *definiens*.

Partiendo de este modelo, el concepto de definición se ha ido elaborando con la inclusión de distintos factores. Por ejemplo, Trimble (1985) propone una clasificación basada en el grado de formalidad y la información que se transmite, Auger y Knecht (1997) se refieren a *enunciados de interés definitorio* como la inclusión de aspectos como el sentido o contexto de uso, entre otros, y su relación con un concepto o idea específicos. Un estudio destacable en esta línea es el de Meyer (2001), que acuña el concepto de *contexto rico en conocimiento* (CRC), definido como "naturally occurring utterances that explicitly describe attributes of domain-specific concepts or semantic relations holding between them at a certain point in time, in a manner that is likely to help the reader of the context understand the concept in question". Destacamos en el ámbito del estudio y extracción automática de CRCs el trabajo de Feliu y Cabré (2002), en el que se propone un sistema basado en reglas para la identificación y clasificación automática de CRCs para el catalán, basado en la combinación de patro-

[1] https://ca.wikipedia.org/

nes léxico-sintácticos y medidas estadísticas para evaluar la prominencia de un candidato a término en un contexto concreto. Por otra parte, en ED existe una creciente tendencia a aplicar algoritmos de aprendizaje automático (Del Gaudio, Batista, y Branco, 2013).

A continuación se describe el método empleado para la construcción del corpus utilizado para entrenar y evaluar nuestro sistema, así como la descripción de los rasgos lingüísticos y estadísticos utilizados para el modelado de datos.

3 Método

En esta sección se describen los datos utilizados para entrenar y evaluar el sistema de ED y los rasgos utilizados en el modelado de los datos.

3.1 Los datasets

Tomamos como base un corpus de definiciones y oraciones no definitorias sobre un determinado término (Navigli, Velardi y Ruiz-Martínez, 2010). A partir de él, se realiza un proceso de mapeo entre aquellos términos que aparecen en el corpus original y su equivalente en Viquipèdia. Se aplican una serie de reglas para evitar ruido y evitar mapeos en blanco dado que existen términos en Wikipedia sin una entrada equivalente en otro idioma, en este caso el catalán.

A continuación se muestran dos ejemplos de oraciones definitorias y no definitorias en catalán, presentes en nuestro corpus, para el término *iot* (yate).

- **Def**: Un iot és una embarcació d'esbarjo o esportiva propulsada a vela o a motor amb coberta i amb cabina per a viure-hi.
- **Def**: *Un yate es una embarcación de recreo o deportiva propulsada a vela o a motor con cubierta y cabina para vivir.*
- **Nodef**: Tot i això la majoria de iots a vela privats solen tenir una eslora de 7 a 14m, ja que el seu cost augmenta ràpidament en proporció a l'eslora.
- **Nodef**: *Sin embargo, la mayoría de yates a vela privados suelen tener una eslora de 7 a 14m, ya que su coste aumenta rápidamente en proporción a la eslora.*

Para llevar a cabo el entrenamiento del sistema, compilamos un corpus a partir de la rama catalana de Wikicorpus (Reese et al., 2010), siguiendo el mismo método que para el corpus de evaluación. Para cada término y su correspondiente artículo se extrae una oración definitoria (la primera). Para obtener las no definitorias se extraen aquellas oraciones en las que el término también aparece, con el fin de introducir un contexto en el que el número de distractores sea elevado e incrementar así la dificultad de la tarea.

3.2 Diseño Experimental

El preprocesado de los corpus se realiza con el etiquetador morfológico presente en Freeling (Carreras et al., 2004). Dado que vamos a explotar el potencial de Conditional Random Fields para etiquetado secuencial, los rasgos propuestos se aplican a nivel de token, y no a nivel de oración.

Partimos de una oración $s = f_1, f_2, ...f_n$, en la que cada f_i es un vector de rasgos que se corresponde con una palabra, y que recibe una etiqueta BIO dependiendo de si se encuentra al principio (Beginning), dentro (Inside) o fuera (Outside) de una definición. Este esquema de etiquetado permitirá, a posteriori, evaluar el rendimiento del algoritmo para cada etiqueta, siendo la etiqueta "B" un elemento clave en el ámbito de la detección de definiciones, dado que la primera palabra de una frase que contiene una definición es probable que sea (parte del) definiendum. A continuación, se describen los rasgos utilizados durante la fase de entrenamiento.

- **Surface**: La forma superficial de la palabra, tal y como aparece originalmente en el texto.
- **Lemma**: Forma lematizada de la palabra.
- **PoS**: Categoría gramatical.
- **Pos_Prob**: Probabilidad asignada por Freeling a la categoría gramatical de cada palabra.
- **BIO_NP**: En primer lugar, se aplica un filtro lingüístico para identificar sintagmas nominales. A continuación, se asignan etiquetas BIO a dichos sintagmas. Así, una oración quedaría etiquetada de la siguiente manera:
  - El[B-NP] verd[I-NP] és[O-NP] un[O-NP] dels[O-NP] tres[O-NP] colors[B-NP] primaris[I-NP] additius[I-NP] .[O-NP]

- *El[B-NP] verde[I-NP] es[O-NP] uno[O-NP] de[O-NP] los[O-NP] tres[O-NP] colores[B-NP] primarios[I-NP] aditivos[I-NP] .[O-NP]*

- **Def-TF**: La frecuencia de la palabra en las definiciones del corpus de entrenamiento.
- **Gen-TF**: La frecuencia de la palabra en un corpus de ámbito general, extraído del subcorpus catalán de *HC Corpora*[2], formado por documentos del género periodístico. Partimos de la hipótesis de que ciertas palabras o expresiones utilizados habitualmente en definiciones pueden ser poco frecuentes en texto del dominio general, como "es considera" (*se considera*) o "es defineix com" (*se define como*).
- **Def-TFIDF**: Se computa la métrica *Term Frequency - Inverse Document Frequency* para la palabra, considerando su frecuencia en las definiciones del corpus de entrenamiento, y tomando cada oración como documento. Esta métrica se define como:

$$\text{tfidf}(w, d, D) = \text{tf}(w, d) \times \text{idf}(w, D)$$

donde $\text{tf}(w, d)$ es la frecuencia de la palabra w en el documento d. Asimismo, $\text{idf}(w, D)$ se define como

$$\frac{|D|}{|\{d \in D : w \in d\}|}$$

donde D es el la colección de documentos y $|D|$ su cardinalidad.

- **Gen-TFIDF**: Tomando el mismo enfoque que en el rasgo anterior, computamos esta métrica para cada palabra tomando como referencia el corpus mencionado en el rasgo **Gen-TF**.
- **Termhood**: Esta métrica determina el grado de importancia de un candidato unipalabra a término en un dominio concreto (Kit y Liu, 2008), midiendo su frecuencia en un corpus general y un corpus específico. Se obtiene a través de la siguiente fórmula:

$$\text{Termhood(w)} = \frac{\text{r}_\text{D}(\text{w})}{|\text{V}_\text{D}|} - \frac{\text{r}_\text{B}(\text{w})}{|\text{V}_\text{B}|}$$

Donde r_D es el ránking por frecuencia de la palabra w en un corpus específico (en nuestro, caso, el corpus de entrenamiento), y r_B es el ránking por frecuencia de dicha palabra en el corpus general. Los denominadores se refieren al tamaño de cada corpus.

- **BIO_D y BIO_d**: En cada oración (definición o no) del corpus de entrenamiento, tomamos el primer verbo y asignamos la etiqueta *definiendum* (D) a lo que le precede. Por su parte, las palabras que le siguen reciben la etiqueta *definiens* (d). A continuación, identificamos los sintagmas nominales siguiendo el mismo procedimiento que en el rasgo BIO_NP. Así, por ejemplo, la oración anterior quedaría etiquetada:

El[B-definiendum] verd[i-definiendum] és[O-definiens] un[O-definiens] dels[O-definiens] tres[O-definiens] colors[B-definiens] primaris[I-definiens] additius[I-definiens] .[O-definiens]

- **Definitional prominence**: Introducimos la noción de prominencia definitoria, con el objetivo de establecer la probabilidad de una palabra w de aparecer en una oración definitoria ($s = \text{def}$). Para ello, consideramos su frecuencia en definiciones y oraciones no definitorias del corpus de entrenamiento en la siguiente ecuación:

$$\text{DefProm(w)} = \frac{\text{DF}}{|\text{Defs}|} - \frac{\text{NF}}{|\text{Nodefs}|}$$

donde $\text{DF} = \sum_{i=0}^{i=n}(s_i = \text{def} \wedge w \in s_i)$ y $\text{NF} = \sum_{i=0}^{i=n}(s_i = \text{nodef} \wedge w \in s_i)$.

- **Definiendum prominence**: Partiendo de la hipótesis de que el hecho de que una palabra aparezca frecuentemente en posición de posible *definiendum* puede ser un indicador de su papel en definiciones, este rasgo viene dado por

$$\text{DP(w)} = \frac{\sum_{i=0}^{i=n} w_i \in \text{term}_D}{|DT|}$$

[2] http://www.corpora.heliohost.org/

donde $term_D$ es un sintagma nominal (es decir, un candidato a término) que aparece en posición de definiendum. Finalmente, |DT| se refiere al tamaño del corpus de terminología de definienda.

- **Definiens prominence**: Este rasgo consiste en la misma ecuación que en el caso anterior, esta vez considerando términos que aparecen en posición de posible definiens.

El algoritmo CRF itera sobre cada uno de los vectores y aprende combinaciones de los rasgos descritos. Estas combinaciones se establecen de antemano, por ejemplo, para aprender como rasgo la combinación lema + termhood de la palabra anterior y la combinación lema + categoría gramatical + definitional_prominence de la palabra actual. La Sección 4 describe los resultados obtenidos tras llevar a cabo experimentos realizados con varias configuraciones de rasgos.

4 Evaluación

Se han realizado experimentos que combinan los rasgos descritos en la sección 3.2, así como su combinatoria. Ofrecemos resultados en términos de Precisión, Cobertura, y F-Measure para cada una de las clases consideradas (B, O, I), además de una media de las 3. Éstas se aplican a nivel de palabra, y de la forma más restrictiva posible. Es decir, que se considera un error cuando el algoritmo predice correctamente que una palabra se encuentra en una definición, pero asigna una categoría incorrecta (es decir, Beginning en vez de Inside o viceversa). Con estas consideraciones, realizamos cuatro configuraciones experimentales, a saber:

- **Baseline**: Esta configuración sólo considera la forma superficial del token en la iteración actual.
- **C-1**: Se toman en cuenta rasgos lingüísticos (forma superficial, lema, categoría gramatical y pertenencia a sintagma nominal) sobre una ventana [i-3:i+3], siendo i la posición de la iteración actual.
- **C-2**: Se toman únicamente rasgos estadísticos (tf-def, tf-gen, tfidf-def, tfidf-gen, termhood, definitional prominence, definiendum prominence y definiens prominence). La ventana es la misma que para **C-1**.
- **C-3**: Toma en cuenta todos los rasgos.

La Tabla 1 muestra los resultados obtenidos con estas cuatro configuraciones. Los identificadores de las filas se refieren a: (1) Si el resultado es en precisión (P), cobertura (C), o F-Measure (F), y (2) si la evaluación corresponde a la etiqueta Beginning (B), Inside (I), Outside (O), o a la media de las tres (M), que en definitiva refleja el comportamiento general del sistema propuesto en este artículo. Se puede observar que a partir de una baseline que obtiene un 67.31 de F-Measure, ésta es superada por la combinación de rasgos lingüísticos (en C-1), que obtiene un F=75.85, y rasgos estadísticos (C-2), que llega a F=75.68. La combinación de ambos conjuntos de rasgos obtiene resultados altamente competitivos (F=86.69), lo cual sugiere que ambos conjuntos de rasgos son informativos y contribuyen al proceso de aprendizaje.

	Baseline	C-1	C-2	C-3
P-B	67.50	89.29	80.68	**93.60**
C-B	51.72	57.47	**88.62**	85.87
F-B	58.57	69.93	84.47	**89.57**
P-I	58.49	84.89	72.25	**90.71**
C-I	49.58	51.82	**88.80**	83.48
F-I	53.67	64.35	79.68	**86.95**
P-O	88.03	**89.19**	77.24	79.36
C-O	91.43	**97.76**	53.08	88.23
F-O	89.79	**93.28**	62.92	83.56
P-M	71.34	87.78	76.72	**87.89**
C-M	64.24	69.01	76.83	**85.85**
F-M	67.31	75.85	75.68	**86.69**

Tabla 1: Resultados obtenidos en término de precisión, cobertura y F-measure

4.1 Discusión

A la luz de los resultados obtenidos, cabe destacar el importante papel que juegan los rasgos lingüísticos (categoría gramatical, lema y pertenencia o no a un sintagma nominal), ya que observamos un mejor rendimiento de un modelo entrenado sólo con rasgos de este tipo, en comparación con un sistema entrenado sólo con rasgos estadísticos. No obstante, la combinación de rasgos de ambos tipos contribuye al mayor rendimiento de las tres configuraciones propuestas.

Observando algunos de las instancias incorrectamente clasificadas, se observa una tendencia al sobre-entrenamiento con respecto a las definiciones basadas en el modelo *genus et differentia* propio de Viquipèdia. Por ejemplo, la definición del término "gas natural" es:

- El gas natural és una font d'energia fòssil que, com el carbó o el petroli, està constituïda per una barreja d'hidrocarburs, unes molècules formades per àtoms de carboni i hidrogen.
- *El gas natural es una fuente de energía fósil que, como el carbón o el petróleo, está constituída por una mezcla de hidrocarburos, unas moléculas formadas por átomos de carbono e hidrógeno.*

Por su parte, en nuestro test set, además de la oración anterior, existe el siguiente distractor:

- **El gas natural és una** energia primària, o que es pot obtenir directament sense **transformació**
- ***El gas natural es una*** *energía primaria, o que se puede obtener directamente sin* ***transformación***

Las palabras resaltadas en negrita fueron incorrectamente marcadas como pertenecientes a una definición por nuestro clasificador. Dos conclusiones se pueden extraer de casos como éste: (1) Existen relaciones semánticas entre conceptos que podrían ser consideradas como definitorias, según un criterio ligeramente más laxo, y esto se refleja en algunos de los falsos negativos obtenidos en nuestra evaluación; (2) Asumimos que en el futuro sería deseable contar con una heurística post-clasificación (o un segundo proceso de clasificación) para realizar un segundo proceso de clasificación sobre palabras que o bien han sido clasificadas con poca probabilidad, o bien tienen palabras próximas clasificadas de otra manera. Por ejemplo, en una oración en la que el 80 % de las palabras han sido clasificadas como no definitorias, es razonable asumir que si el 20 % fueron clasificadas como definitorias, la probabilidad de que ésta sea una clasificación incorrecta es elevada.

Finalmente, con respecto a los rasgos estadísticos, su introducción al proceso de entrenamiento, si bien contribuyen a mejorar el sistema, no parece que puedan constituir la base de un sistema de extracción de definiciones, o al menos, no sería recomendable descartar rasgos estadísticos. A continuación, se describen algunas de las posibles razones por las que los rasgos estadísticos propuestos son susceptibles de mejora:

- La falta de un paso previo en identificación de terminología para generalizar los términos en posición de definiendum. Esto provoca que sólo aquellos términos multipalabra con alguna palabra repetida se benefician de métricas que toman en cuenta la frecuencia de sus componentes. Éste es el caso, por ejemplo, de nombres propios que comparten algún apellido o definienda que comparten algún término (por ejemplo, en especies de peces como: **ammodytes** tobianus, **ammodytes** immaculatus o **ammodytes** marinus, entre otros).
- Posible falta de representatividad de los corpus de referencia. Si bien el dominio en el que se ha desarrollado este estudio es homogéneo en el género textual (enciclopédico), podemos afirmar que se trata de un estudio no delimitado a un dominio concreto. Nuestra hipótesis es que métricas como *tfidf* o *definitional_prominence* serían más informativas aplicadas a dominios concretos. De hecho, en el campo de ED, salvo contadas excepciones (Snow, Jurafsky y Ng, 2004; Velardi, Navigli y D'Amadio, 2008; Cui, Kan y Chua, 2005), la tendencia es desarrollar y evaluar sistemas en corpora pertenecientes a un dominio específico.

5 Conclusiones

En este trabajo se ha descrito un sistema de extracción de definiciones para el catalán. Los corpora de entrenamiento y test, similar a los datasets descritos en Navigli, Velardi y Ruiz-Martínez (2010), son obtenidos de Viquipèdia. El corpus de entrenamiento es un subcorpus de la rama catalan de Wikicorpus (Reese et al., 2010), mientras que el corpus de test ha sido validado manualmente. Afrontamos el problema como una tarea de clasificación secuencial supervisada, en la que a cada palabra se le asigna la etiqueta BIO, dependiendo de si el sistema predice que se encuentra al principio (Beginning), dentro (Inside) o fuera (Outside) de una definición.

Utilizamos el algoritmo Conditional Random Fields, y combinamos rasgos lingüísticos y estadísticos, obteniendo resultados razonablemente prometedores.

6 Trabajo Futuro

Tras haber explorado el proceso de desarrollar un sistema de ED para el catalán, a continuación se enumeran algunos aspectos susceptibles de mejora, además de posibles líneas de trabajo futuro: (1) un estudio de la relevancia de los distintos rasgos y su combinatoria ayudaría a desarrollar un sistema más eficiente y que no considerara rasgos escasamente discriminatorios. (2) Realizar experimentos con distintos corpus de referencia, y con distintos ratios entre definiciones y no definiciones ayudaría a valorar la influencia de estos datasets en el proceso de aprendizaje. Finalmente, (3) aplicar distintos algoritmos de ED que existen para el inglés daría una idea del comportamiento de nuestro sistema en un contexto comparativo.

También creemos que el proceso de evaluación (en este trabajo y en ED en general) se puede desarrollar y especificar más si: (1) se realiza una evaluación de contenido. Este enfoque nos permitiría determinar si se han dado casos de definiciones extraídas parcialmente. Y (2), si se realizara una evaluación en distintas etapas del flujo (*pipeline*) del sistema, como por ejemplo, para la identificación de definienda y definiens.

Agradecimientos

Agradecemos a los revisores anónimos sus comentarios y sugerencias. Este trabajo ha sido parcialmente financiado por el proyecto número TIN2012-38584-C06-03 del Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, España.

Bibliografía

Alarcón, Rodrigo. 2009. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Ph.D. tesis, Universitat Pompeu Fabra.

Auger, Alain y Pierre Knecht. 1997. Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles.

Bontas, Elena Paslaru y Malgorzata Mochol. 2005. Towards a cost estimation model for ontology engineering. En Rainer Eckstein y Robert Tolksdorf, editores, *Berliner XML Tage*, páginas 153–160.

Carreras, Xavier, Isaac Chao, Lluis Padró, y Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. En *LREC*.

Cui, Hang, Min-Yen Kan, y Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. En *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 384–391. ACM.

Del Gaudio, Rosa, Gustavo Batista, y António Branco. 2013. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, páginas 1–33.

Espinosa-Anke, Luis. 2013. Towards definition extraction using conditional random fields. En *Proceedings of RANLP 2013 Student Research Workshop*, páginas 63–70.

Feliu, Judit y M Teresa Cabré. 2002. Conceptual relations in specialized texts: new typology and an extraction system proposal. En *TKE 02, 6th International Conference in Terminology and Knowledge Engineering*, páginas 45–49.

Kit, Chunyu y Xiaoyue Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2).

Lafferty, John D., Andrew McCallum, y Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. En *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, páginas 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279.

Muresan, A y Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. En *Proceedings*

of the Language Resources and Evaluation Conference (LREC.

Nakamura, Jun-ichi y Makoto Nagao. 1988. Extraction of semantic information from an ordinary english dictionary and its evaluation. En *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, páginas 459–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, Roberto y Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, páginas 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, Roberto, Paola Velardi, y Juana María Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 3716–3722, Valletta, Malta. European Language Resources Association (ELRA).

Okazaki, Naoaki. 2007. Crfsuite: a fast implementation of conditional random fields (CRFs).

Park, Youngja, Roy J. Byrd, y Branimir K. Boguraev. 2002. Automatic Glossary Extraction: Beyond Terminology Identification. En *Proceedings of the 19th International Conference on Computational Linguistics*, páginas 1–7. Association for Computational Linguistics.

Reese, Samuel, Gemma Boleda, Montse Cuadros, Lluís Padró, y German Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. En *LREC*. European Language Resources Association.

Saggion, Horacio y Robert Gaizauskas. 2004. Mining on-line sources for definition knowledge. En *17th FLAIRS*, Miami Bearch, Florida.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, y Alberto Barrón. 2006. Towards the building of a corpus of definitional contexts. En *Proceeding of the 12th EURALEX International Congress, Torino, Italy*, páginas 229–40.

Snow, Rion, Daniel Jurafsky, y Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge Language Teaching Library.

Velardi, Paola, Roberto Navigli, y Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25, Septiembre.

Westerhout, Eline y Paola Monachesi. 2007. Extraction of Dutch definitory contexts for elearning purposes. *Proceedings of the Computational Linguistics in the Netherlands (CLIN 2007), Nijmegen, Netherlands*, páginas 219–34.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 77-84 recibido 15-04-14 revisado 30-06-14 aceptado 02-07-14

The aid of machine learning to overcome the classification of real health discharge reports written in Spanish

Aportaciones de las técnicas de aprendizaje automático a la clasificación de partes de alta hospitalarios reales en castellano

Alicia Pérez, Arantza Casillas, Koldo Gojenola, Maite Oronoz, Nerea Aguirre, Estibaliz Amillano
IXA Taldea, University of the Basque Country (UPV-EHU).
{*alicia.perez*, *arantza.casillas*, *koldo.gojenola*, *maite.oronoz*}*@ehu.es*

Resumen: La red de hospitales que configuran el sistema español de sanidad utiliza la Clasificación Internacional de Enfermedades Modificación Clínica (ICD9-CM) para codificar partes de alta hospitalaria. Hoy en día, este trabajo lo realizan a mano los expertos. Este artículo aborda la problemática de clasificar automáticamente partes reales de alta hospitalaria escritos en español teniendo en cuenta el estándar ICD9-CM. El desafío radica en que los partes hospitalarios están escritos con lenguaje espontáneo. Hemos experimentado con varios sistemas de aprendizaje automático para solventar este problema de clasificación. El algoritmo Random Forest es el más competitivo de los probados, obtiene un F-measure de 0.876.
Palabras clave: Procesamiento del Lenguaje Natural, Biomedicina, Aprendizaje Automático

Abstract: Hospitals attached to the Spanish Ministry of Health are currently using the International Classification of Diseases 9 Clinical Modification (ICD9-CM) to classify health discharge records. Nowadays, this work is manually done by experts. This paper tackles the automatic classification of real Discharge Records in Spanish following the ICD9-CM standard. The challenge is that the Discharge Records are written in spontaneous language. We explore several machine learning techniques to deal with the classification problem. Random Forest resulted in the most competitive one, achieving an F-measure of 0.876.
Keywords: Natural Language Processing, Biomedicine, Machine Learning

1 Introduction

Thousands of Discharge Records and, in general, Electronic Health Records (EHRs) are produced every year in hospitals. These records contain valuable knowledge sources for further diagnoses, association and allergy development reporting in a population. Apart from the documentation services from the hospitals there are other interests behind mining biomedical records, amongst others from the insurance services. In (Lang, 2007) it is stated that the cost of assigning ICD-9 codes to clinical free texts is $25 billion per year in the US. In particular, this work tackles EHR classification according to Diagnostic Terms (DT). The task deals with Spanish DTs written in spontaneous language.

1.1 Bridging the gap between spontaneous and standard written language

In this particular task we deal with real files written by doctors at the consultation time. The language is not the same as that found in the biomedical literature (e.g. PubMed), in the sense that the language in these records is almost free, including misspells and syntactically incorrect phrases. Being both natural language, we shall refer to the former as *spontaneous* and to the latter as *standard* jargon. At the consultation-time the doctor is devoted to the attention and care of the patient rather than filling the record. As a result, the spontaneous language used by doctors differs from the standardly accepted jar-

ISSN 1135-5948

gon (or the written language they would use in other more relaxed circumstances).

The language gap between these two language-varieties is self-evident in many ways:

- Acronyms: the adoption of non standard contractions for the word-forms.
- Abbreviations: the prefix of the words terminated with a dot.
- Omissions: often prepositions and articles are omitted in an attempt to write the word-form quickly. The verbs are often omitted.
- Synonyms: some technical words are typically replaced by others apparently more frequently used while possibly not that specific.
- Misspells: sometimes words are incorrectly written.

Examples of the aforementioned issues are gathered in Table 1.

1.2 Goals and challenges

In this work we tackle a particular classification problem associated with written spontaneous language processing. We devote to the classification of the discharge records. We are focusing on the records produced at the Galdakao-Usansolo Hospital (attached to the Spanish public hospital-system). These records convey information relative to:

- Personal details of the patient, admission and discharge date, counsellor, etc. In order to preserve the confidentiality, we do not count on this part of the records (it was removed beforehand).
- A narrative summary of the admission details, antecedents, referred main problems, treatment undergone, findings, recommended care plan, etc. This body-part is completely unstructured, since it does not count on sections to extract particular information from. Besides, not all the aforementioned pieces of information are necessarily present in all the records.
- The diagnostic terms together with their associated code in the International Classification of Diseases 9 Clinical Modification[1] (ICD-9-CM). Note that it is the ICD-9-CM that is being followed so far in the hospitals attached to the Spanish Ministry of Health, Social Services and Equality. Admittedly, in some countries the ICD-10 is being used.

453.40 Embolia y trombosis venosa aguda de vasos profundos no especificados de extremidad inferior ▪ TVP MID ▪ TVP POPLITEO_FEMORAL MII
600.00 Hipertrofia (benigna) de próstata sin obstrucción urinaria ni otros síntomas del tracto urinario inferior (STUI) ▪ HBP ▪ Hipertrofia de Prostata
530.81 Reflujo esofágico ▪ E.R.G.E.
332 Enfermedad de Parkinson ▪ Enf de Parkinson
536.8 Dispepsia y otros trastornos especificados del funcionamiento del estómago ▪ Dispesia alta
185 Neoplasia maligna de la próstata ▪ ca prostata

Table 1: Examples revealing the differences between standard and spontaneous writing. The ICD-9 code appears next to the standard DT, and below spontaneous forms that were assigned the same ICD-9 code are shown.

The aim of the text mining task in which we are focusing on is to get the discharge reports automatically classified by their diagnostic term (DT). That is, the goal is to de-

[1]The International Classification of Diseases 9 Clinical Modification in Spanish is accessible through the web in the Spanish Ministry `http://eciemaps.mspsi.es/ecieMaps/browser/index_9_mc.html`

sign a decision support system to assign an ICD-9-CM code to each DT in the records. So far, a set of experts are in charge of getting the records classified. Hence, all the work is carried out by hand, and our goal is to help to automatize this process. Our aim is to develop a computer aided classification system with very high precision. Addressing this process as a classification problem entails a major challenge: given that the entire ICD-9-CM is being considered, the problem conveys a very large-scale classification system (note that the ICD-9-CM gathers thousands of different classes). Moreover, precision is crucial in this process, and that is, indeed, why we do not aspire to get a fully automatic system.

1.3 State of the art and contributions

Since 1990 the task of extracting ICD-9 codes from clinical documents has become relevant. In 2007 the BioNLP workshop a shared task on multi-label classification of clinical texts was organised (Pestian et al., 2007). For this task it was developed the CMC dataset, consisting of 1954 radiology reports arising from outpatient chest X-ray and renal procedures, observed to cover a substantial portion of paediatric radiology activity. It covered a total of 45 unique codes. The best system of the competition achieved a micro-average F-score of 0.89 and 21 of the 44 participating systems scored between 0.8 and 0.9.

By contrast to the works presented in BioNLP, our work focuses on automatic generation of ICD-9 codes from DTs written in spontaneous Spanish language. We do not examine the whole document. Another relevant difference is that we deal with a problem of an order of magnitude bigger (we envisage more than 678 classes and achieve similar performance). We have also tried different inferred classifiers.

In (Ferrao et al., 2012), they propose a methodology encompassing EHR data processing to define a feature set and a supervised learning approach to predict ICD-9-CM code assignment. Four supervised learning models decision trees, naïve Bayes, logistic regression and support vector machines were tested and compared using fully structured EHR data. By contarst, our data lacks of structure.

The contribution of this work is to delve into real EHR classification on Spanish language. First, we collected a set of real EHRs written in Spanish, and got them fully anonimized. There are works in the literature aiming at overcoming the gap between spontaneous and standard language on the biomedical domain, yet, few of them deal with real EHRs. In this work we explore several machine learning techniques, train them on real EHRs, and assess their performance. Some machine-learning techniques have proven to be able to deal with this big-scale classification problem with quite high precision.

1.4 Arrangement

The rest of the paper is arranged as follows: Section 2 presents the inferred classifiers used in this work and also the means of representing the instances to get them inferred; Section 3 is devoted to present the experimental layout; finally, concluding remarks and some ideas for future work are given in Section 4.

2 Machine Learning

In brief, given a set of discharge records, we focus on the DTs and try to automatically assign the associated ICD-9 code. At first, we thought (and possibly the reader might do now) that this task could be neatly tackled by means of quite a naive system that would simply look up the given DT in the ICD-9-CM catalogue. Nevertheless, we were not aware yet of the aforementioned gap between spontaneous and standard jargon. Indeed, we proceed with this approach and extremely poor results were achieved: only 0.96% of the DTs within the evaluation set were found in the ICD-9-CM catalogue even after applying little modifications such as re-casing, accepting omission of write-accents, getting rid of multiple spaces and allowing to delete the punctuation marks (amongst others). As an alternative, we applied several machine learning techniques in this task.

Bearing in mind the language gap, we tried to approach this task by matching the spontaneous DTs not against the standard DTs from the ICD-9-CM catalogue, but against other sets of data in spontaneous language. That is, the system would learn from previously classified records. All together, this problem can be seen as a supervised classification process, and to that end, we count, in fact, on a set of previously classified set of data.

In this work, we explore four inferred clas-

sifiers that have proven successful in text mining problems. All of them were implemented using the libraries available in Weka-6.9 (Hall et al., 2009). Weka is an open-source software that implements a number of machine-learning algorithms, evaluation metrics and other helpful methods.

The machine learning approaches considered in this work are the following ones:

NB Naive Bayes

DT Decision Tree

RF Random Forest

SVM Support Vector Machines

Next, a full description of the machine learning schemes explored, as well as the motivation to do so are presented. The learning scheme and a few important details on the parameters selected for each of them are given. Also in this section, the operational description of the instances used to train the models are given.

2.1 Naïve Bayes

Within a general-framework on a probabilistic approach, the classification problem could be tackled as a maximum likelihood estimation problem:

$$\widehat{C} = \arg\max_{C_k \in \mathcal{C}} p(C_k|\mathbf{x}) = \quad (1)$$

$$= \arg\max_{C_k \in \mathcal{C}} \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_{C_j \in \mathcal{C}} p(\mathbf{x}|C_j)} \quad (2)$$

Where the $\mathcal{C}$ is the set of possible classes (in our problem, the set of all the ICD-9 codes that can be given as output), and $\mathbf{x}$ represents the observations (in our problem, the operational representation of the input DT). In our context, each instance $\mathbf{x} \in \Sigma^N$ being Σ the input vocabulary. Besides, $\mathcal{C}$ comprises all the ICD-9 codes (since we are not restricting ourselves to any particular subset such as paediatrics as other works in the literature did).

Admittedly, we are dealing with a large-scale classification problem. In fact, if there are $D = |\mathbf{x}|$ inputs and each of them might take $|\Sigma|$ values, a general distribution would correspond to an application of Σ^D possible values for each class (with a constraint imposed by the total probability theorem). In an attempt to make this problem affordable, the naive-Bayes assumption is made: the features in $\mathbf{x}$ are conditionally independent given the class $C_k \in \mathcal{C}$.

These models were explored as a baseline, since they are efficient and besides they were successful in a number of text mining problems such as spam classifiers in short messages (Sriram et al., 2010; Peng et al., 2012) and also in biomedical classification (Soni et al., 2011; Rodríguez et al., 2012). Nevertheless, for our task it did not result to be competitive enough.

These models were implemented by means of the the `classifiers.bayes.NaiveBayes` library included in Weka (Hall et al., 2009).

2.2 Decision Tree

Decision Tree inference is based on the C4.5 algorithm (Quinlan, 1993). This technique follows a divide and conquer strategy recursively. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the Information Gain (IG), as described in eq. (3).

$$IG(\mathcal{X}, A) = H(\mathcal{X}) - H(\mathcal{X}|A) = \quad (3)$$

$$= H(\mathcal{X}) - \sum_{v \in Val(A)} \frac{|\mathcal{X}_v|}{|\mathcal{X}|} H(\mathcal{X}_v)$$

where:

- $H(\mathcal{X})$ represents the entropy of the set of instances $\mathcal{X}$ with respect to the class. Likewise, $H(\mathcal{X}|A)$ represents the entropy of the set given the attribute A.
- $Val(A)$ is the set of all the possible values for attribute A.
- $\mathcal{X}_v = \{\mathbf{x} \in \mathcal{X} : \mathbf{x} \cdot A = v\}$ represents the set of instances that take the value v on the attribute A.

In plain words, IG measures the expected reduction in the entropy of the set $\mathcal{X}$ given an attribute A (Mitchell, 1997), and hence, it quantitatively measures the worth of keeping that attribute.

Once an attribute is selected, the set of training samples is divided into sub-sets (according to the value that the samples take for that attribute). The same criterion is recursively applied to each sub-set until convergence according to a an impurity measure (a threshold on the IG). As a result, a tree

structure is generated, where the attribute with the highest IG is chosen to make the decision at each stage.

These models were implemented by means of the the `classifiers.trees.J48` library included in Weka (Hall et al., 2009). Besides, a parameter dealing with the impurity, the minimum number of instances in the leaf nodes, was fine-tuned so as to optimize the f-measure on the training set. As a result this parameter was set to 2.

2.3 Random Forest

Random Forest (RF) consists of a variety of ensemble models. RF combines a number of decision trees. The trees involved were close to the optimum tree, yet some randomness was introduced in the order in which the nodes are generated. Particularly, each time a node is generated in the tree, instead of choosing the attribute that minimized the error (instead of Information Gain), the attribute is randomly selected amongst the k best attributes. This randomness enhances the generalization ability of the trees, while the overfitting is avoided. Next, consensus is achieved to decide which class to vote.

These models were implemented by means of the the `classifiers . trees . RandomForests` library included in Weka (Hall et al., 2009). Besides, a parameter relative to the number of trees comprised in the forest was fine tuned so as to optimize the f-measure on the training set. As a result, 9 trees were selected.

2.4 Support Vector Machines

Support Vector Machines (SVMs) are kernel-based models that lay on sparse solutions. The predictions for new inputs rely upon the kernel function evaluated at a subset of the training data points. The parameters defining the model are chosen in a convex optimization problem (local solution is also a global optimum). In SVMs the decision boundary is chosen in such a way that the margin is maximized. That is, if there are multiple solutions that cope with the training data set without errors, the one with the smallest generalization error is chosen (Bishop, 2006).

These models were implemented by means of the the `classifiers.functions.SMO` library included in Weka (Hall et al., 2009). It implements John Platt's sequential minimal optimization algorithm for training a support vector classifier (Platt, 1999). Nevertheless, there exist other more powerful approaches such as LibSVM (Chang and Lin., 2001).

2.5 Operational description of the instances

As it is well-known, the success of the techniques based on Machine Learning relies, amongst others, upon the features used to describe the instances. In this work the operational description of the DTs was done in the same way for all the techniques explored. Admittedly, each technique would be favored by one or another sort of features. Thus, in order to make the most of each learning scheme, appropriate features should be adopted for each of them.

Originally, in the training set the samples are described using a string of variable length to define the DT and a nominal class. That is, while the set of DTs might be infinite, the classes belong to a finite-set of values (all of the ICD-codes admitted within the ICD-9-CM catalogue). In brief, each instance from the supervised set consists of a tuple $(\mathbf{s}, C) \in \Sigma^* \times \mathcal{C}$ being Σ the input vocabulary or a finite-set of words in the input language (hence, Σ^* represents its free monoid) and $\mathcal{C}$ a finite-set of classes.

First of all, a pre-processing was defined to deal with simple string formatting operations. This pre-processing is denoted as h in eq. (4). The application h defines an equivalence class between: lower/upper-case words; strings with and without written accents;...

$$\begin{array}{rcl} h : \Sigma^* \times \mathcal{C} & \longrightarrow & \Sigma^* \times \mathcal{C} \\ (\mathbf{s}, C) & \longrightarrow & (\mathbf{s'}, C) \end{array} \quad (4)$$

The pre-processing defined by h enables the mapping of equivalent strings written in slightly different ways (as it is frequent in spontaneous writing).

Due to the fact that many methods are not able to deal with string-type of features, the transformation f, defined in eq. (5) was applied next.

$$\begin{array}{rcl} f : \Sigma^* \times \mathcal{C} & \longrightarrow & \mathcal{X} \times \mathcal{C} \\ (\mathbf{s}, C) & \longrightarrow & (\mathbf{x}, C) \end{array} \quad (5)$$

Where $\mathcal{X} = 2^{|\Sigma|}$.

The application f acts as a filter. It transforms each string $\mathbf{s}$ (a sequence of words with

precedence constraint) into a binary vector referred to the terms of the vocabulary Σ. That is, the element x_i is a binary feature that expresses whether the term $t_i \in \Sigma$ is present in the string $\mathbf{s}$ or not.

The application f is capable of describing each instance by their words as elements. This approach is also referred to as *Bag of Words* (BOW) in the sense that the string is mapped as a set of words without preserving the order, and thus, loosing the involved n-grams. While the precedence of the words is lost, the application allows a simple though effective representation for the instances. Besides, this approach enables a computationally efficient data structure representing the instances as sparse vectors. Hence, the dimensionality of Σ does not convey any computational course.

Note that the DTs consist of short strings with a high semantic component in each word and simple syntax. Intuitively, it is the keywords that matters above the syntax, this is the motivation behind using the BOW as operational description for the instances. Moreover, applying the filter f to the set of instances the dimension of the problem is made affordable since the free monoid comprises all the string whatever their length: $\Sigma^* = \bigcup_{i=0}^{\infty} \Sigma^i$

3 Experimental framework

3.1 Task and corpus

We count on a set of DTs written in spontaneous language extracted from real discharge records that were manually coded. The entire set of instances was randomly divided into two disjoint sets for training and evaluation purposes, referred to as Train and Eval respectively.

Table 2 provides a quantitative description of the Train and Eval sets. Each (DT, ICD-9) pair belongs to the pre-processed set of instances, formally denoted as $\mathcal{X} \times \mathcal{C}$ (with the preprocess described in Section 2.5). The first row shows the number of different instances, formally denote as $|\mathcal{X} \times \mathcal{C}|$; the second row, shows the number of different DTs, formally denoted as $|\mathcal{X}|$; the third row, shows the number of different ICD-9 codes, denoted as $|\mathcal{C}|$; the fourth row shows the number of features or relevant words in the vocabulary of the application, formally denoted as $|\Sigma|$.

Note that the number of instances is higher than the number of different ICD-9 codes. This means that some DTs are taken as equivalent, in the sense that different strings were assigned the same ICD-9 code.

	Train	**Eval**
Different instances	6,302	1,588
Different DTs	6,085	1,554
Different ICD-9 codes	1,579	678
$\|\Sigma\|$	4,539	

Table 2: Quantitative description of the training and evaluation sets.

On the other hand, since we are working with real data some diseases are more frequent than the others, this makes that some pairs appear more frequently. For example, there are around 3,500 pairs occurring only once, 500 occurring 3 times, and the ratio decreases exponentially, that is, there are very few pairs with high frequency. The distribution is not exactly the same for the DTs or for the ICD-codes, hence, the corpus shows some ambiguities. For example, the code `185` mentioned in Table 1, appears 22 times in the corpus, 17 DTs are different, amongst them, we can see:

- `Adenocarcinoma de próstata con bloqueo hormonal`
- `Ca. próstata metastásico`

3.2 Evaluation metrics

On what the evaluation metrics regards, the following evaluation metrics were considered:

Pr: precision

Re: recall

F1-m: f1-measure

It must be clarified that, given the large amount of classes, the results associated to each class are not provided, instead, a per-class average (weighted by the number of instances from each class) is given (as it is implemented in Weka libraries denoted as *weighted average* for each metric). Per-class averaging means that the number of instances in each class contributes as a weighting factor on the number of true-positives and negatives for that class.

3.3 Results

A twofold evaluation was carried out:

1. **Hold-out evaluation:** the model was trained on the Train set and the predictive power assessed on the Eval set.

2. **Re-substitution error:** the model was trained on the Train set and the predictive power assessed on the Train set. The quality of the training data, the difficulty and the ability of the learning techniques are limited and rarely provide an accuracy of 100%. We could not expect to overcome this threshold on an unseen evaluation set. Hence, in an attempt to get to know the maximum performance achievable on this task, we assessed the performance of the models on the Train set. That is, we explored the predictions exactly on the same set used to train the models. The error derived from this method are the so-called re-substitution error.

On account of this, Table 3 shows the performance of each model (the nomenclature for each model was given in Section 2) on either the Eval or the Train set.

Set	Model	Pr	Re	F1-m
Eval	NB	0.163	0.181	0.131
	DT	0.854	0.851	0.843
	RF	**0.883**	0.881	0.876
	SVM	0.880	**0.889**	**0.878**
Train	NB	0.328	0.394	0.312
	DT	0.905	0.909	0.902
	RF	**0.969**	**0.970**	**0.967**
	SVM	0.959	0.964	0.959

Table 3: Performance of different inferred classifiers on both the evaluation set and also on the training set itself as an upper threshold of the performance.

3.4 Discussion

We proposed the use of Naive Bayes as a baseline system (since it has proven useful in other text mining tasks such as in spam detection), yet, for this task with so many classes has resulted in very poor results. Amongst the explored ML techniques Random Forest presents the highest quality, yet with no significant difference with respect to Support Vector Machines. It is well worth mentioning that the highest f1-measure resulted in 0.876, satisfactorily enough, the upper threshold is not far from that (to be precise, the highest achievable f1-measure is 0.967).

Random Forest comprises 9 Decision Trees, and can be seen as an ensemble model made up of homogeneous classifiers (that is, Decision Trees). Note that the quality provided by a single Decision Tree is nearly the precision achieved by the Random Forest with substantially lower cost.

For this task it is crucial to achieve very high precision, and the Random Forest offers very high precision. Still, on a decision support system we would strive towards 100% precision. Hence, the presented system seems to be much benefitial as a computer aided decision support system, but not yet as an automatic classification system.

It is well-worth endeavoring towards an automatic classification system. Nevertheless, there are evident shortcomings, there are pragmatic limits on this task as it can be derived from the upper performance achievable (see Table 3). Admittedly, it is disappointing not to get an almost-null re-substitution error. A manual inspection of the Train set revealed that the corpus itself had several errors, in the sense that we observed that almost identical DTs had associated different ICD-9 codes. It is quite common not to get flawless datasets, and above all, when they are spontaneous. Moreover, we presented a source of ambiguity in Section 3.1. Possibly, the cause behind these errors might have to do with the conversion from electronic health records to the set of instances. Hence, for future work we will delve into the outlier detection in our training set.

4 Concluding remarks

4.1 Conclusions

This work tackles the classification of discharge records for their DT following the ICD-9-CM standard. The classification problem is quite tough for several reasons: 1) the gap between spontaneous written language and standard jargon; and 2) it is a large-scale classification system (being the number of possible classes the number of different diseases within the ICD-9-CM catalogue). There are few works facing this problem, and the authors are not aware of any in Spanish.

While a look-up in the standard ICD-9-CM provided very poor results, machine learning techniques, trained on spontaneous data resulted very competitive. Due to patient privacy it is difficult to find datasets of clinical documents for free use, this is most evident in the case of clinical text written in Spanish. We would like to remark the impor-

tance of harvesting this sort of corpus on the quality of the developed systems.

Amongst the techniques explored, Random Forest resulted the most competitive one (slightly over Support Vector Machines). The best system showed high-quality, an f1-measure of 0.876, being 0.967 the upper threshold for the expected achievable f-1 measure. It would be a great deal to strive towards improving both the hold-out evaluation and its upper boundary.

4.2 Future work

Currently we are working on enhancing the set of features by defining an equivalence class between synonyms derived from the SNOMED-CT (SNOMED-CT, 2012).

In the near future we will delve into the outlier detection in our training set so as to strive into 100% precision on the Train set. The aim will be to filter the outliers so that they do not do harm the inference process.

In this work we explored several ML schemes working alone. Nevertheless, ensemble learning has proven successful in recent research-challenges or competitions. For future work, we mean to double-check if the aforementioned classifiers complement each other and jointly get to improve the performance. Together with this, it could be useful to adapt the features to describe the DTs to each particular learning scheme and also to apply feature subset selection techniques.

As it is the case for speech recognition, we might try to overcome the spontaneous language gap by means of a language model trained on spontaneous data.

Acknowledgments

Authors would like to thank the Hospital Galdakao-Usansolo for their contributions and support, in particular to Javier Yetano, responsible of the Clinical Documentation Service. We would like to thank Jon Patrick for his kind comments on the feature transformation stages and assessment. This work was partially supported by the European Commission (SEP-210087649), the Spanish Ministry of Science and Innovation (TIN2012-38584-C06-02) and the Industry of the Basque Government (IT344-10).

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Chang, C. C. and C. J. Lin. 2001. Libsvm: a library for support vector machines.

Ferrao, J. C., M. D. Oliveira, F. Janela, and H.M.G. Martins. 2012. Clinical coding support based on structured data stored in electronic health records. In *Bioinformatics and Biomedicine Workshops, 2012 IEEE International Conference on*, pages 790–797.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Lang, D. 2007. Natural language processing in the health care industry. Consultant report, Cincinnati Children's Hospital Medical Center.

Mitchell, T. 1997. *Machine Learning*. McGraw Hill.

Peng, H., C. Gates, B. Sarma, N. Li, Y. Qi, R. Potharaju, C. Nita-Rotaru, and I. Molloy. 2012. Using probabilistic generative models for ranking risks of android apps. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 241–252. ACM.

Pestian, J. P., C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. Bretonnel Cohen, and W. Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104. Association for Computational Linguistics.

Platt, J. C. 1999. Fast training of support vector machines using sequential minimal optimization. *MIT press*.

Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rodríguez, J. D., A. Pérez, D. Arteta, D. Tejedor, and J. A. Lozano. 2012. Using multidimensional bayesian network classifiers to assist the treatment of multiple sclerosis. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):1705–1715.

SNOMED-CT. 2012. SNOMED CT User Guide. January 2012 International Release. Technical report, International Health Terminology Standards Development Organisation.

Soni, J., U. Ansari, D. Sharma, and S. Soni. 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17.

Sriram, B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Herramientas de Procesamiento del Lenguaje Natural

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 87-94 recibido 10-04-14 revisado 15-07-14 aceptado 15-07-14

ParTes. Test Suite for Parsing Evaluation *

ParTes: Test suite para evaluación de analizadores sintácticos

Marina Lloberes
Irene Castellón
GRIAL-UB
Gran Via Corts Catalanes 585
08007 Barcelona
marina.lloberes@ub.edu
icastellon@ub.edu

Lluís Padró
TALP-UPC
Jordi Girona 1-3
08034 Barcelona
padro@lsi.upc.edu

Edgar Gonzàlez
Google Research
1600 Amphitheatre Parkway
94043 Mountain View - CA
edgargip@google.com

Resumen: En este artículo se presenta ParTes, el primer test suite en español y catalán para la evaluación cualitativa de analizadores sintácticos automáticos. Este recurso es una jerarquía de los fenómenos representativos acerca de la estructura sintáctica y el orden de argumentos. ParTes propone una simplificación de la evaluación cualitativa contribuyendo a la automatización de esta tarea.
Palabras clave: test suite, evaluación cualitativa, analizador sintáctico, español, catalán

Abstract: This paper presents ParTes, the first test suite in Spanish and Catalan for parsing qualitative evaluation. This resource is a hierarchical test suite of the representative syntactic structure and argument order phenomena. ParTes proposes a simplification of the qualitative evaluation by contributing to the automatization of this task.
Keywords: test suite, qualitative evaluation, parsing, Spanish, Catalan

1 Introduction

Qualitative evaluation in Natural Language Processing (NLP) is usually excluded in evaluation tasks because it requires a human effort and time cost. Generally, NLP evaluation is performed with corpora that are built over random language samples and that correspond to real language utterances. These evaluations are based on frequencies of the syntactic phenomena and, thus, on their representativity, but they usually exclude low-frequency syntactic phenomena. Consequently, current evaluation methods tend to focus on the accuracy of the most frequent linguistic phenomena rather than the accuracy of both high-frequent and low-frequent linguistic phenomena.

This paper takes as a starting point these issues related to qualitative evaluation. It presents ParTes, the first parsing test suite in Spanish and Catalan, to allow automatic qualitative evaluation as a complementary task of quantitative evaluation. This resource is designed to simplify the issues related to qualitative analysis reducing the human effort and time cost. Furthermore, ParTes provides a set of representative linguistic utterances based on syntax. The final result is a hierarchical test suite of syntactic structure and argument order phenomena defined by means of syntactic features.

2 Evaluation databases

Traditionally, two analysis methods have been defined: the quantitative analysis and the qualitative analysis. Both approaches are complementary and they can contribute to a global interpretation.

The main difference is that quantitative analysis relies on statistically informative data, while qualitative analysis talks about richness and precision of the data (McEnery and Wilson, 1996).

Representativeness by means of frequency is the main feature of quantitative studies. That is, the observed data cover the most frequent phenomena of the data set. Rare phenomena are considered irrelevant for a quantitative explanation. Thus, quantitative descriptions provide a close approximation of

* The resource presented in this paper arises from the research project SKATeR (Ministry of Economy and Competitiveness, TIN2012-38584-C06-06 and TIN2012-38584-C06-01). Edgar Gonzàlez collaborated in the ParTes automatization process. We thank Marta Recasens for her suggestions.

ISSN 1135-5948

the real spectrum.

Qualitative studies offer an in-depth description rather than a quantification of the data (McEnery and Wilson, 1996). Frequent phenomena and marginal phenomena are considered items of the same condition because the focus is on providing an exhaustive description of the data.

In terms of analysis methods and databases, two resources have been widely used: corpora and test suites. Language technologies find these resources a reliable evaluation test because they are coherent and they are built over guidelines.

A corpus contains a finite collection of representative real linguistic utterances that are machine readable and that are a standard reference of the language variety represented in the resource itself (McEnery and Wilson, 1996). From this naive conceptualization, Corpus Linguistics takes the notion of representativeness as a presence in a large population of linguistic utterances, where the most frequent utterances are represented as a simulation of the reality and they are annotated according to the resource goals. That is why corpora are appropriate test data for quantitative studies.

On the other hand, test suites are structured and robust annotated databases which store an exhaustive collection of linguistic utterances according to a set of linguistic features. They are built over a delimited group of linguistic utterances where every utterance is detailed and classified according to rich linguistic and non-linguistic annotations (Lehmann et al., 1996). Thus, the control over test data and their detailed annotations make test suites a perfect guidance for qualitative studies.

Corpora have also been used in qualitative analysis, but they collect representative linguistic utterances by means of frequency rather than the representative linguistic utterances by means of exhaustiveness. Then, they are not the most appropriate tool for qualitative studies.

3 Existing test suites

Traditional test suites were simple collections of linguistic test cases or interesting examples. However, with the success of the NLP technologies, there was a real need for developing test suites based on pre-defined guidelines, with a deep structure, richly annotated and not necessarily developed for a particular tool (Flickinger, Nerbonne, and Sag, 1987). For this reason, the new generation of test suites are databases that cover the real needs of the NLP software evaluation (Lehmann et al., 1996).

The HP test suite (Flickinger, Nerbonne, and Sag, 1987) is an English and general purpose resource developed to diagnose and monitor the progress of NLP software development. The main goal of this test suite is to evaluate the performance of heuristic-based parsers under development. The suite contains a wide-range collection of linguistic examples that refer to syntactic phenomena such as argument structure verbs and verbal subcategorization among others. It also includes some basic anaphora-related phenomena. Furthermore, these phenomena are represented by a set of artificially constructed sentences and the annotations are shallow. This resource has a minimal internal classification since the suite organizes the test data under headings and sub-headings.

In order to step further, subsequent test suites have been developed as in-depth resources with rich structure and annotations. One of the groups of EAGLES proposes a set of guidelines for evaluating grammar checkers based on test suites (EAGLES, 1994). The test suite is a collection of attributes that allow to validate the quality of the functions of the evaluated tool. It is derived from a taxonomy of errors, where each error class is translated into a feature which is collected in the test suite. The final result is a classification of sentences containing an error, the corresponding sentence without the error, the name of the error and the guidelines for the correction process.

The TSNLP (Lehmann et al., 1996) is a multilingual test suite (English, French and German) richly annotated with linguistic and meta-linguistic features. This test suite is a collection of test items with general, categorial and structural information. Every test item is classified according to linguistic and extra-linguistic features (e.g. number and type of arguments, word order, etc.). These test items are also included in test sets by means of positive and negative examples. Furthermore, the TSNLP includes information about frequency or relevance for a particular domain.

In Spanish, a previous test suite exists

for NLP software evaluation, the SPARTE test suite (Peñas, Álvaro, and Verdejo, 2006). Specifically, it has been developed to validate Recognizing Textual Entailment systems and it is a collection of text and hypothesis pairs with true/false annotations. Although SPARTE and the presented ParTes in Spanish (ParTesEs) are resources for the same language, both test suites have been developed for different purposes which make both resources unique. With respect to the Catalan language, the version of ParTes in Catalan (ParTesCa) is the first test suite for this language.

4 The construction of ParTes

ParTes is a new test suite in Spanish and Catalan for qualitatively evaluating parsing systems. This test suite follows the main trends on test suite design, so that it shares some features with the EAGLES test suite (EAGLES, 1994) and the TSNLP (Lehmann et al., 1996).

Additionally, ParTes adds two new concepts in test suite design concerning how the data are classified and which data are encoded. The test suite is seen as a hierarchy where the phenomenon data are explicitly connected. Furthermore, representativeness is the key-concept in ParTes to select the phenomenon-testing data that configure the test suite.

The ParTes guidelines are created to ensure the coherence, the robustness and the easy implementation of this resource.

Specific purpose. While some test suites are general purpose like TSNLP, ParTes is a specific purpose test suite. Particularly, it is focused to validate the accuracy of the syntactic representations generated by parsers. For this reason, the test cases are related to syntactic phenomena and the test suite has been annotated with several syntactic features.

Test suite of syntactic phenomena. ParTes is not a simple collection of linguistic test cases nor a set of linguistic features, actually. This resource lists the syntactic phenomena that configure a language by a set of syntactic features.

For example, ParTes collects syntactic structures based on head-child relation. It also contains several features that syntactically define every phenomenon (e.g. the syntactic category of the head or the child, the syntactic relation with the node that governs it, etc.). Complementarily, every phenomenon is associated with a test case that corresponds to the linguistic utterance of the actual phenomenon described and that is used to evaluate the accuracy of the performance of the parser.

Hierarchy of syntactic phenomena. Previous test suites were a collection of test sentences, optionally structured (EAGLES and TSNLP). ParTes proposes a hierarchically-structured set of syntactic phenomena to which tests are associated.

Polyhedral hierarchy. Test suites can define linguistic phenomena from several perspectives (e.g. morphologic features, syntactic structures, semantic information, etc.). Because ParTes is built as a global test suite, it defines syntactic phenomena from two major syntactic concepts: syntactic structure and argument order (Section 5).

Exhaustive test suite. In order to evaluate NLP tools qualitatively, test suites list exhaustively a set of linguistic samples that describe in detail the language(s) of the resource, as discussed in Section 2. ParTes is not an exception and it contains an exhaustive list of the covered syntactic phenomena of the considered languages. However, some restrictions are applied to this list. Otherwise, listing the whole set of syntactic phenomena of a language is not feasible, and it is not one of the goals of the test suite's design.

Representative syntactic phenomena. As mentioned, lists of test cases need to be delimited because test suites are controlled data sets. Similarly to corpora development, the syntactic phenomena to be included in the test suite can be selected according to a certain notion of representativeness. Consequently, representative syntactic phenomena are relevant for testing purposes and they should be added in the test suite, whereas peripheral syntactic phenomena can be excluded. The next section (Section 5) details the definition of representativeness in ParTes and how it is implemented.

Rich annotations. Every syntactic phenomenon of ParTes is annotated with precise information that provides a detailed description and that allows the qualitative interpretation of the data. The annotations refer to

several linguistic and extra-linguistic features that determine the syntactic phenomena.

Controlled data. As argued in Section 2, there is a direct relation between qualitative evaluation, test suites and controlled test data. Because ParTes is a test suite for qualitative evaluation, there is a strong control over the test data and, specifically, the control is applied in a double way. The number of test cases is limited to human-processing size. The sentences of the test cases are controlled to avoid ambiguities and interactions with other linguistic utterances. For this reason, test cases are artificially created.

Semi-automatically generated. Linguistic resources usually have a high cost in terms of human effort and time. For this reason, automatic methods have been implemented whenever it has been possible. Manual linguistic description of the syntactic structure has been the main method to annotate the syntactic phenomena related to the structure. On the other hand, argument order annotations have been automatically generated and manually reviewed, using the automatization process of the SenSem corpus (Fernández and Vàzquez, 2012).

Multilingual. The architecture of this resource allows it to be developed in any language. The current version of ParTes includes the Spanish version of the test suite (ParTesEs) and the Catalan version (ParTesCa).

5 The results of ParTes

The final result of ParTes is an XML hierarchically and richly annotated test suite of the representative syntactic phenomena of the Spanish (ParTesEs) and Catalan (ParTesCa) languages. This resource is the first test suite for the evaluation of parsing software in the considered languages. It is freely available[1] and distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.

ParTes is built over two kinds of information: the test suite module with the syntactic phenomena to be evaluated and the test data module with the linguistic samples to evaluate over. Since it is a polyhedral test suite, it is organized according to two major concepts in Syntax: structure and order. Table 1 gives the size of the current version of ParTes.

[1] http://grial.uab.es/descarregues.php

Section	**ParTesEs**	**ParTesCa**
Structure	99	101
Order	62	46
Total	**161**	**147**

Table 1: ParTes in numbers

5.1 Syntactic structure

The structure section is a hierarchy of syntactic levels where each level receives a tag and it is associated to a set of attributes that define several aspects about the syntactic structure. This section is placed between the `<structure></structure>` tags and it is organized into the following parts:

`<level>` It can be intrachunk (i.e. any structure inside a chunk) or intraclause (i.e. any connection between a clause marker and a grammatical category, phrase or clause).

`<constituent>` Phrase or clause that determines the nature of the constituent (e.g. noun phrase, verb phrase, infinitive clause, etc.). The head of the constituent corresponds to the parent node.

`<hierarchy>` Given two connected constituents, it defines which one occurs in the parent position and which other one in the child position.

`<realization>` Definition of the attributes of the head or child:

- `id`: Numerical code that identifies every `<realization>`.
- `name`: Name of the gramatical category, phrase or clause that occurs in head or child position (e.g. noun, pronoun, etc., as heads of noun phrase).
- `class`: Specifications about the gramatical category, the phrase or the clause that occurs in head or child position (e.g. a nominal head can be a common noun or a proper noun).
- `subclass`: Sub-specifications about the gramatical category, the phrase or the clause that occur in head or child position (e.g. a nominal head can be a bare noun).
- `link`: Arch between parent and child expressed by Part of Speech tags (e.g. the link between a nominal head and a modifying adjective is 'n-a').

```
<constituent name="verbphrase">
   <hierarchy name="head">
      <realization id="0001" name="verb" class="finite" subclass="default" link="null"
                   parent="salir" child="null" freq="null"
                   test="Saldrán"/>
      <realization id="0002" name="verb" class="nonfinite" subclass="default" link="null"
                   parent="viajar" child="null" freq="null"
                   test="Hubiesen viajado"/>
   </hierarchy>
   <hierarchy name="child">
      <realization id="0003" name="verb" class="auxiliar" subclass="haber" link="v-v"
                   parent="vender" child="haber" freq="0.010655" test="Habrán vendido la casa"/>
      <realization id="0004" name="verb" class="auxiliar" subclass="ser" link="v-v"
                   parent="acusar" child="ser" freq="0.010655"
                   test="Es acusada de robo"/>
      ...
      <realization id="0009" name="noun" class="null" subclass="default" link="v-n"
                   parent="romper" child="taza" freq="0.131629"
                   test="La taza se rompió"/>
      <realization id="0010" name="adjective" class="null" subclass="default" link="v-a"
                   parent="considerar" child="innovador" freq="0.010373"
                   test="Se considera una propuesta innovadora"/>
            ...
   </hierarchy>
</constituent>
```

Figure 1: Syntactic structure of the verb phrase in ParTesEs

- `parent`: Lemma of the upper level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the parent is 'casa').
- `child`: Lemma of the lower level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the child is 'caro').
- `freq`: Relative frequency in the AnCora corpus of the link between the two nodes defined in `link`.
- `test`: Linguistic test data that illustrates the syntactic structure.

For example, in the definition of verb phrase as `<constituent name="verbphrase">` (Figure 1), the possible grammatical categories, phrases and clauses that can form a verb phrase are detected and classified into two categories: those pieces that can be the head of the verb phrase (`<hierarchy name="head">`) and those that occur in child position (`<hierarchy name="child">`).

Next, the set of the possible heads of the verb phrase are listed in the several instances of `<realization>`. Furthermore, all the candidates of the child position are identified.

Every realization is defined by the previous set of attributes. In the Figure 1, in the case where the realization of one of the verb phrase children is a noun (`<realization ... name="noun".../>`), the frequency of occurrence of this link (i.e. the link of a verbal head and a nominal child, `link="v-n"`) is 0.131629 (in a scale between 0 and 1) and the test case to represent this structure is 'La taza se rompió' ('The cup broke'). Furthermore, the parent of the link 'v-n' of the test case is the lemma of the finite verb form 'rompió' (`parent="romper"`, 'to break') and the child of this link is the substantive 'taza' (`child="taza"`, 'cup'). The rest of this realization's attributes are empty.

As mentioned in Section 4, the most representative syntactic structure phenomena have been manually collected. In order to determine which phenomena are relevant to be included in ParTes, linguistic descriptive grammars have been used as a resource in the decision process. Thus, the syntactic phenomena that receive a special attention in the descriptive grammars can be considered candidates in terms of representativeness. In particular, the constructions described in *Gramática Descriptiva de la Lengua Española* (Bosque and Demonte, 1999) and in *Gramàtica del Català Contemporani* (Solà et al., 2002), for Spanish and Catalan respectively, have been included.

In addition, the representativeness of the selected syntactic phenomena is supported by the frequencies of the syntactic head-child re-

lations of the AnCora corpus (Taulé, Martí, and Recasens, 2008). These frequencies are automatically extracted and they are generalizations of the Part of Speech tag of both head and child given a link: all the main verb instances are grouped together, the auxiliaries are recognized into the same class, etc. Some frequencies are not extracted due to the complexity of certain constructions. For example, comparisons are excluded because it is not possible to reliable detect them by automatic means in the corpus.

The representation of the syntactic structures in ParTes follows the linguistic proposal implemented in FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). This proposal states that the nature of the lexical unit determines the nature of the head and it determines the list of syntactic categories that can occur in the head position.

5.2 Argument order

Similarly to the syntactic structure section, the argument order schemas are also a hierarchy of the most representative argument structures that occur in the SenSem corpus. This section is organized in ParTes as follows:

`<class>` Number and type of arguments in which an order schema is classified. Three classes have been identified: monoargumental with subject expressed (`subj#V`), biargumental where subject and object are expressed (`subj#V#obj`), and monoargumental with object expressed (`V#obj`).

`<schema>` Sub-class of `<class>` where the argument order and the specific number of arguments are defined. For example, ditransitive verbs with an enclitic argument (e.g. '[El col·leccionista$_{subj}$] no [li$_{iobj}$] [ven$_{v}$] [el llibre$_{dobj}$]' - 'The collector to him do not sell the book') are expressed by the schema `subj#obj#V#obj` (Figure 2).

`<realization>` Specifications of the argument order schema, which are defined by the following set of attributes (Figure 2):

- `id`: Numerical code that identifies every `<realization>`.
- `func`: Syntactic functions that define every argument of the argument order schema. In Figure 2, the argument schema is composed by subject (`subj`), preverbal indirect object (`iobj`) and postverbal direct object (`dobj`).
- `cat`: Grammatical categories, phrases or clauses that define every argument of the argument order schema. For example, the three arguments of Figure 2 are realized as noun phrases (`np`).
- `parent`: Lemma of the upper level node of the argument order schema. In the case illustrated in Figure 2, the parent corresponds to the lemma of the verbal form of the test case (i.e. 'vendre'-'to sell').
- `children`: Lemmas of the lower level nodes of the argument order schema. In the test case of Figure 2, the children are the head of every argument (i.e. 'col·leccionista'-'collector', 'ell'-'him', 'llibre'-'book').
- `constr`: Construction type where a particular argument order schema occurs (active, passive, pronominal passive, impersonal, pronominal impersonal). In Figure 2, the construction is in active voice.
- `sbjtype`: Subject type of a particular argument order schema (semantically full or empty and lexically full or empty). The subject type of Figure 2 is semantically and lexically full so the value is `full`.
- `freq`: Relative frequency of the argument order schema in the SenSem corpus (Fernández and Vàzquez, 2012). The frequency of the ditransitive argument schema in Figure 2 is 0.005176, which means that the realization `subj#iobj#V#dobj` occurs 0.005176 times (in a scale between 0 and 1) in the SenSem corpus.
- `idsensem`: Three random SenSem id sentences have been linked to every ParTes argument order schema.
- `test`: Linguistic test data of the described realization of the argument order schema (in Figure 2, 'El col·leccionista no li ven el llibre'-'The collector to him do not sell the book').

The ParTes argument order schemas have been automatically generated from the syntactic patterns of the annotations of the SenSem corpus (Fernández and Vàzquez, 2012). Specifically, for every annotated verb

```
<class name="subj#V#obj">
   <schema name="subj#obj#V#obj">
      <realization id="0140" func="subj#iobj#v#dobj" cat="np#np#v#np" parent="vendre"
                   children="col·leccionista#ell#llibre" constr="active" sbjtype="full"
                   freq="0.005176" idsensem="43177#45210#52053"
                   test="El col·leccionista no li ven el llibre"/>
   </schema>
</class>
```

Figure 2: Argument order of ditransitive verbs in ParTesCa

in the corpus, the argument structure has been recognized. This information has been classified into the ParTes argument order schemas. Finally, the most frequent schemas have been filtered and manually reviewed, considering those schemas above the average. The total set of candidates is 62 argument order schemas for Spanish and 46 for Catalan.

5.3 Test data module

ParTes contains a test data set module to evaluate a syntactic tool over the phenomena included in the test suite. For the sentences in the data set, both plain text and syntactic annotations are available. The test data set is controlled in size: ParTesEs contains 94 sentences and ParTesCa is 99 sentences long. It is also controlled in terms of linguistic phenomena to prevent the interaction with other linguistic phenomena that may cause incorrect analysis. For this reason, test cases are artificially created.

A semi-automated process has been implemented to annotate ParTesEs and ParTesCa data sets. Both data sets have been automatically analyzed by the FreeLing Dependency Parser (Lloberes, Castellón, and Padró, 2010). The dependency trees have been mapped to the CoNLL format (Figure 3) proposed for the shared task on multilingual dependency parsing (Buchholz and Marsi, 2006). Finally, two annotators have reviewed and corrected the FreeLing Dependency Parser mapped outputs.

6 ParTes evaluation

To validate that ParTes is a useful evaluation parsing test suite, an evaluation task has been done. ParTes test sentences have been used to evaluate the performance of Spanish and Catalan FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). The accuracy metrics have been provided by the CoNLL-X Shared Task 2007 script (Buchholz and Marsi, 2006), in which the syntactic analysis generated by the FreeLing Dependency Grammars (*system output*) are compared to ParTes data sets (*gold standard*).

The global scores of the Spanish Dependency Grammar are 82.71% for LAS[2], 88.38% for UAS and 85.39% for LAS2. Concerning to the Catalan FreeLing Dependency Grammar, the global results are 76.33% for LAS, 83.38% for UAS and 80.98% LAS2.

A detailed observation of the ParTes syntactic phenomena shows that FreeLing Dependency Grammars recognize successfuly the root of the main clause (Spanish: 96.8%; Catalan: 85.86%). On the other hand, subordinate clause recognition is not perfomed as precise as main clause recognition (Spanish: 11%; Catalan: 20%) because there are some limitations to determine the boundaries of the clause, and the node where it should be attached to.

Noun phrase is one of the most stable phrases because it is formed and attached right most of times (Spanish: 83%-100%; Catalan: 62%-100%). On the contrary, prepositional phrase is very unstable (Spanish: 66%; Catalan: 49%) because the current version of the grammars deals with this syntactic phenomenon shallowly.

This evaluation has allowed to determine which FreeLing Dependency Grammars syntactic phenomena are also covered in ParTes (*coverage*), how these syntactic phenomena are performed (*accuracy*) and why these phenomena are performed right/wrong (*qualitative analysis*).

7 Conclusions

The resource presented in this paper is the first test suite in Spanish and Catalan for parsing evaluation. ParTes has been de-

[2] *Labeled Attachment Score* (LAS): the percentage of tokens with correct head and syntactic function label; *Unlabeled Attachment Score* (UAS): the percentage of tokens with correct head; *Label Accuracy Score* (LAS2): the percentage of tokens with correct syntactic function label.

```
1       Habrán          haber       VAIF3P0     _       _       2       aux
2       vendido         vender      VMP00SM     _       _       0       top
3       la              el          DA0FS0      _       _       4       espec
4       casa            casa        NCFS000     _       _       2       dobj
5       .               .           Fp          _       _       2       term
```

Figure 3: Annotation of the sentence 'Habrán vendido la casa' ('[They] will have sold the house')

signed to evaluate qualitatively the accuracy of parsers.

This test suite has been built following the main trends in test suite design. However, it also adds some new functionalities. ParTes has been conceptualized as a complex structured test suite where every test case is classified in a hierarchy of syntactic phenomena. Furthermore, it is exhaustive, but exhaustiveness of syntactic phenomena is defined in this resource as representativity in corpora and descriptive grammars.

Despite the fact that ParTes is a polyhedral test suite based on the notions of structure and order, there are more foundations in Syntax, such as syntactic functions that currently are being included to make ParTes a more robust resource and to allow more precise evaluation tasks.

In addition, the current ParTes version contains the test data set annotated with syntactic dependencies. Future versions of ParTes may be distributed with other grammatical formalisms (e.g. constituents) in order to open ParTes to more parsing evaluation tasks.

References

Bosque, I. and V. Demonte. 1999. *Gramática Descriptiva de la Lengua Española*. Espasa Calpe, Madrid.

Buchholz, S. and E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164.

EAGLES. 1994. Draft Interim Report EAGLES. Technical report.

Fernández, A. and G. Vàzquez. 2012. Análisis cuantitativo del corpus SenSem. In I. Elorza, O. Carbonell i Cortés, R. Albarrán, B. García Riaza, and M. Pérez-Veneros, editors, *Empiricism and Analytical Tools For 21st Century Applied Linguistics*. Ediciones Universidad Salamanca, pages 157–170.

Flickinger, D., J. Nerbonne, and I.A. Sag. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories, Cambridge, England.

Lehmann, S., S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvy, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP – Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2, pages 711–716.

Lloberes, M., I. Castellón, and L. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 693–699.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Peñas, A., R. Álvaro, and F. Verdejo. 2006. SPARTE, a Test Suite for Recognising Textual Entailment in Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pages 275–286.

Solà, J., M.R. Lloret, J. Mascaró, and M. Pérez-Saldanya. 2002. *Gramàtica del Català Contemporani*. Empúries, Barcelona.

Taulé, M., M.A. Martí, and M. Recasens. 2008. AnCora: Multi level annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation*, pages 96–101.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 95-101 recibido 15-04-14 revisado 06-06-14 aceptado 12-06-14

PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems *

Anotación morfosintáctica de la Web en portugués. Variedades nacionales, tipologías textuales y sistemas ortográficos

Marcos Garcia and **Pablo Gamallo**
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
{marcos.garcia.gonzalez, pablo.gamallo}@usc.es

Iria Gayo
Cilenis Language Technology
iria.gayo@cilenis.com

Miguel A. Pousada Cruz
Universidade de Santiago de Compostela
miguelangel.pousada@usc.es

Resumen: La gran cantidad de texto producido diariamente en la Web ha provocado que ésta sea utilizada como una de las principales fuentes para la obtención de corpus lingüísticos, posteriormente analizados utilizando técnicas de Procesamiento del Lenguaje Natural. En una escala global, idiomas como el portugués —oficial en 9 estados— aparecen en la Web en diferentes variedades, con diferencias léxicas, morfológicas y sintácticas, entre otras. A esto se suma la reciente aprobación de una ortografía unificada para las diferentes variedades del portugués, cuyo proceso de implementación ya ha comenzado en varios países, pero que se prolongará todavía durante varios años, conviviendo por lo tanto también diferentes ortografías. Una vez que los etiquetadores morfosintácticos existentes para el portugués están adaptados específicamente para una variedad nacional concreta, el presente trabajo analiza diferentes combinaciones de corpus de aprendizaje y de léxicos con el fin de obtener un modelo que mantenga una alta precisión de anotación en diferentes variedades y ortografías de esta lengua. Además, se presentan diferentes diccionarios adaptados a la nueva ortografía (Acordo Ortográfico de 1990) y un nuevo corpus de evaluación con diferentes variedades y tipologías textuales, disponibilizado libremente.
Palabras clave: anotación morfosintáctica, portugués, Web as Corpus, Acordo Ortográfico

Abstract: The great amount of text produced every day in the Web turned it as one of the main sources for obtaining linguistic corpora, that are further analyzed with Natural Language Processing techniques. On a global scale, languages such as Portuguese —official in 9 countries— appear on the Web in several varieties, with lexical, morphological and syntactic (among others) differences. Besides, a unified spelling system for Portuguese has been recently approved, and its implementation process has already started in some countries. However, it will last several years, so different varieties and spelling systems coexist. Since PoS-taggers for Portuguese are specifically built for a particular variety, this work analyzes different training corpora and lexica combinations aimed at building a model with high-precision annotation in several varieties and spelling systems of this language. Moreover, this paper presents different dictionaries of the new orthography (Spelling Agreement) as well as a new freely available testing corpus, containing different varieties and textual typologies.
Keywords: PoS-tagging, Portuguese, Web as Corpus, Spelling Agreement

* This work has been supported by the HPCPLN project – Ref: EM13/041 (Galician Government) and by the Celtic – Ref: 2012-CE138 and Plastic – Ref: 2013-CE298 projects (Feder-Interconnecta).

ISSN 1135-5948

1 Introduction

In recent years, the Web has turned the main source of corpora for extracting information about different topics in many languages. Thus, Natural Language Processing (NLP) applications take advantage of web crawlers and data mining strategies in order to analyze large amounts of textual data.

In this respect, one of the main NLP tasks to be performed is Part-of-Speech (PoS) tagging, which consists of labeling every token of a text with its correct morphosyntactic category. Specifically for Portuguese, there are several state-of-the-art PoS-taggers for different varieties, such as the European (EP) (Bick, 2000; Branco and Silva, 2004) and the Brazilian one (BP) (Aires, 2000).

However, using the Web for obtaining corpora in Portuguese involves the crawling of texts from different varieties, including the African ones, as well as the use of sources which contain a mixture of national varieties, such as the Wikipedia.

Apart from that, a Spelling Agreement (Acordo Ortográfico de 1990, AO90) for Portuguese, which unifies the spelling system of the different national varieties, has been recently approved, and its implementation process has already started in some countries. The chronology of the process differs in each country, but it is expected that the new orthography will be mandatory in Brazil and Portugal before 2016, as well as in the other countries with Portuguese as official language (ending in Cape Verde in 2020).[1]

Furthermore, some of the main journals of Brazil and Portugal adopted the AO90 spelling system since 2010 (e.g., *Diário de Notícias* and *Jornal de Notícias* in Portugal, or *Folha de São Paulo* in Brazil), while others did not (e.g. *Público*, in Portugal), so large amounts of texts are published every day using this new orthography.

Taking the above facts into account, this paper evaluates the use of different PoS-taggers, trained with several combinations of European and Brazilian resources, for analyzing the Web in Portuguese, including various linguistic varieties, textual typologies and spelling systems.

In order to carry out the evaluation, this paper also presents new lexica of the AO90 and a manually revised corpus which represents in some way the (journalistic and encyclopedic) Web in Portuguese. The corpus includes European, Brazilian and African (from Angola and Mozambique) texts with samples before and after the AO90, and is freely distributed.

The experiments show that the consistency between the training corpus and the dictionary has the major effect in the PoS-tagger performance. Concerning the lexica, it is shown that the new dictionaries can be combined to better analyze texts using the AO90 orthography, without losing precision when PoS-tagging documents in different spelling systems.

Apart from this introduction, Section 2 includes the Related Work. In Section 3, the different resources used for training the PoS-taggers are presented. Then, Section 4 describes the performed experiments and their results, while Section 5 outlines the main conclusions of this paper.

2 Related Work

Several PoS-taggers were developed for Portuguese language, namely for the European and Brazilian varieties. Some of them are statistical models trained with specific resources for each variety, while others use rule-based approaches.

Among the latter ones, PALAVRAS uses large sets of rules and a lexicon of about 50,000 lemmas for PoS-tagging (and also parsing) European Portuguese.

Marques and Lopes (2001) presented a neural-network approach for PoS-tagging, which obtain high-precision results ($\approx 96\%$) with small training corpora.

Ribeiro, Oliveira, and Trancoso (2003) compared Markov models and a transformation-based tagger (based in Brill (1995)) for PoS-tagging EP, focused on pre-processing data for a Text-To-Speech system.

In Branco and Silva (2004), the authors compare different algorithms (transformation-based (Brill, 1995), Maximum Entropy (Ratnaparkhi, 1996), Hidden Markov Models (HMM) (Tufis and Mason, 1998) and second order Markov models (Brants, 2000)) for analyzing EP. The best results (97.09%) were obtained with the transformation-based system.

In Garcia and Gamallo (2010), the HMM

[1] http://pt.wikipedia.org/wiki/Acordo_Ortografico_de_1990

tagger (Brants, 2000) of FreeLing (Padró and Stanilovsky, 2012) was adapted for European Portuguese (and also for Galician), achieving results up to 96.3%.

For Brazilian Portuguese, Aires (2000) also compared several PoS-taggers, with the best results of 90.25% with the MXPoST algorithm (Ratnaparkhi, 1996). Further development (with simplified tagsets) improved the precision up to 97%.[2]

MXPoST also obtained the best results for BP in (Aluísio et al., 2003), with a precision of about 95.92%.

Concerning annotated corpora for Portuguese, the Bosque corpus[3] contains about 138,000 tokens (20,884 unique token-tag pairs) for European Portuguese. For the Brazilian variety, the Mac-Morpho[4] corpus has 1,167,183 tokens (73,955 unique token-tag pairs) and it was used for training different models in Aluísio et al. (2003).

Finally, some lexica for Portuguese are available, and were also used for building PoS-taggers for this language. For EP, LABEL-LEX (SW)[5] Eleutério et al. (2003) includes 1,257,000 forms from about 120,000 different lemma-tag pairs.

In Brazilian Portuguese, Muniz (2004) presented the DELAF_PB[6] lexicon, which contains 878,651 forms from 61,095 lemmas.

Recently, some projects started to compile new resources for analyzing the new spelling system (AO90), such as Almeida et al. (2013) or the Portal da Língua Portuguesa.[7]

3 Linguistic Resources

In order to evaluate various models for PoS-tagging different varieties of Portuguese, the following resources were used.

3.1 Corpora

For European Portuguese, the Bosque corpus (footnote 3) was used. In particular, a version based on the one used in Garcia and Gamallo (2010) was adapted for a new tagset (see Table 1).

Category	Tag
Adjective	AD
Adverb	AV
Coordinating Conjunction	CC
Subordinating Conjunction	CS
Determiner (Definite/Indefinite)	DT
Demonstrative Determiner	DD
Possessive Determiner	DP
Preposition	PS
Verb	VB
Participle	VP
Common Noun	NC
Proper Noun	NP
Demonstrative Pronoun	PD
Exclamative Pronoun	PE
Indefinite Pronoun	PI
Personal Pronoun	PP
Relative Pronoun	PR
Interrogative Pronoun	PT
Possessive Pronoun	PX
Interjection	I
Numbers	Z
Contractions with preposition *de*	DC
Contractions with preposition *por*	PC
Punctuation	F*
Dates/Hours	W
Numerical Expressions/Quantities	Z*

Table 1: Tagset used in the experiments. Top categories are the main ones (both in lexica and in corpora). Bottom categories appear in some corpora, but not in the lexica. "*" indicates that there are several PoS-tags both for Punctuation (24 types) and for different Numerical Expressions (5 types).

For Brazilian Portuguese, the Mac-Morpho (footnote 4) was also adapted to the same tagset as the EP resource.

Finally, a new corpus (Web) containing several varieties and text typologies of Portuguese was used for evaluating the different PoS-taggers (Garcia and Gamallo, 2014). The corpus has about 52,000 tokens, and includes the following sources: three Portuguese journals, two Brazilian journals, a journal from Angola, a journal from Mozambique, and texts from the Wikipedia in Portuguese, containing texts from different varieties. Table 2 shows the details of this new resource.

[2] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

[3] http://www.linguateca.pt/Floresta/corpus.html#bosque

[4] http://www.nilc.icmc.usp.br/lacioweb/corpora.htm

[5] http://label.ist.utl.pt/pt/labellex_pt.php

[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

[7] http://www.portaldalinguaportuguesa.org/

Variety	Size	Vocab
Brazil	11,460	3,137
Portugal	13,987	3,637
Angola	4,180	1,403
Mozambique	5,517	1,700
Wikipedia	17,187	4,003
Total	52,331	9,873

Table 2: Size (in number of tokens) and vocabulary (*Vocab*, number of different token-tag pairs) of the Web corpus (and subcorpora).

The journals from Mozambique and Angola, and one from Portugal do not use the AO90 orthography, while the other Portuguese corpora and the Brazilian ones use this new spelling system. Also, Mozambique and Angola have traditionally used the EP orthography (even though they have lexical and syntactic variations). Wikipedia corpus contains texts from Brazil and Portugal, with both pre-AO90 and post-AO90 spellings.

The PoS-tags were manually corrected and also converted to the same tagset as the above mentioned corpora.

3.2 Lexica

Concerning the lexica, different resources were also used:

The version of LABEL-LEX (SW) (Eleutério et al., 2003) for FreeLing suite[8] was adapted as the lexicon for EP. This lexicon was already used in Garcia and Gamallo (2010), and it has a strong consistency with the Bosque corpus.

The DELAF_PB (footnote 6) was the lexicon used for BP.

Apart from that, two new lexica were created for evaluating their influence when PoS-tagging different varieties:

PEB_Dict: PEB_Dict is a lexicon built by merging the EP and BP ones. In order to do that, all the token-lemma-tag triples of the EP and BP dictionaries were selected. Then, every triple in the EP dictionary was added to the PEB_Dict. After that, BP triples not included in EP were also added to PEB_Dict. For functional words, which sometimes had different PoS-tags in EP and BP, the European version was preferred.

The resulting dictionary contains about $1,254,000$ token-lemma-tag triples and $1,179,000$ token-tag pairs, from $112,000$ different lemmas.

It is worth noting that this fusion may increase the ambiguity of the PoS-tagger, since some entries belong to a higher number of token-lemma-tag triples.

AO+_Dict: AO+_Dict is another merged resource containing lexical units from different varieties. In order to create it, the strategy described for the PEB_Dict was followed, merging in this case a dictionary of the AO90 (developed by the authors) and the PEB_Dict. This way, AO+_Dict consists of a PEB_Dict enriched with the new forms of the Acordo Ortográfico de 1990. Note that AO+_Dict includes entries which are not correct in AO90. AO+_Dict has about $1,277,000$ token-lemma-tag triples and $\approx 1,200,000$ token-tag pairs. The number of lemmas of this resource is about $119,000$.

The tagsets of these two lexica were also unified (Table 1). As the tagset is simpler than the original one, the number of triples was reduced in about $50,000$ in each dictionary.

4 Experiments

In order to evaluate the performance of several PoS-taggers for analyzing different varieties of Portuguese, the following experiments were carried out.

First, both European Portuguese and Brazilian Portuguese resources were used for training and testing specific EP and BP taggers (`EPtag` and `BPtag`). The performance of these models was also evaluated with the Web corpus, which contains different varieties of Portuguese before and after the AO90.

Then, several training corpora and dictionaries were combined in order to evaluate (i) how they behave with the new corpora and (ii) whether they increase or decrease the PoS-tagging precision in EP and BP corpora before the AO90.[9]

4.1 Models

The HMM PoS-tagger of FreeLing (Padró and Stanilovsky, 2012) was the selected al-

[8] `http://nlp.lsi.upc.edu/freeling/`

[9] Both training and testing corpora, labeled with the different dictionaries are freely available at `http://gramatica.usc.es/~marcos/pt_tag_corpora.tar.bz2`

gorithm for doing the experiments. It is a state-of-the-art PoS-tagger algorithm implemented in an open-source suite of linguistic analysis which also contains other modules for previous and further NLP tasks.

The European Portuguese model (`EPtag`) was trained with ≈ 83% of the EP corpus (120,007 tokens and 18,035 unique token-tag pairs), and tested in the remaining 17% (with 23,102 tokens and 5,873 unique token-tag pairs).

The Brazilian tagger (`BPtag`) uses ≈ 79% for training (1,000,044 tokens and 62,762 unique token-tag pairs) and ≈ 21% (267,845 tokens and 30,848 unique token-tag pairs) for testing. As the BP corpus is much larger than the EP one, two sub-corpora were extracted from the former, in order to obtain balanced datasets for doing more tests: (i) a short version of the training (with ≈ 150,000 tokens and 16,395 unique token-tag pairs) and (ii) a reduced version for testing (≈ 23,000 tokens and 5,690 uniq token-tag pairs). Thus, these short BP datasets have a similar size than the EP ones. Every extracted sub-corpus for both EP and BP were randomly selected, and the testing datasets were never used for training.

`ALLtag` model uses for training both the EP and BP training corpora, and the PEB_Dict lexicon. `ALLtag+` was trained with the same corpora than `ALLtag`, but with the AO+_Dict.

Finally, the `PEBtag` taggers use the EP training corpus and the short version of the BP one, thus having a more balanced dataset. `PEBtag` and `PEBtag+` models also differ in the dictionary: the former uses the PEB_Dict while the latter was trained with the AO+_Dict.

The tagset (Table 1) contains 23 tags, apart from punctuation (24 tags), dates and hours (1 tag), and numerical expressions (5 tags). During the experiments, only the FreeLing PoS-tagger was used, so other modules (Recognition of Dates, Numbers, Currencies, etc.) were not applied.

For testing the performance of the PoS-taggers with different varieties, the new Web corpus was used (Section 3.1). Different experiments were carried out using the sub-corpora from Angola (AN), Mozambique (MO), Brazil (BP_AO), Portugal (EP_AO) and from the Wikipedia (Wiki).

The total micro-average of the evaluation was computed by replacing the BP testing corpus with the shorter version of the same dataset, in order to reduce bias in the results.

4.2 Results and Discussion

Table 3 contains the results of the different PoS-taggers evaluated. Here, precision is the number of correctly labeled tokens in the test set divided by the total number of tokens in the same dataset.

`BPtag` and `EPtag` models obtained 95.96% and 97.46% precision values in their respective corpora, but their results are 1.4% and 0.6% (respectively) worse when analyzing the other variety. On these (EP and BP) corpora, the performance of the `ALLtag` and `PEBtag` models depends on the distribution of the training corpora. Thus, `ALLtag` models (with more BP data) analyze better the BP corpus, while the precision of `PEBtag` models is higher when tagging EP.

When comparing both versions of `ALLtag` and `PEBtag` models with the `BPtag` and `EPtag` ones, the combined taggers achieve a better tradeoff in the annotation of BP and EP corpora.

Apart from that, the impact of the AO+_Dict lexicon is null, because BP and EP corpora do not contain texts with the AO90 spelling.

Concerning the Web corpus, `EPtag` model is still the best in every sub-corpora, except for the Wikipedia one. In this respect, it is worth noting that the annotation consistency between the EP training corpus and the EP dictionary is higher than the other varieties, and that a large part of the Web corpus follows the EP orthography. Also, remember that AN, MO and one EP_AO sub-corpora use the EP spelling system, so the results follow similar tendencies than those in the EP corpus.

In general, `PEBtag` models behave slightly better than `ALLtag` ones (except in the Wikipedia dataset), but they do not overcome the performance of the `EPtag` model.

The results in the Web corpus show that using the AO+_Dict has low (but positive) impact in the annotation. Its effect is only perceived in some texts whose spelling system had more changes due to the use of the AO90 orthography (EP_AO and Wikipedia), with small improvements (≈ 0.3) when using the larger version, which includes the AO90 entries.

Model	BP	EP	AN	MO	BP_AO	EP_AO	Wiki	Web	Total
BPtag	95.96	96.03	97.06	96.39	96.35	96.88	95.52	96.28	96.13
EPtag	95.35	**97.46**	**98.18**	**97.76**	**97.29**	**97.80**	96.25	**97.20**	**96.85**
ALLtag	**96.07**	96.94	97.30	96.91	96.68	97.18	96.50	96.83	96.64
ALLtag+	**96.07**	96.94	97.30	96.91	96.68	97.21	**96.53**	96.86	96.65
PEBtag	95.74	97.04	97.37	97.06	96.97	97.28	96.43	96.92	96.65
PEBtag+	95.74	97.04	97.37	97.06	96.97	97.31	96.45	96.93	96.66

Table 3: Precision of 6 PoS-taggers on different testing corpora. *Web* is the micro-average of the *AN*, *MO*, *BP_AO*, *EP_AO* and *Wiki* results. *Total* values are the micro-average of all the results, except for *BP*, replaced by the shorter version (see Section 4.1) in order to avoid bias.

However, even though these new dictionaries increase the ambiguity of the PoS-tagging (since they contain more token-lemma-tag triples), their influence was always positive in the tests.

In conclusion, it must be said that the consistency between the training corpus and dictionary was crucial in these experiments, with the `EPtag` models achieving the best results in almost every dataset. Apart from that, the bias between different linguistic varieties in both training and test corpora has also impact in the results. Finally, the experiments also showed that the new dictionaries have a positive influence when PoS-tagging both pre-AO90 and post-AO90 corpora in Portuguese.

5 Conclusions

Natural Language Processing tools for languages with different varieties and spelling systems —such as Portuguese–, are often built just for one of these varieties. But current NLP tasks often use a Web as Corpus approach, so there is a need of adaptation of tools for different varieties and spelling systems of the same language.

This paper has evaluated the use of several combinations of lexica and corpora for training HMM PoS-taggers aimed at analyzing different varieties of the Portuguese language.

The combinations have been focused on the analysis of Web corpora, including different text typologies (journalistic and encyclopedic), national varieties (from Portugal, Brazil, Angola and Mozambique) and spelling systems (before and after the Spelling Agreement of Portuguese: Acordo Ortográfico de 1990).

Moreover, new resources has been presented: (i) manually revised corpora for the above mentioned varieties and text typologies and (ii) two different dictionaries for Portuguese, with various combinations of European and Brazilian forms before and after the Acordo Ortográfico de 1990.

The results of the different evaluations indicate that models built with consistent training data (both corpora and lexica) achieve the highest precision.

Concerning the lexica, it has been shown that using dictionaries enriched with AO90 entries allows PoS-taggers for Portuguese to better analyze corpora from different varieties.

Finally, using a balanced training data from different varieties also helps to build a generic PoS-tagger for different linguistic varieties and text typologies.

References

Aires, Raquel V. Xavier. 2000. Implementação, adaptação, combinação e avaliação de etiquetadores para o Português do Brasil. Master's thesis, Instituto de Ciências Matemáticas, Universidade de São Paulo, São Paulo.

Almeida, Gladis Maria de Barcellos, José Pedro Ferreira, Margarita Correia, and Gilvan Müller de Oliveira. 2013. Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *ESTUDOS LINGUÍSTICOS*, 42(1):204–215.

Aluísio, Sandra M., Gisele M. Pinheiro, Marcelo Finger, M. Graças Volpe Nunes, and Stella E. Tagnin. 2003. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In *Proceedings of Corpus Linguistics*, volume 2003, pages 14–21.

Bick, Eckhard. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, University of Aarhus, Denmark.

Branco, António and João Silva. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th edition of the Language Resources and Evaluation Conference (LREC 2004)*, pages 507–510, Paris. European Language Resources Association.

Brants, Thorsten. 2000. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*. Association for Computational Linguistics.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota, and Paula Carvalho. 2003. Dicionários Electrónicos do Português. Características e Aplicações. In *Actas del VIII Simposio Internacional de Comunicación Social*, pages 636–642, Santiago de Cuba.

Garcia, Marcos and Pablo Gamallo. 2010. Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática*, 2(2):59–67.

Garcia, Marcos and Pablo Gamallo. 2014. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 3229–3233, Reykjavik. European Language Resources Association.

Marques, Nuno and Gabriel Lopes. 2001. Tagging with Small Training Corpora. In *Proceedings of the International Conference on Intelligent Data Analysis*, volume 2189 of *Lecture Notes on Artificial Intelligente (LNAI)*, pages 63–72. Springer-Verlag.

Muniz, Marcelo Caetano Martins. 2004. A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Paulo.

Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 1996)*, volume 1, pages 133–142. Association for Computational Linguistics.

Ribeiro, Ricardo, Luís C. Oliveira, and Isabel Trancoso. 2003. Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. In *Proceedings of the 6th Workshop on Computational Processing on the Portuguese Language (PROPOR 2003)*, pages 143–150, Faro. Springer-Verlag.

Tufis, Dan and Oliver Mason. 1998. Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger. In *Proceedings of the 1st edition of the Language Resources and Evaluation Conference (LREC 1998)*, volume 1, pages 589–596. European Language Resources Association.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 103-110 recibido 17-04-14 revisado 07-06-14 aceptado 10-06-14

Document-Level Machine Translation as a Re-translation Process*

Traducción Automática a Nivel de Documento como Proceso de Retraducción

Eva Martínez Garcia
Cristina España-Bonet
TALP Research Center
Univesitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
emartinez@lsi.upc.edu y cristinae@lsi.upc.edu

Lluís Màrquez
Qatar Computing Research Institute
Qatar Foundation
Tornado Tower, Floor 10,
P.O. Box 5825, Doha, Qatar
lluism@lsi.upc.edu

Resumen: Los sistemas de Traducción Automática suelen estar diseñados para traducir un texto oración por oración ignorando la información del discurso y provocando así la aparición de incoherencias en las traducciones. En este artículo se presentan varios sistemas que detectan incoherencias a nivel de documento y proponen nuevas traducciones parciales para mejorar el nivel de cohesión y coherencia global. El estudio se centra en dos casos: palabras con traducciones inconsistentes en un texto y la concordancia de género y número entre palabras. Dado que se trata de fenómenos concretos, los cambios no se ven reflejados en una evaluación automática global pero una evaluación manual muestra mejoras en las traducciones.
Palabras clave: Traducción Automática Estadística, Discurso, Coreferencia, Coherencia

Abstract: Most of the current Machine Translation systems are designed to translate a document sentence by sentence ignoring discourse information and producing incoherencies in the final translations. In this paper we present some document-level-oriented post-processes to improve translations' coherence and consistency. Incoherences are detected and new partial translations are proposed. The work focuses on studying two phenomena: words with inconsistent translations throughout a text and also, gender and number agreement among words. Since we deal with specific phenomena, an automatic evaluation does not reflect significant variations in the translations. However, improvements are observed through a manual evaluation.
Keywords: Statistical Machine Translation, Discourse, Coreference, Coherence

1 Introduction

There are many different Machine Translation (MT) systems available. Differences among systems depend on their usage, linguistic analysis or architecture, but all of them translate documents sentence by sentence. For instance, in rule-based MT systems, rules are defined at sentence level. In data-based MT systems, the translation of a document as a whole make the problem computationally unfeasible. Under this approach, the wide-range context and the discourse information (coreference relations, contextual coherence, etc.) are lost during translation.

Since this is one of the limitations for current MT systems, it is interesting to explore the possibility of improving the quality of the translations at document level. There are several phenomena that confer coherence to final translations that cannot be seen in an intra-sentence scope, for instance, some pronouns or corefered words spanning several sentences, or words that depend on a specific topic and should be translated in the same way through a document.

Following the path of some recent works (Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Xiao et al., 2011; Hardmeier, Nivre, and Tiedemann, 2012), we study some phenomena paying special attention to lexical, semantic and topic cohesion, coreference

* Supported by an FPI grant within the OpenMT2 project (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation (MICINN) and by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC).

ISSN 1135-5948

and agreement. In general, these tasks can be done following two different approaches. On the one hand, discourse information can be integrated inside a decoder, this is, trying to improve translations quality at translation time. On the other hand, the translation can be thought as a two-pass process where the characteristic phenomena are detected in a first step and re-translated afterwards.

In this work we start from a standard phrase-based Statistical Machine Translation system (SMT) from English to Spanish, and design and develop post-process architectures focusing on the phenomena just mentioned. We introduce a method to detect inconsistent translations of the same word through a document and propose possible corrections. We also present an approach to detect gender and number disagreements among coreferred words, which is extended to deal with intra-sentence disagreements.

The paper is organized as follows. We revisit briefly the related work in Section 2. Section 3 contains the description of the resources that we used to design and run the experiments explained in Section 4. Section 5 shows the results of the different translation systems together with a complete manual evaluation of the selected phenomena. Finally, we draw our conclusions and describe some lines of future work in Section 6.

2 Related Work

In the last years approaches to document-level translation have started to emerge. The earliest approaches dealt with pronominal anaphora within an SMT system (Hardmeier and Federico, 2010; Nagard and Koehn, 2010). These authors develop models that, with the help of coreference resolution methods, identify links among words in a text and use them for a better translation of pronouns. The authors in (Gong, Zhang, and Zhou, 2011) approach the problem of topic cohesion by making available the previous translations at decoding time by using a cache system. In this way, one can bias easily the system towards the lexicon already used.

Document-level translation can be also seen as the post-process of an already translated document. In (Xiao et al., 2011), the authors study the translation consistency of a document and re-translate source words that have been translated in different ways within a same document. The aim is to incorporate document contexts into an existing SMT system following 3 steps. First, they identify the ambiguous words; then, they obtain a set of consistent translations for each word according to the distribution of the word over the target document; and finally, generate the new translation taking into account the results of the first two steps.

All of these works are devoted to improve the translation in one particular aspect (anaphora, lexicon, ambiguities) but do not report relevant improvements as measured by an automatic metric, BLEU (Papineni et al., 2002).

Recently, the authors in (Hardmeier, Nivre, and Tiedemann, 2012) presented Docent, an SMT document-level decoder. The decoder is built on top of an open-source phrase-based SMT decoder, Moses (Koehn et al., 2007). The authors present a stochastic local search decoding method for phrase-based SMT systems which allows decoding complete documents. Docent starts from an initial state (translation) given by Moses and this one is improved by the application of a hill climbing strategy to find a (local) maximum of the score function. The score function and some defined change operations are the ones encoding the document level information. The Docent decoder is introduced in (Hardmeier et al., 2013).

3 Experimental Setup

In order to evaluate the performance of a system that deals with document-level phenomena, one needs to consider an adequate setting where the involved phenomena appear.

3.1 Corpora and systems

Most of the parallel corpora used to train SMT systems consist of a collection of parallel sentences without any information about the document structure. An exception is the News Commentary corpus given within the context of the workshops on Statistical Machine Translation[1]. The corpus is build up with news, that is, coherent texts with a consistent topic throughout a document. Besides, one can take advantage from the XML tags of the documents that identify the limits of the document, paragraphs and sentences.

This corpus is still not large enough to train an SMT system, so, for our baseline system we used the Europarl corpus (Koehn, 2005) in its version 7. All the experiments

[1] http://www.statmt.org/wmt14/translation-task.html

are carried out over translations from English to Spanish.The different morphology between these two languages should contribute to obtain troublesome translations which can be tackled with our methodology.

Our baseline system is a Moses decoder trained with the Europarl corpus. For estimating the language model we use SRILM (Stolcke, 2002) and calculate a 5-gram language model using interpolated Kneser-Ney discounting on the target side of the Europarl corpus. Word alignment is done with GIZA++ (Och and Ney, 2003) and both phrase extraction and decoding are done with the Moses package. The optimization of the weights of the model is trained with MERT (Och, 2003) against the BLEU measure on the News Commentary corpus of 2009 (NC-2009, see Table 1 for the concrete figures of the data).

3.2 Data annotation

Besides the aforementioned markup, the 110 documents in the test set have been annotated with several linguistic processors. In particular, we used the Part-of-Speech (PoS) tagger and dependency parser provided by the Freeling library (Padró et al., 2010), the coreference resolutor RelaxCor (Sapena, Padró, and Turmo, 2010) and the named entity recognizer of BIOS (Surdeanu, Turmo, and Comelles, 2005). Whereas the PoS has been annotated in both the source (English) and the target (Spanish) sides of the test set, named entities, dependency trees and coreferences have been annotated in the source side and projected into the target via the translation alignments (see Section 5).

4 Systems Description

4.1 Document-level phenomena

4.1.1 Lexical coherence

The first characteristic we want to improve is the lexical coherence of the translation. Ambiguous words are source words with more than one possible translation with different meanings. Choosing the right translation is,in this case, equivalent to disambiguate the word in its context. Taking the assumption of "one-sense-per-discourse", the translation of a word must be the same through a document. So, our problem is not related to Word Sense Disambiguation, but we want to identify the words in the source document that are translated in different ways in the target. For example, the English word "desk" appearing in the first News can be translated as "ventanilla", "escritorio", "mesa" or "mostrador" according to the translation table of our baseline system, where the different options are not exact synonyms. The aim of our system is to translate "desk' as "mesa" homogeneously throughout the document as explained in Section 4.2.1. This is an example from the 488 instances of words whit inconsistent translations that we found in our corpus using our baseline system.

4.1.2 Coreference and agreement

It is easy to find words that corefer in a text. A word corefers with another if both refer to the same entity. These words must in principle agree in gender and number since they are representing the same concept (person, object, etc.). For instance, if the term "the engineer" appears referring to a girl as it is the case in News 5, the correct translation in Spanish would be "la ingeniera" and not "el ingeniero".

Once we identify and try to fix incoherences in gender or number inside coreference chains, we can take advantage of the analysis and the applied strategies in the coreference problem to correct agreement errors in the intra-sentential scope. This is, in fact, a simpler problem because it is not affected by possible errors given by the coreference resolutor. However, since dependencies among words are shorter, the expressions tend to be translated correctly by standard SMT engines. In our corpus, we only found two instances where the agreement processing could be applied to coreference chains, so most of our analysis finally corresponds to the intra-sentence agreement case.

4.2 Re-translation systems

Given these phenomena we design two main post-processes to detect, re-translate and fix the interesting cases. Figure 1 shows the basic schema of the post-processes.

4.2.1 Lexical coherence

The first step of the post-process that deals with lexical coherence is to identify those words from the source document translated in more than one way. We use the PoS tags to filter out only the nouns, adjectives and main verbs in English. Then, the word alignments given by a first-pass Moses' translation are used to link every candidate token to its translation, so those tokens aligned with more than one different form in the target

	Corpus	News	Sentences	English Tokens	Spanish Tokens
Training	Europarl-v7	–	1,965,734	49,093,806	51,575,748
Development	NC-2009	136	2,525	65,595	68,089
Test	NC-2011	110	3,003	65,829	69,889

Table 1: Figures on the corpora used for training, development and test.

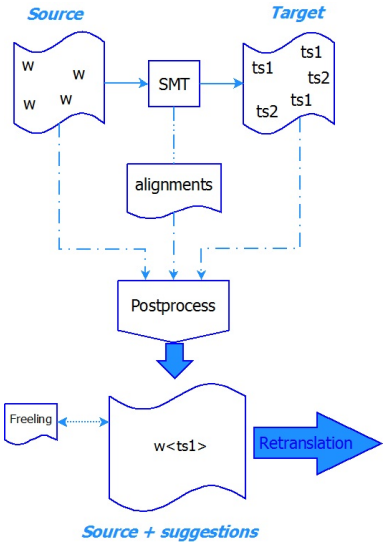


Figure 1: Structure of the post-processes that rewrite the source document in order to prepare it to a retranslation step.

side can be identified. Following the example of Section 4.1.1, if the word "desk" appears three times in the text and it is translated two times as "mesa" and one as "mostrador", then the pair (desk, desk) (mesa, mostrador) will be selected for re-translation.

Re-translation is done in two different ways: *restrictive* and *probabilistic*. The *restrictive* way forces[2] as a possible translation the most used option in the current document; in case of tie, there is no suggestion in order to avoid biasing the result in a wrong way. By doing this, we somehow control the noise introduced by the post-process but we also lose information given by the decoder in the available translation. On the other hand, the *probabilistic* way suggests the most used options as possible translations assigning them a probability estimated by frequency counts within the options. So, in this case, one feeds the decoder with the most suitable options and let it choose among them. This option introduces more noise because the system is managing more possible translations than in the previous situation, sometimes as many as in the initial state translation.

[2] *Forcing* or *suggesting* a translation is a feature available in the Moses decoder that involves an XML markup of the source sentence with the information of translation options.

4.2.2 Gender and number agreement

The post-process for disagreements in gender and number analyses every source document and annotates it with PoS, coreference chains, and dependency trees. The main structure in this case is the tree since it is the one that allows to link the elements that need to agree. A tree traversal is performed in order to detect nouns in the source. When a noun is found and its children are determiners and/or adjectives, the matching subtree is projected into the target via the SMT word alignments. In the target side, one can check the agreement among tokens by using the PoS tags. If there is a disagreement, the correct tag for the adjective or determiner is built using the corresponding Freeling library, which allows to get the correct form in the target language for the translation.

The system implements a similar strategy to check the agreement among subject and verb. A tree traversal allows to detect the node that represents the verb of the sentence and the child corresponding to the subject. The structure is projected into the target via the alignments and the agreement is verified using the PoS information. If the subject is a noun, we assume that the verb must be conjugated in third person plural or singular depending on the number of the noun; if it is a pronoun, gender, person and number must agree. As before, if there is a disagreement, the system assigns the correct tag to the verb, and the form is generated using the Freeling library.

In both cases (determiner–adjective(s)–noun(s) and subject–verb) the output of the pre-process is a proposed new translation for translations that show a disagreement. Similarly to the *restrictive* and *probabilistic* systems of the previous subsections, here we run the re-translation step in two ways: forcing an output with new translation or allowing the interaction of this new translation

System	BLEU	NIST	TER	METEOR	ROUGE	SP-Op	ULC
Baseline	26.73	7.34	55.45	27.78	29.36	31.53	85.01
Lex_R	26.76	7.34	55.39	27.80	29.39	31.60	83.26
Lex_P	26.73	7.34	55.41	27.77	29.38	31.58	85.07
Agr_R	26.66	7.33	55.46	27.75	29.41	31.69	85.10
Agr_P	26.73	7.33	55.45	27.75	29.41	31.64	85.05
Seq_R_R	26.65	7.32	55.46	27.74	29.40	31.68	85.08
Seq_R_P	26.73	7.33	55.45	27.75	29.40	31.63	85.05
Seq_P_R	26.64	7.32	55.48	27.74	29.38	31.67	79.28
Seq_P_P	26.72	7.32	55.46	27.74	29.40	31.63	85.04

Table 2: Automatic evaluation of the systems. See text for the system and metrics definition.

with the remaining translation options of the phrase table. In both cases, the full sentence can be re-translated to accommodate the new options.

The following section shows an automatic and manual evaluation of these systems for the English–Spanish language pair.

5 Experimental Results

The most straightforward way to evaluate translation engines is using automatic metrics on the full test sets. However, in our case, the measures are not informative enough considering that we apply small modifications to previous translations. As an example, Table 2 shows the automatic evaluation obtained with the Asiya toolkit (González, Giménez, and Màrquez, 2012) for several lexical metrics (BLEU, NIST, TER, METEOR and ROUGE), a syntactic metric based on the overlap of PoS elements (SP-Op), and an average of a set of 27 lexical and syntactic metrics (ULC).

The first row shows the results for the baseline system built with Moses without any re-translation step (*Baseline*). The second block includes the experiments on lexical coherence alone both restrictive and probabilistic (*Lex_R* and *Lex_P*) and the third block the experiments on agreement alone also in the two cases (*Agr_R* and *Agr_P*). Finally, the last block shows the result of the sequential process with the four combination of systems (*Seq_R_R*, *Seq_R_P*, *Seq_P_R*, *Seq_P_P*). As it can be seen, the scores do not show any systematic preference for a system and it is necessary a manual evaluation of the outputs to study the performance of the re-translation systems.

System	BLEU	tags	words	OK/ch	linTags	linDif
News20bl	13.40					
News20_R	13.56	26	8	5/9	13	6
News20_P	13.22	45	15	7/11	19	8
News25bl	14.42					
News25_R	14.45	18	4	4/4	16	3
News25_P	14.52	38	10	5/5	28	7
News39bl	28.49					
News39_R	28.20	16	5	5/5	15	4
News39_P	28.56	34	11	6/8	25	7
News48bl	30.05					
News48_R	30.06	42	3	3/3	23	10
News48_P	29.83	53	7	4/5	24	15
News49bl	25.54					
News49_R	25.87	24	5	5/5	17	8
News49_P	25.83	42	12	7/8	23	10

Table 3: Manual evaluation of the system for lexical coherence (*Lex* in Table 2) for a subset of news with restrictive and probabilistic systems. See text for column's meaning.

5.1 Manual evaluation

In order to manually evaluate the output of the previous systems, we chose those documents where we include more suggestions into the re-translation step.

In Table 3, one can see the results of evaluating the system devoted to improve the global lexical coherence for the five news where the post-process introduces more changes. For every document, the *News*bl* row represents the scores for the translations obtained using the baseline system. Column *tags* shows the number of introduced tags, *words* the number of different words involved in the tags, *OK/ch* shows the number of changes made with respect to the first translation and how many are correct attending to our criteria of having one-sense-per-discourse and the word appearing in the reference translation. Note that the tags in the probabilistic approach (_P) include the ones of the restrictive approach (_R) since the first strategy allows us to suggest possible new translations of a word in more cases,

not only when we find the strictly most used translation option for a word. In order to see the scope of the introduced changes, *linTags* shows the number of tagged lines in the source text and *linDif* shows the number of different lines between the final translation and the translation the system uses at the beginning. In general, in all our experiments we could see very local changes due to the retranslation step that affected mostly the tagged words without changing the main structure of the target sentence nor the surrounding ones.

We observe that as with the full automatic evaluation, the BLEU scores of our experiments differ in a non-significant way from the baseline and this is because we are introducing only a few changes in a document. For instance, when we re-translate News20, the one that makes the largest number of changes, we only change 9 words using the *restrictive* approach and 11 using the *probabilistic* one. In this concrete document the accuracy of our changes is above the 50%, but in general, the restrictive approach obtains a high performance and, in the rest of the documents evaluated (News25, News39, News48 and News49), the accuracy in the changes is of a 100%. The probabilistic approach shows a slightly lower performance with accuracies around 80%.

A clear example of how our system works can be observed in a document that talks about a judgement. The document contains the phrase "the trial coverage" translated in first place as "la cobertura de prueba" where the baseline system is translating wrongly the word "trial". But, our post-process sees this word translated more times through the document as "juicio", identifies it as an ambiguous word and tags it with the good translation form "juicio". But not all the changes are positive as we have seen. For example, in a document appears the word "building" five times, being translated three times as "construcción" and two times as "edificio". For our system, the first option is better as long as it appears more times in the translation than the second one. So, we suggested the decoder to use "construcción" when translates "building" in the document. Doing that, we produce two changes in the final translation that generate two errors with respect to the reference translation although both translation options are synonyms. So, in this case our system moves away the translation from the reference although both translations should be correct.

system	BLEU	OK/ch	dets	adjs	verbs
News5bl	13.74				
News5_R	14.06	23/26	17/19	6/7	0/0
News5_P	13.79	15/26	12/19	3/7	0/0
News6bl	11.06				
News6_R	11.22	19/23	8/11	11/11	0/1
News6_P	11.10	10/23	4/11	6/11	0/1
News22bl	16.23				
News22_R	14.74	17/25	4/8	13/17	0/0
News22_P	14.89	10/25	2/8	8/17	0/0
News27bl	13.15				
News27_R	12.35	22/28	14/19	7/8	1/1
News27_P	12.76	21/28	14/19	7/8	0/1
News33bl	15.09				
News33_R	16.05	18/22	14/16	3/3	1/3
News33_P	15.97	11/22	7/16	2/3	2/3

Table 4: Manual evaluation of the system that deals with the agreement (*Agr* in Table 2) for a subset of news with restrictive and probabilistic systems. See text for column's meaning.

Regarding to the errors introduced by the systems, we find that they are caused mainly by bad alignments which provoke an erroneous projection of the structures annotated on the source, errors in the PoS tagging, untranslated words, or, finally, a consequence of the fact that the most frequent translation for a given word in the initial state is wrong.

If we move on now to the agreement experiment, we observe the results from the manual evaluation of checking number and gender agreement in Table 4. Column *OK/ch* shows the number of introduced changes (correct/total), the *dets* column shows the changes over determiners, *adjs* over adjectives and *verbs* over verb forms.

In this set of experiments, we observe that the changes induced by our post-process have an impact in the BLEU score of the final translation because in this case the number of changes is higher. For instance, in News22, we observe a drop of almost two points in the BLEU score after applying the post-process although many of the changes made after the re-translation are correct. We observe the same behaviour in News27, although in the rest of news is shown an opposite trend. According to the manual evaluation, the restrictive system is better than the probabilistic one and reaches accuracies above 80% in the selected news.

A positive example of the performance of the system is the re-translation of the source

system	BLEU	OK/ch	dets	adjs	verbs
News20bl	13.40				
News20_R_R	13.38	17/19	14/15	3/3	0/1
News20_R_P	13.44	14/19	11/15	2/3	1/1
News20_P_R	13.21	16/17	13/14	3/3	0/0
News20_P_P	13.44	12/17	10/14	2/3	0/0
News25bl	14.42				
News25_R_R	14.68	12/19	9/13	3/6	0/0
News25_R_P	15.09	15/19	10/13	5/6	0/0
News25_P_R	14.39	10/17	6/11	4/6	0/0
News25_P_P	14.82	13/17	8/11	5/6	0/0
News39bl	28.49				
News39_R_R	30.02	20/22	14/16	6/6	0/0
News39_R_P	29.59	18/22	13/16	5/6	0/0
News39_P_R	29.94	19/21	14/16	5/5	0/0
News39_P_P	29.59	17/21	13/16	4/5	0/0
News48bl	30.05				
News48_R_R	29.57	6/6	5/5	1/1	0/0
News48_R_P	29.60	4/6	4/5	0/1	0/0
News48_P_R	29.57	6/6	5/5	1/1	0/0
News48_P_P	29.60	4/6	4/5	0/1	0/0
News49bl	25.54				
News49_R_R	25.82	9/11	3/4	6/7	0/0
News49_R_P	26.02	9/11	3/4	6/7	0/0
News49_P_R	25.63	8/11	3/4	5/6	0/1
News49_P_P	26.02	9/11	3/4	5/6	1/1

Table 5: Manual evaluation of the translation after combining sequentially both post-processes, first applying the disambiguation post-process and, afterwards, checking for the agreement. The notation is the same as in previous tables.

phrase "the amicable meetings". This phrase is translated by the baseline as "el amistosa reuniones", where one can find disagreements of gender and number among the determiner, the adjective and the noun. The system detects these disagreements and after tagging the source with the correct forms and re-translating, one obtains the correct final translation "las reuniones amistosas", where we observe also that the decoder has reordered the sentence.

Regarding to the errors introduced by the system, we observe again that many of them are caused by wrong analysis. For instance, in the sentence "all (the) war cries" which should be translated as "todos los gritos de guerra", the dependence tree shows that the determiner depends on the noun "war" and not on "cries", so, according to this relation, our method identifies that the determiner and the translation do not agree and produces the wrong translation "todos (la) guerra gritos".

These results also show that for our approach it is easier to detect and fix disagreements among determiners or adjectives and nouns than among subjects and their related verbs. In general, this is because our current system does not take into account subordinated sentences, agent subjects and other complex grammatical structures, and therefore the number of detected cases is smaller than for the determiner–adjective–noun cases. Further work can be done here to extend this post-process in order to identify disagreements among noun phrases and other structures in the sentence that appear after the verb.

In order to complete this set of experiments, we run sequentially both systems. Table 5 shows the results for the combination of systems in the same format as in the previous experiment. Once again, we observe only slight variations in BLEU scores but, manually, we see that when the systems introduce changes, they are able to fix more translations than the ones they damage. Also as before, it is easier to detect and fix disagreements among determiners, adjectives and nouns than those regarding verbs because of the same reason as in the independent system.

6 Conclusions and Future Work

This work presents a methodology to include document-level information within a translation system. The method performs a two-pass translation. In the first one, incorrect translations according to predefined criteria are detected and new translations are suggested. The re-translation step uses this information to promote the correct translations in the final output.

A common post-process is applied to deal with lexical coherence at document level and intra- and inter-sentence agreement. The source documents are annotated with linguistic processors and the interesting structures are projected on the translation where inconsistencies can be uncovered. In order to handle lexical coherence, we developed a post-process that identifies words translated with different meanings through the same document. For treating disagreements, we developed a post-process that looks for inconsistencies in gender, number and person within the structures determiner–adjective(s)–noun(s) and subject–verb.

Because we are treating concrete phenomena, an automatic evaluation of our systems does not give us enough information to assess the performance of the systems. A detailed manual evaluation of both systems

shows that we only introduce local changes. The lexical-coherence-oriented post-process induces mostly correct translation's changes when using our restrictive system, improving the final coherence of the translation. On the other hand, for the post-process focused on the analysis of the number and gender agreement, it achieves more than 80% of accuracy over the introduced changes in the manually-evaluated news documents. We also observed that some of the negative changes are consequence of bad word alignments which introduce noise when proposing new translations.

A natural continuation of this work is to complete the post-processes by including in the study new document-level phenomena like discourse markers or translation of pronouns. On the other hand, we aim to refine the methods of suggestion of new possible translations and to detect bad word alignments. As a future work, we plan to introduce the analysis of these kind of document-level phenomena at translation time, using a document-level oriented decoder like Docent.

References

Gong, Z., M. Zhang, and G. Zhou. 2011. Cache-based document-level statistical machine translation. In *Proc. of the 2011 Conference on Empirical Methods in NLP*, pages 909–919, UK.

González, M., J. Giménez, and L. Màrquez. 2012. A graphical interface for MT evaluation and error analysis. In *Proc. of the 50th ACL Conference, System Demonstrations*, pages 139–144, Korea.

Hardmeier, C. and M. Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proc. of the 7th International Workshop on Spoken Language Translation*, pages 283–289, France.

Hardmeier, C., J. Nivre, and J. Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proc. of the Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 1179–1190, Korea.

Hardmeier, C., S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proc. of the 51st ACL Conference*, pages 193–198, Bulgaria.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proc.: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th ACL Conference*, pages 177–180, Czech Republic.

Nagard, R. Le and P. Koehn. 2010. Aiding pronouns translation with co-reference resolution. In *Proc. of Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Sweden.

Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the ACL Conference*.

Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Padró, L., S. Reese, E. Agirre, and A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, India. Global Wordnet Conference.

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL Conference*, pages 311–318.

Sapena, E., L. Padró, and J. Turmo. 2010. A global relaxation labeling approach to coreference resolution. In *Proceedings of 23rd COLING*, China.

Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*.

Surdeanu, M., J. Turmo, and E. Comelles. 2005. Named entity recognition from spontaneous open-domain speech. In *Proc. of the 9th Interspeech*.

Xiao, T., J. Zhu, S. Yao, and H. Zhang. 2011. Document-level consistency verification in machine translation. In *Proc. of Machine Translation Summit XIII*, pages 131–138, China.

Extracción de Terminología y Léxicos de Opinión

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 113-120 recibido 11-04-14 revisado 30-06-14 aceptado 02-07-14

ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas

ML-SentiCon: *a multilingual, lemma-level sentiment lexicon*

Fermín L. Cruz, José A. Troyano, Beatriz Pontes, F. Javier Ortega
Universidad de Sevilla
Escuela Técnica Superior de Ingeniería Informática, Av. Reina Mercedes s/n
{fcruz,troyano,bepontes,javierortega}@us.es

Resumen: En este trabajo presentamos un conjunto de lexicones de polaridades semánticas a nivel de lemas para inglés, español, catalán, gallego y euskera. Estos lexicones están estructurados en capas, lo que permite seleccionar distintos compromisos entre la cantidad de estimaciones de positividad y negatividad y la precisión de dichas estimaciones. Los lexicones se han generado automáticamente a partir de una mejora del método utilizado para generar SentiWordNet, un recurso ampliamente utilizado que recoge estimaciones de positividad y negatividad a nivel de synsets. Nuestras evaluaciones sobre los lexicones para inglés y español muestran altos niveles de precisión en todas las capas. El recurso que contiene todos los lexicones obtenidos, llamado ML-SentiCon, está disponible de forma pública para su uso.
Palabras clave: Análisis del sentimiento, recursos léxicos, recursos multilingüe

Abstract: In this work, we present a set of lemma-level sentiment lexicons for English, Spanish, Catalan, Basque and Galician. These lexicons are layered, allowing to trade off between the amount of available words and the accuracy of the estimations. The lexicons have been automatically generated from an improved version of SentiWordNet, a very popular resource which contains estimations of the positivity and negativity of synsets. Our evaluations on the English and Spanish lexicons show high accuracies in all cases. The resource containing all the lexicons, ML-SentiCon, is publicly available.
Keywords: Sentiment analysis, lexical resources, multilingual resources

1 Introducción

El Análisis del Sentimiento es una disciplina enmarcada en el Procesamiento del Lenguaje Natural que se ocupa del tratamiento computacional de fenómenos como la subjetividad, las emociones y las opiniones en textos; una buena introducción al estado de la cuestión se encuentra en(Liu y Zhang, 2012). Se trata de un área de investigación muy activa en los últimos años, debido en gran parte a que las opiniones expresadas en Internet por los usuarios constituyen una fuente de información de incalculable utilidad para las Administraciones Públicas, las grandes y medianas compañías y los consumidores. Los contenidos de esta naturaleza generados por usuarios son de un volumen y velocidad de aparición tales que imposibilitan su análisis exclusivamente manual y requieren de la aplicación de métodos automáticos de apoyo. Algunas de las tareas definidas en el área son la detección de textos subjetivos, la clasificación de textos basada en la polaridad de las opiniones expresadas (positiva/negativa), o la extracción de opiniones individuales y sus participantes. Muchas de las soluciones propuestas a estas tareas se apoyan en lexicones de opiniones o de polaridad (*opinion lexicons* o *sentiment lexicons* en inglés), recursos léxicos que contienen información sobre las implicaciones emocionales de las palabras. Generalmente, esta información consiste en la polaridad a priori de las palabras, es decir, las connotaciones positivas o negativas de dichas palabras en ausencia de contexto.

En este trabajo presentamos nuevos lexicones para inglés, español, catalán, gallego y euskera. Los lexicones están organizados en varias capas, lo que permite a las aplicaciones que los utilicen seleccionar distintos compromisos entre la cantidad de palabras disponibles y la precisión de las estimaciones de

ISSN 1135-5948

sus polaridades a priori. Para generar estos lexicones, como paso previo, hemos reproducido el método utilizado en (Baccianella, Esuli, y Sebastiani, 2010) para construir SentiWordNet 3.0, un recurso léxico construido sobre WordNet y ampliamente utilizado en el área del Análisis del Sentimiento. Hemos incorporado diversas mejoras al método original, que repercutieron positivamente en la calidad del recurso obtenido, según nuestras evaluaciones.

El resto del artículo se estructura como sigue. En la sección 2 repasamos algunos ejemplos representativos de trabajos sobre lexicones de opinión, tanto en inglés como en otros idiomas, centrándonos especialmente en el español. En las secciones 3 y 4 describimos el proceso de construcción de nuestro recurso; dicho proceso está dividido en dos partes fundamentales, descritas y evaluadas cada una por separado. Por último, en la sección 5 resumimos el contenido anterior y señalamos algunas conclusiones y líneas de trabajo futuro.

2 Trabajos relacionados

Existen muchos trabajos que abordan la creación de lexicones de opinión, con enfoques muy distintos. *General Inquirer* (Stone, Dunphy, y Smith, 1966) puede ser considerado, entre otras cosas, el primer lexicón que contiene información afectiva. Se trata de un lexicón construido manualmente, formado por lemas, es decir, unidades semánticas que pueden aparecer en múltiples formas lexicalizadas (por ejemplo, el verbo *approve* es un lema que puede aparecer en los textos con diferentes flexiones, como *approved* o *approving*). *General Inquirer* contiene una gran cantidad de información asociada a cada lema, de tipo sintáctico, semántico y pragmático. Entre toda esta información, existen un total de 4206 lemas que están etiquetados como positivos o negativos. A pesar de su antigüedad, *General Inquirer* es aún usado en muchos trabajos de análisis del sentimiento.

El *MPQA Subjectivity Lexicon* (Wilson, Wiebe, y Hoffmann, 2005) es un ejemplo de lo anterior. Se trata de un lexicón que aglutina palabras subjetivas obtenidas mediante un método automático (Riloff y Wiebe, 2003), términos obtenidos a partir de un diccionario y un *tesaurus*, y las listas de palabras positivas y negativas de *General Inquirer*. En total contiene 8221 palabras, cuyas polaridades (positiva, negativa o neutra) fueron anotadas manualmente. La lista final contiene 7631 entradas positivas y negativas, y es muy heterogénea, conteniendo palabras flexionadas y otras lematizadas. Al igual que *General Inquirer*, la lista no contiene términos formados por más de una palabra.

Los dos lexicones de opinión más utilizados en la actualidad, atendiendo al número de citas, son *Bing Liu's Opinion Lexicon* (Hu y Liu, 2004; Liu, Hu, y Cheng, 2005) y SentiWordNet (Esuli y Sebastiani, 2006; Baccianella, Esuli, y Sebastiani, 2010). Ambos presentan enfoques muy distintos y en cierto sentido opuestos. Mientras que el *Bing Liu's Opinion Lexicon* está formado por unas 6800 palabras flexionadas, incluyendo incluso faltas de ortografía y *slangs* (expresiones informales utilizadas frecuentemente en Internet), SentiWordNet está construido sobre WordNet (Fellbaum, 1998), un recurso léxico en el que las unidades básicas, llamadas *synsets*, aglutinan a distintas palabras que comparten un mismo significado. El *Bing Liu's Opinion Lexicon* está construido a partir de un método automático, pero las listas de palabras positivas y negativas han sido actualizadas manualmente de forma periódica, hasta la versión actualmente disponible en la web, que data de 2011. Por su parte, SentiWordNet asigna valores reales de positividad y negatividad, entre 0 y 1, a cada uno de los más de 100 mil *synsets* presentes en WordNet. Los valores han sido calculados mediante un método automático, partiendo de unos conjuntos semilla de *synsets* positivos y negativos.

Es importante señalar la diferencia entre los lexicones a nivel de palabras o lemas, como *General Inquirer*, el *MPQA Subjectivity Lexicon* o el *Bing Liu's Opinion Lexicon*, y los lexicones a nivel de *synsets*, como SentiWordNet. Los primeros están formados por términos con ambigüedad semántica, debido a la polisemia de muchas de las palabras. Sin embargo, los lexicones basados en *synsets* no presentan este problema, puesto que las unidades que los forman representan unívocamente un sólo significado. Por contra, para poder emplear un lexicón basado en *synsets* es necesario aplicar a los textos que se van a analizar una herramienta de desambigüación de significados, las cuales tienen una precisión aún relativamente baja hoy por hoy. La mayoría de los trabajos que utilizan Senti-

WordNet optan por calcular valores agregados de polaridad a nivel de palabras o lemas, a partir de los valores de todos los posibles *synsets* correspondientes (Taboada et al., 2011; Denecke, 2008; Martín-Valdivia et al., 2012; Agrawal y others, 2009; Saggion y Funk, 2010; Kang, Yoo, y Han, 2012; Desmet y Hoste, 2013). En este trabajo hemos construido lexicones de ambos tipos: a nivel de *synsets*, introduciendo algunas mejoras al método empleado para la construcción de SentiWordNet 3.0, y a nivel de lemas, a partir de los valores obtenidos en el lexicón anterior.

2.1 Lexicones de opiniones en otros idiomas

Aunque aún son pocos los lexicones de opinión disponibles para idiomas distintos al inglés, poco a poco van creciendo en número. Existen trabajos que se centran en la creación de lexicones para idiomas tan diversos como el hindú y el francés (Rao y Ravichandran, 2009), el árabe (Abdul-Mageed, Diab, y Korayem, 2011), el alemán (Clematide y Klenner, 2010), el japonés (Kaji y Kitsuregawa, 2007), el chino (Lu et al., 2010; Xu, Meng, y Wang, 2010), el rumano(Banea, Mihalcea, y Wiebe, 2008) y el español. Creemos que en el caso del español son pocos aún los lexicones disponibles actualmente, teniendo en cuenta que el español es el segundo idioma más hablado en el mundo y el tercero más utilizado en Internet.

En (Brooke, Tofiloski, y Taboada, 2009) se utilizan dos recursos (un diccionario bilingüe[1] y Google Translator[2]) para obtener, de manera automática y a partir de un lexicón en inglés, dos lexicones en español. El trabajo no muestra resultados de evaluación para los lexicones obtenidos, sino que presenta los resultados de una herramienta de clasificación basada en la polaridad que se apoya en los lexicones. Una técnica similar es empleada en (Molina-González et al., 2013), donde se aplica traducción automática de inglés a español al Bing Liu's Opinion Lexicon. Algunas de las ambigüedades inherentes a la traducción fueron resueltas manualmente. También se añadieron manualmente traducciones de algunas palabras informales y *slangs* contenidos en el lexicón de partida. Tampoco en este trabajo se reportan resultados de evaluación directa del lexicón obtenido, sino que se presentan resultados de evaluación extrínseca basados en clasificación binaria del sentimiento. Pérez-Rosas, Banea, y Mihalcea (2012) parten de dos lexicones en inglés: el MPQA Subjectivity Lexicon (el cual es mapeado sobre WordNet mediante un método automático) y SentiWordNet. Después utilizan una versión en español de WordNet que forma parte del proyecto EuroWordNet (Vossen, 1998) para obtener listas de términos positivos y negativos en español. El trabajo plantea dos variaciones del método, a partir de las que se obtienen dos lexicones en español, uno formado por 1347 términos y otro por 2496. Las evaluaciones realizadas mediante la anotación manual de una muestra de 100 términos de cada uno de los lexicones arrojan un 90 % y un 74 % de *accuracy* respectivamente.

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

[2] http://translate.google.com

3 Estimación de la polaridad a nivel de synsets

Nuestro objetivo es la obtención de lexicones a nivel de lemas en varios idiomas. Como paso previo a la obtención de dichos lexicones, hemos construido un lexicón en inglés a nivel de *synsets*. Para la creación de este lexicón, partimos del método utilizado para la construcción de SentiWordNet 3.0, introduciendo mejoras en cada una de las etapas del método. Una vez obtenido este recurso, generamos lexicones a nivel de lemas para inglés, español, catalán, gallego y euskera. Las evaluaciones llevadas a cabo muestran mejoras significativas tanto en el lexicón a nivel de *synsets* (comparándonos con SentiWordNet 3.0), como en el lexicón a nivel de lemas, con valores de precisión y volumen para el lexicón en español superiores a los de (Pérez-Rosas, Banea, y Mihalcea, 2012).

Basándonos en el método empleado por (Baccianella, Esuli, y Sebastiani, 2010) para la generación de SentiWordNet 3.0 e incorporando diversas modificaciones, hemos calculado valores reales entre 0 y 1 de positividad, negatividad y objetividad para cada uno de los *synsets* de WordNet 3.0. Al igual que el método en el que nos basamos, nuestro método se divide en dos partes claramente diferenciadas: un primer cálculo individual de la polaridad, y un segundo cálculo global de la polaridad a partir de los valores obtenidos en la primera etapa.

3.1 Cálculo individual de polaridad

El cálculo individual de la polaridad se basa en la construcción de clasificadores ternarios, capaces de decidir si un *synset* es positivo, negativo o neutro a partir de los textos de sus glosas (las glosas son definiciones contenidas en WordNet para cada uno de los *synsets*). Para entrenar estos clasificadores, se parte de distintos conjuntos de *synsets* considerados a priori positivos, negativos o neutros. En (Baccianella, Esuli, y Sebastiani, 2010) se utilizaron los *synsets* correspondientes a palabras positivas y negativas usadas por Turney y Littman (2003)[3] En nuestro caso, hemos utilizado también WordNet-Affect(Strapparava, Valitutti, y Stock, 2006) como fuente de semillas positivas y negativas. Los clasificadores entrenados a partir de las distintas fuentes de información, y usando dos algoritmos de clasificación distintos (Rocchio y SVM), fueron combinados en una etapa de meta-aprendizaje, obteniéndose finalmente tres clasificadores regresionales capaces de inducir valores de positividad, negatividad y objetividad en el intervalo $[0, 1]$. Los detalles de esta etapa pueden consultarse en (Cruz et al., 2014).

3.2 Cálculo global de polaridad

El cálculo global de la polaridad trata de refinar en su conjunto los valores de positividad y negatividad asignados a cada *synset*, a partir de distintos tipos de relaciones entre ellos. Estas relaciones se modelan mediante un grafo en el que los *synsets* son representados mediante nodos y las aristas dirigidas indican algún tipo de relación entre los valores de positividad y negatividad de dichos *synsets*. Se asigna a cada nodo un valor numérico (por ejemplo, los valores de positividad), y se aplica entonces al grafo un algoritmo de paseo aleatorio. Estos algoritmos son capaces de computar iterativamente las interacciones que se producen entre los valores asignados a los nodos: los valores inicialmente asignados a los nodos "fluyen" a lo largo del grafo a través de las aristas. Una vez ha convergido, el algoritmo obtiene valores finales para los nodos, que dependen tanto de los valores inicialmente asignados a los mismos como de las relaciones existentes entre los nodos a nivel global.

Las diferencias fundamentales de nuestra propuesta con respecto a SentiWordNet 3.0 en este paso son dos. En primer lugar, construimos dos tipos de grafos distintos, uno a partir de las glosas y otro a partir de las relaciones semánticas de WordNet (en SentiWordNet se emplea únicamente un grafo basado en las glosas). En ambos casos, los grafos incluyen aristas con peso positivo, que representan una transferencia directa entre los valores de positividad y negatividad de los *synsets* conectados, y aristas con peso negativo, que indican una transferencia cruzada entre ambos tipos de valores (en SentiWordNet sólo se contemplan aristas sin pesos). En segundo lugar, aplicamos PolarityRank(Cruz et al., 2012), un algoritmo de paseo aleatorio sobre grafos que permite computar los valores finales de positividad y negatividad en una sola ejecución, existiendo además una interdependencia entre los valores finales positivos y negativos (en SentiWordNet se llevaban a cabo dos ejecuciones independientes del algoritmo PageRank, una para los valores de positividad y otra para los de negatividad). Los detalles de esta etapa también están explicados más ampliamente en (Cruz et al., 2014).

3.3 Evaluación

En la tabla 1 se muestran los valores de la distancia τ_p de *Kendall* (Fagin et al., 2004) entre un *gold standard* y los valores obtenidos con nuestro método. Esta medida estima la similitud entre un *ranking* modelo o *gold standard* y otro *ranking* candidato. Cuanto más cercano a cero, más parecidos son ambos *rankings*. Hemos usado el mismo *gold standard* usado en (Baccianella, Esuli, y Sebastiani, 2010), por lo que los resultados son comparables con los mostrados en dicho trabajo. Como puede apreciarse, hemos conseguido mejoras significativas en ambas etapas, con estimaciones finales de positividad y negatividad usando nuestro método más precisas que las de SentiWordNet 3.0 (se reduce τ_p un 24,2 % y un 7,4 %, respectivamente).

4 *Inducción de* ML-SentiCon

Para facilitar el uso del recurso por parte de aquellos investigadores que no deseen utilizar desambigüación de significados, hemos generado un lexicón a nivel de lemas partiendo del

[3]Positivas: good, nice, excellent, positive, fortunate, correct, superior. Negativas: bad, nasty, poor, negative, unfortunate, wrong, inferior.

Etapa	Recurso	Positiv.	Negativ.
1	SWN	0,339	0,286
	ML-SC	0,238	0,284
2	SWN	0,281	0,231
	ML-SC	**0,213**	**0,214**

Tabla 1: Valores de τ_p de SentiWordNet (SWN) y ML-SentiCon (ML-SC) obtenidos en cada etapa del método de cálculo de valores de positividad y negatividad de *synsets* (1: Cálculo individual; 2: Cálculo global).

lexicón a nivel de *synsets* anterior. Además, usando recursos que nos permiten conectar los synsets con lemas en otros idiomas, hemos generado versiones del lexicón en español, catalán, gallego y euskera. El recurso finalmente obtenido, llamado ML-SentiCon, está disponible de forma pública para su uso [4].

Cada lexicón a nivel de lemas está formado por ocho capas. Las capas están ordenadas, desde la primera hasta la octava, de manera que las capas posteriores contienen todos los lemas de las anteriores, y añaden algunos nuevos. Los lemas que conforman cada una de las capas son obtenidos rebajando progresivamente una serie de restricciones, de manera que el número de lemas que las cumplen va aumentando capa tras capa, a costa de una bajada paulatina en la fiabilidad de dichos lemas como indicadores de positividad y negatividad.

4.1 Definición de las capas

Cada *synset* s_i en WordNet tiene asociado un conjunto de lemas $L_i = \{l_1, l_2, \ldots, l_n\}$ (también llamados *variants*), todos con la misma categoría morfosintáctica (nombre, adjetivo, verbo o adverbio). Además, cada *synset* s_i tiene un valor de positividad p_i y un valor de negatividad n_i en el recurso obtenido anteriormente. Diremos que la polaridad del *synset* s_i es $pol_i = p_i - n_i$. De cara a la definición del lexicón a nivel de lemas, consideremos que un *synset* s_i es la tupla formada por el conjunto de lemas y la polaridad, es decir, $s_i = (L_i, pol_i)$. Invirtiendo esta asociación, a cada lema l le corresponde un conjunto de *synsets* $S_l = \{s_i : l \in L_i\}$. Denotamos con $\overline{pol}_l$ a la media de las polaridades pol_i de los *synsets* $s_i \in S_l$. Cada una de las ocho capas está formada por un conjunto de lemas positivos l tales que $\overline{pol}_l > 0$, y otro conjunto de lemas negativos l tales que $\overline{pol}_l < 0$.

[4] http://www.lsi.us.es/~fermin/index.php/Datasets

Capa	en	sp	ca	eu	gl
1	157	353	512	138	49
2	982	642	530	278	223
3	1600	891	699	329	370
4	2258	1138	860	404	534
5	3595	1779	1247	538	883
6	6177	2849	1878	888	1429
7	13517	6625	4075	2171	2778
8	25690	11918	7377	4323	4930

Tabla 2: Distribución de lemas por capas en los lexicones en inglés (en), español (es), catalán (ca), euskera (eu) y gallego (gl).

Las dos primeras capas están formadas exclusivamente por lemas $l \in L_i$ de *synsets* s_i que pertenecen a alguno de los conjuntos de entrenamiento usados en la etapa de cálculo individual de la polaridad. El resto de capas están formadas por lemas $l \in L_i$ de cualquier *synset* s_i; en cada capa se exigen valores mínimos diferentes sobre el valor absoluto de $\overline{pol}_l$ de los lemas l que las conforman. Estos valores mínimos se han escogido tratando de obtener una progresión geométrica en el número de lemas que componen cada una de las capas.

4.2 Obtención de lemas en otros idiomas

Para obtener correspondencias entre los *synsets* y lemas en otros idiomas distintos al inglés, hemos utilizado el *Multilingual Central Repository 3.0* (MCR 3.0) (Gonzalez-Agirre, Laparra, y Rigau, 2012). Este recurso se compone de WordNets incompletos para cuatro idiomas: español, catalán, euskera y gallego. Los *synsets* de estos WordNets están conectados con los de WordNet 3.0, lo que nos permite replicar el mismo procedimiento de construcción de las capas explicado en la sección anterior. Para el caso del español y el catalán, hemos utilizado también la información generada por el proyecto *EuroWordNet* (Vossen, 1998) a fecha de noviembre de 2006, lo que nos permite aumentar el número de lemas para estos idiomas. *EuroWordNet* se basa en WordNet 1.6, por lo que hemos tenido que realizar un mapeo a WordNet 3.0 mediante *WN-Map*[5] (Daudé, Padró, y Rigau, 2003) . En la tabla 2 se muestra la distribución de lemas por capas e idiomas de los lexicones obtenidos.

[5] http://nlp.lsi.upc.edu/tools/download-map.php

4.3 Evaluación

Para evaluar la calidad de los lexicones a nivel de lemas, hemos revisado manualmente las listas de lemas positivos y negativos de cada una de las capas, etiquetando cada entrada como correcta o incorrecta. Hemos evaluado de esta forma los lexicones en inglés y español. Para los cuatro primeros niveles (niveles 1-4), hemos revisado las listas completas[6]. Para el resto de niveles (niveles 5-8), hemos revisado una muestra aleatoria estadísticamente representativa de cada uno de los niveles. Hemos calculado el tamaño de la muestra que garantiza un error muestral menor a ±5 % en la estimación de la proporción de elementos correctos, suponiendo una distribución binomial de la variable aleatoria con $p = q = 0{,}5$ (el peor caso posible), y con un intervalo de confianza del 95 % ($\alpha = 0{,}05$). Con estos parámetros, hemos obtenido tamaños muestrales de entre 300 y 400 elementos, según el nivel.

En la tabla 3 se muestra el *accuracy* estimado (porcentaje de elementos correctos frente al total) para cada capa de los lexicones en inglés y español, respectivamente. Los resultados confirman una gran fiabilidad de las listas de lemas positivos y negativos generadas, con valores por encima del 90 % en las capas 1-6 del lexicón en inglés y las capas 1-5 del lexicón en español.

Como puede observarse, el accuracy del lexicón en inglés es mayor al del lexicón en español, lo cual es lógico puesto que el lexicón en español se ha construido a partir de recursos generados mediante métodos semiautomáticos y por tanto no carentes de errores. La diferencia de *accuracy* entre ambos lexicones va en aumento a lo largo de las capas, pasando de 1,63 puntos porcentuales en la primera capa a 12,27 puntos en la última capa. A pesar de esto, creemos que los valores medidos para el lexicón en español son buenos, si comparamos las capas 5 y 6 (91,75 % en un lexicón de 1779 y 86,09 % en un lexicón de 2849) con los lexicones en español de (Pérez-Rosas, Banea, y Mihalcea, 2012) (90 % en un lexicón de 1347 lemas y 74 % en uno de 2496 lemas). Más aún, la siguiente capa de nuestro recurso continúa teniendo un mejor nivel de acierto (77,69 %) que el mayor de los lexicones de (Pérez-Rosas, Banea, y Mihalcea, 2012), con un número de lemas muy superior (6625).

	Inglés		Español	
Capa	**Acc.**	**Tam.**	**Acc.**	**Tam.**
1	99,36 %	157	97,73 %	353
2	98,88 %	982	97,20 %	642
3	97,75 %	1600	94,95 %	891
4	96,24 %	2258	93,06 %	1138
5	93,95 %	3595	91,75 %	1779
6	91,99 %	6177	86,09 %	2849
7	85,29 %	13517	77,69 %	6625
8	74,06 %	25690	61,29 %	11918

Tabla 3: Estimación muestral del porcentaje de lemas con polaridad correcta (Acc.) y número de lemas total (Tam.) de cada una de las capas de los lexicones en inglés y español.

5 Conclusiones

En este trabajo presentamos un nuevo recurso, llamado ML-SentiCon, formado por lexicones a nivel de lemas en inglés, español, catalán, gallego y euskera. Los lexicones están formados por capas, lo que permite seleccionar distintos compromisos entre la cantidad de términos disponibles y la precisión de las estimaciones de sus polaridades a priori. Para cada lema, el recurso proporciona un valor real que representa dicha polaridad, entre -1 y 1, y un valor de desviación estándar que refleja la ambigüedad resultante del cómputo de la polaridad a partir de los valores de los distintos significados posibles del lema. El recurso está disponible de forma pública para su uso[7].

Como paso previo en la obtención de estos lexicones, hemos presentado también una versión mejorada del método usado para construir SentiWordNet 3.0(Baccianella, Esuli, y Sebastiani, 2010). Hemos llevado a cabo evaluaciones de nuestro método similares a las realizadas en el artículo anterior, cuyos resultados reflejan mejoras significativas en las estimaciones de polaridad obtenidas con nuestro método con respecto al método original. Los detalles de cada una de las mejoras implementadas se pueden consultar en (Cruz et al., 2014).

Creemos que el recurso obtenido puede ser útil en multitud de aplicaciones relacionadas con el análisis del sentimiento, tanto para inglés como para español. Aunque es

[6] En el recurso hecho público, hemos incluido las versiones libres de lemas erróneos de las cuatro primeras capas.

[7] http://www.lsi.us.es/~fermin/index.php/Datasets

presumible que las conclusiones en cuanto al porcentaje de acierto en la polaridad de los lemas del lexicón en español sean extrapolables al resto de idiomas incluidos (puesto que se han utilizado recursos y métodos equivalentes), sería deseable que otros investigadores más familiarizados con estos idiomas que los autores del presente trabajo estimaran la calidad de dichos lexicones. El método aquí propuesto puede ser reproducido para otros idiomas, siempre que existan WordNets disponibles. En este sentido, puede ser útil el recurso *Open Multilingual WordNet*(Bond y Foster, 2013), que reúne WordNets para multitud de idiomas procedentes de distintos proyectos internacionales.

Agradecimientos

Este trabajo ha sido financiado a través de los proyectos ATTOS (TIN2012-38536-C03-02) y DOCUS (TIN2011-14726-E) concedidos por el Ministerio de Ciencia e Innovación del Gobierno de España, y del proyecto AORESCU (P11-TIC-7684 MO) concedido por la Consejería de Innovación, Ciencia y Empresas de la Junta de Andalucía.

Bibliografía

Abdul-Mageed, Muhammad, Mona T Diab, y Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. En *ACL (Short Papers)*, páginas 587–591.

Agrawal, Shaishav y others. 2009. Using syntactic and contextual information for sentiment polarity analysis. En *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, páginas 620–623. ACM.

Baccianella, Stefano, Andrea Esuli, y Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh conference on International Language Resources and Evaluation*. ELRA, may.

Banea, Carmen, Rada Mihalcea, y Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. En *LREC*.

Bond, Francis y Ryan Foster. 2013. Linking and extending an open multilingual wordnet. En *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*.

Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*, páginas 50–54.

Clematide, Simon y Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for german. En *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, página 7.

Cruz, Fermín L, José A Troyano, Beatriz Pontes, y F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Cruz, Fermín L., Carlos G. Vallejo, Fernando Enríquez, y José A. Troyano. 2012. Polarityrank: Finding an equilibrium between followers and contraries in a network. *Inf. Process. Manage.*, 48(2):271–282.

Daudé, Jordi, Lluis Padró, y German Rigau. 2003. Making wordnet mapping robust. *Procesamiento del lenguaje natural*, 31.

Denecke, Kerstin. 2008. Using sentiwordnet for multilingual sentiment analysis. En *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, páginas 507–512. IEEE.

Desmet, Bart y Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*.

Esuli, Andrea y Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. En *Proceedings of Language Resources and Evaluation (LREC)*.

Fagin, Ronald, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, y Erik Vee. 2004. Comparing and aggregating rankings with ties. En *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, páginas 47–58, New York, NY, USA. ACM.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gonzalez-Agirre, Aitor, Egoitz Laparra, y German Rigau. 2012. Multilingual central repository version 3.0. En *LREC*, páginas 2525–2529.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *KDD '04: Proceedings of the tenth ACM SIGKDD*, páginas 168–177, New York, NY, USA. ACM.

Kaji, Nobuhiro y Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. En *EMNLP-CoNLL*, páginas 1075–1083.

Kang, Hanhoon, Seong Joon Yoo, y Dongil Han. 2012. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010.

Liu, Bing, Minqing Hu, y Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. En *Proceedings of WWW*.

Liu, Bing y Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. En Charu C. Aggarwal y ChengXiang Zhai, editores, *Mining Text Data*. Springer US, páginas 415–463.

Lu, Bin, Yan Song, Xing Zhang, y Benjamin K Tsou. 2010. Learning chinese polarity lexicons by integration of graph models and morphological features. En *Information retrieval technology*. Springer, páginas 466–477.

Martín-Valdivia, María-Teresa, Eugenio Martínez-Cámara, Jose-M Perea-Ortega, y L Alfonso Ureña-López. 2012. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*.

Molina-González, M Dolores, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, y José M Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Pérez-Rosas, Verónica, Carmen Banea, y Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. En *LREC*, páginas 3077–3081.

Rao, Delip y Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 675–682. Association for Computational Linguistics.

Riloff, Ellen y Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. En *Proceedings of EMNLP*.

Saggion, Horacio y A Funk. 2010. Interpreting sentiwordnet for opinion classification. En *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*.

Stone, Philip J, Dexter C Dunphy, y Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Strapparava, Carlo, Alessandro Valitutti, y Oliviero Stock. 2006. The affective weight of lexicon. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, páginas 423–426.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, y Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Turney, Peter D. y Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston.

Wilson, Theresa, Janyce Wiebe, y Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of the HLT/EMNLP*, páginas 347–354.

Xu, Ge, Xinfan Meng, y Houfeng Wang. 2010. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. En *Proceedings of the 23rd International Conference on Computational Linguistics*, páginas 1209–1217. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 121-128 recibido 15-04-14 revisado 01-07-14 aceptado 02-07-14

Unsupervised acquisition of domain aspect terms for Aspect Based Opinion Mining

Adquisición no supervisada de aspectos de un dominio para Minería de Opiniones Basada en Aspectos

Aitor García-Pablos, Montse Cuadros, Seán Gaines
Vicomtech-IK4 research centre
Mikeletegi 57, San Sebastian, Spain
{agarciap,mcuadros,sgaines}@vicomtech.org

German Rigau
IXA Group
Euskal Herriko Unibertsitatea
San Sebastian, Spain
german.rigau@ehu.es

Resumen: El análisis automático de la opinión, que usualmente recibe el nombre minería de opinión o análisis del sentimiento, ha cobrado una gran importancia durante la última década. La minería de opinión basada en aspectos se centra en detectar el sentimiento con respecto a "aspectos" de la entidad examinada (i.e. características o partes concretas evaluadas en una sentencia). De cara a detectar dichos aspectos se requiere una cierta información sobre el dominio o temática del contenido analizado, ya que el vocabulario varía de un dominio a otro. El objetivo de este trabajo es generar de manera automática una lista de aspectos del dominio partiendo de un set de textos sin etiquetar, de manera completamente no supervisada, como primer paso para el desarrollo de un sistema más completo.
Palabras clave: aspectos de dominio, adaptación a dominio, minería de opinión

Abstract: The automatic analysis of opinions, which usually receives the name of opinion mining or sentiment analysis, has gained a great importance during the last decade. This is mainly due to the overgrown of online content in the Internet. The so-called aspect based opinion mining systems aim to detect the sentiment at "aspect" level (i.e. the precise feature being opinionated in a clause or sentence). In order to detect such aspects it is required some knowledge about the domain under analysis. The vocabulary in different domains may vary, and different words are interesting features in different domains. We aim to generate a list of domain related words and expressions from unlabeled domain texts, in a completely unsupervised way, as a first step to a more complex opinion mining system.
Keywords: aspect based sentiment analysis, unsupervised lexicon generation

1 Introduction

Opinion mining and sentiment analysis has attracted the attention of the research community during the last decade (Pang and Lee, 2008; Liu, 2012; Zhang and Liu, 2014). Specially during the last years, when the opinionated content flows thanks to the so called Web 2.0. Review web sites, blogs and social networks, are producing everyday a massive amount of new content, much of it bearing opinions about different entities, products or services. Trying to cope with this data is infeasible without the help of automatic Opinion Mining tools which try to detect, identify, classify, aggregate and summarize the opinions expressed about different topics. The opinion mining systems can be roughly classified into two types, supervised, and unsupervised or semi-supervised since some level of supervision is almost always required to guide or initialize most of the existing systems. Supervised systems require training data, which usually includes manually annotated data, in order to train a model that can "learn" how to label new unseen data. These systems perform quite well, but it is difficult to port to different domains or languages due to the cost of obtaining such manually annotated data. Unsupervised methods (or semi-supervised) try to leverage the vast amount of unlabeled data (i.e. all the content that is constantly generated over the Internet) to infer the required information without the need of big amounts of hand-crafted resources. These systems have the clear advantage of being much more portable to other languages or domains. In this work we will briefly introduce the concept of "aspect based

ISSN 1135-5948

opinion mining" and some of the existing approaches in the literature. Then we will introduce the Semeval 2014 task 4, which is about detecting opinionated aspect targets and their categories and polarities in customrer review sentences. After that we will explain our approach to generate a list of aspect terms for a new domain using a collection of unlabeled domain texts. Finally we show our results after evaluating the approach against Semeval 2104 task 4 datasets, and our conclusions and future work.

2 Related Work

Customer reviews are full of fine grained opinions and sentiments towards different aspects, features or parts of a product or service. In order to discover which aspects are being praised and which are being criticized a fine grained analysis is required. Many approaches have been carried out.

Hu and Liu (2004) try to summarize customer reviews in a aspect level basis. They employ frequent nouns and phrases as potential aspects, and use relations between aspects and opinions to identify infrequent aspects. Popescu and Etzioni (2005) extract high frequent noun phrases in reviews as candidate product aspects. Then, they compute the Pointwise Mutual Information (PMI) score between the candidates and some meronymy discriminators associated with the product class to evaluate each candidate.

Zhuang, Jing, and Zhu (2006) employ certain dependency relations to extract aspect-opinion pairs from movie reviews. They first identify reliable dependency relation templates from training data to identify valid aspect-opinion pairs in test data.Wu et al. (2009) use dependency parsing to extract noun phrases and verb phrases as aspect candidates. Blair-Goldensohn (2008) refine the approach proposed in Hu and Liu (2004) considering only noun phrases inside sentiment-bearing sentences or in some syntactic patternz indicating sentiment, plus some additional filters to remove unlikely aspects.

Qiu et al. (2009) propose a double propagation method to bootstrap new aspect terms and opinion words from a list of seeds using dependency rules. The process is called double propagation because they use opinion words to obtain new aspect terms and aspect terms to obtain new opinion words. The acquired opinion words and aspect terms are added to the seed lists, and used to obtain more words in a new loop. The process stops when no more words can be acquired. In Zhang et al. (2010) the double propagation approach is extended to aspect ranking to deal with the noise that double propagation method tends to generate. The authors model the aspect terms and opinion words as a bipartite graph and use HITS algorithm to rank the aspect terms, also using some linguistics patterns (e.g. part-whole relation patterns).

In this work we reuse some of these ideas to build an unsupervised system that bootstrap a ranked list of domain aspect terms just by using a set of unlabeled domain texts (customer reviews of a particular topic). We evaluate our results against the SemEval 2014 task 4 datasets.

3 SemEval 2014 Task 4

SemEval 2014 task 4[1] *Aspect Based Sentiment Analysis* (Pontiki et al., 2014) provides two training datasets, one of restaurant reviews and other of laptop reviews. The restaurant review dataset consists of over 3,000 English sentences from restaurant reviews borrowed from Ganu, Elhadad, and Marian (2009). The laptop review dataset consist of over 3,000 English sentences extracted from customer reviews. The task is divided in four different subtasks. Subtask 1 is *aspect term extraction*: given a set of sentences referring to pre-identified entities (i.e. restaurants or laptops), return the list of distinct aspect terms present in the sentence. An aspect term names a particular aspect of the target entity (e.g. *menu* or *wine* for restaurants, *hard disk* or *battery life* for laptops). Subtask 2 focuses on detecting the polarity of a given set of aspect terms in a sentence. The polarity in this task can be one of the following: *positive*, *negative*, *neutral* or *conflict*. The objective of subtask 3 is to classify the identified aspect terms into a predefined set of categories. The categories can be seen as a more coarse grained aspects that include the aspect terms. In this SemEval task the predefined set of categories for restaurants are: *food*, *service*, *price*, *ambiance* and *anecdotes/miscellaneous*. No categories have been provided for the laptop do-

[1] http://alt.qcri.org/semeval2014/task4/

main. Subtask 4 is analogous to the subtask 2, but in this case the polarity has to be determined for the aspect categories. Again, only the restaurant dataset is suitable for this task since the laptop dataset does not contain aspect category annotations.

In this paper we focus our attention on subtask 1, *aspect term extraction*. Our aim is to develop an unsupervised system able to extract aspect terms from any domain and evaluate it against the SemEval datasets, using the evaluation tools and metrics provided by the tasks organizers.

4 Our approach

Our aim is to build a system that is capable of generating a list of potential aspect terms for a new domain without any kind of adaptation or tuning. Such a list can be a useful resource to exploit in a more complex system aiming to perform Aspect Based Sentiment Analysis. Aspect terms, also known as *opinion targets* in the literature, generally refer to parts of features of a given entity. For example, *wine list* and *menu* could be aspect terms in a text reviewing a restaurant, and *hard disk* and *battery life* could be aspect terms in a laptop review. Obviously, each domain has its own set of aspect terms, referring to different aspects, parts and features of the entities described in that domain. The only requirement to generate the list of aspect terms for a new domain is a, preferably large, set of unlabelled documents or review describing entities of the domain. Our method combines some techniques already described in the literature with some modifications and additions.

4.1 Used data

Using a web-scraping program we have extracted a few thousand English reviews from a restaurant review website[2], and a similar amount of English reviews from a laptop review website[3]. We have not performed any kind of sampling or preprocessing on the extracted data, it has been extracted "as-is" from the list of entities (restaurants and laptops) available in the respective websites at the time of the scraping. The extracted reviews have been split in sentences unsing Stanford NLP tools and stored into an XML file. A subset of 25,000 sentences have been used to acquire the aspect term lists, combined with the already mentioned 3,000 sentences of the Semeval 2014 task 4 datasets.

4.2 Double propagation

We have adapted the double-propagation technique described in Qiu et al. (2009) and Qiu et al. (2011). This method consists of using an initial seed list of aspect terms and opinion words and propagate them through a dataset using a set of propagation rules. The goal is to expand both the aspect term and opinion word sets. Qiu et al. (2009) define opinion words as words that convey some positive or negative sentiment polarities. They only use nouns as aspect terms, and only adjectives can be opinion words. This is an important restriction that limits the recall of the process, but the double-propagation process is intended to extract only explicit aspects (i.e. aspects that are explicitly mentioned in the text, and not aspects implicitly derived from the context). The detection of implicit aspects (e.g. "The phone fits in the pocket" referring to the size) requires a different set of techniques and approaches that are described in many works in the literature Fei et al. (2012; Hai, Chang, and Cong (2012).

During the propagation process a set of propagation rules are applied to discover new terms (aspect terms or opinion words), and the initial aspect term and opinion word sets are expanded with each new discovery. The newly discovered words are also used to trigger the propagation rules, so in each loop of the process additional words can be discovered. The process ends when no more words can be extracted. Because aspect terms are employed to discover new opinion words, and opinion words are employed to discover new aspect terms, the method receives the name of double-propagation.

The propagation is guided by some propagation rules. When the conditions of a rule are matched, the target word (aspect term or opinion word) is added to its correspondent set.

4.3 Propagation rules

The propagation rules are based on dependency relations and some part-of-speech restrictions. We have mainly followed the same rules detailed in Qiu et al. (2011) with some

[2]Restaurant reviews of different cities from `http://www.citysearch.com`

[3]Laptop reviews from `http://www.toshibadirect.com`

minor modifications. The exact applied rules used this work can be observed in the Table 1.

Some rules extract new aspect terms, and others extract new opinion words. In Table 1, T means *aspect term* (i.e. a word already in the aspect terms set) and O means *opinion word* (i.e. a word already in the opinion words set). W means *any word*. The dependency types used are *amod*, *dobj*, *subj* and *conj*, which stand for *adjectival modifier*, *direct object*, *subject* and *conjunction* respectively. Additional restrictions on the Part-Of-Speech (POS) of the words present in the rule, it is shown in the third column of the table. The last column indicates to which set (aspect terms or opinion words) the new word is added.

To obtain the dependency trees and word lemmas and POS tags, we use the Stanford NLP tools[4]. Our initial seed words are just *good* and *bad*, which are added to the initial opinion words set. The initial aspect terms set starts empty. This way the initial sets are not domain dependent, and we expect that, if the propagation rules are good enough, the propagation should obtain the same results after some extra iterations.

Each sentence in the dataset is analyzed to obtain its dependency tree. Then the rules are checked. If a word and its dependency-related words trigger the rule, and the conditions hold, then the word indicated by the rule is added to the corresponding set (aspect terms or opinion words, depending on the rule). The process continues sentence by sentence adding words to both sets. When the process finishes processing sentences, if new words have been added to any of the two sets, the process starts again from the first sentence with the enriched sets. The process stops when no more words have been added during a full dataset loop.

5 *Ranking the aspect terms*

Although the double-propagation process populates both sets of domain aspect terms and domain opinion words, we focus our attention in the aspect terms set. Depending on the size and content of the employed dataset, the number of potential aspect terms will be quite large. In our case the process generates many thousands of different potential aspect terms. Much of them are incorrect, or very unusual aspect terms (e.g. in the restaurant domain, a cooking recipe written in another language, a typo, etc.). Thus, the aspect terms needs to be ranked, trying to keep the most important aspects on top, and pushing the less important ones to the long tail.

In order to rank the obtained aspect terms, we have modeled the double-propagation process as a graph population process. Each new aspect term or opinion word discovered by applying a propagation rule is added as a vertex to the graph. The rule used to extract the new word is added as an edge to the graph, connecting the originating word and the discovered word.

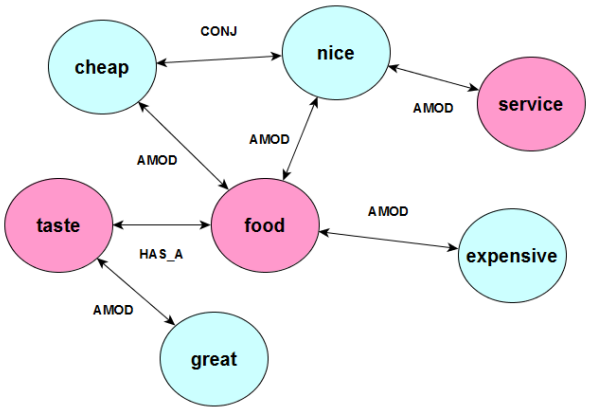


Figure 1: Example of a graph fragment constructed with the bootstrapped words and relations.

Figure 1 presents as an example a small part of the graph obtained by the double-propagation process. Each vertex representing a word maintains the count of how many times that word has appeared in the dataset, and also if it is an aspect term or an opinion word. A word is identified by its lemma and its POS tag. Every edge in the graph also maintains a count of how many times the same rule has been used to connect a pair of words. At the end of the double-propagation process the generated graph contains some useful information: the frequency of appearance of each word in the dataset, the frequency of each propagation rule, the number of different words related to a given word, etc. We have applied the well-known PageRank algorithm on the graph to score the vertices. To calculate the PageRank scores we have

[4] `http://nlp.stanford.edu/software/lex-parser.shtml`

Rule	Observations	Constraints	Action
R11	O → amod→W	W is a noun	W→T
R12	O→dobj→W1 ←subj←W2	W2 is a noun	W2→T
R21	T ← amod ← W	W is an adjective	W→O
R22	T → subj → W1 ← dobj ← W2	W2 is an adjective	W2→ O
R31	T → conj → W	W is a noun	W → T
R32	T → subj → has *gets* dobj ← W	W is a noun	W → T
R41	O → conj → W	W is an adjective	W→ O
R42	O → Dep1 → W1 ← Dep2 ← W2	Dep1==Dep2, W2 is an adjective	W2→ O

Table 1: Propagation rules

Restaurants	Laptops
1- food	1-battery life
2- service	2- keyboard
3- staff	3- screen
4- bar	4- feature
5- drink	5- price
6- table	6- machine
7- menu	7- toshiba laptop
8- dish	8- windows
9- atmosphere	9- performance
10- pizza	10- use
11- meal	11- battery
12- bartender	12- program
13- price	13- speaker
14- server	14- key
15- dinner	15- hard drive

Table 2: Top ranked aspect terms for restaurant and laptop domain using our approach

used the JUNG framework[5], a set of Java libraries to work with graphs. The value of the alpha parameter that represents the probability of a random jump to any node of the graph has been left at 0.15 (in the literature it is recommended an alpha value between 0.1 and 0.2).

The graph is treated as an undirected graph because the propagation rules represented by the graph edges can be interpreted in both directions (e.g. A modifies to B, or B is modified by A). The aspect terms are then ordered using their associated score, being the most relevant aspect term the one with the highest score.

5.1 Filtering undesired words

The double-propagation method always introduces many undesired words. Some of these undesired words appear very frequently and are combined with a large number of words. So, they tend to also appear in high positions in the ranking.

Many of these words are easy to identify, and they are not likely to be useful aspect terms in any domain. Examples of these words are: *nothing*, *everything*, *thing*, *anyone*, *someone*, *somebody*, etc. They are extracted during the double-propagation process because they appear in common expressions like *It was a good thing*, *It is nothing special*, *I like everything*. The process also extract other words, like *year*, *month*, *night*, and other time expressions. Also, some common words, like *boy*, *girl*, *husband* or *wife*. The reason for this is that the input texts are customers reviews, and it is quite common to find anecdotes and personal comments like *I saw a nice girl in the bar*. It would be interesting to find an automatic method to safely remove all these words, valid for many domains. A TF-IDF weigthing of the words may be a useful preprocessing to identify noisy content. For this work we have chosen the simple approach of adding them to a customizable stop word list. The final list contains about one hundred words that are not likely to be aspect terms in any domain. The list has been crafted observing the most common unwanted words after running the system, and using intuition and common sense. Our purpose was not to tune the stop word list to work better with any of our evaluation domains, and the same stop word list has been used in the evaluation in both restaurant and laptop domains.

6 Dealing with multiword terms

Many aspect terms are not just a single word, but compounds and multiword terms. For some domains this is more critical than for others. As it can be observed in Table 2, the top ranked aspect term for laptops is *battery life*. The laptop domain is a very challenging domain due to the amount of technical vocabulary that usually combine several words (e.g. *hard disk drive*, *Intel i7 processor*, etc.). In

[5] `http://jung.sourceforge.net`

order to improve the precision and the recall of the generated set of aspect terms, multiword aspect terms must be detected and included in the resulting sets. We have tried different approaches, trying increase the recall without adding incorrect terms.

6.1 Using WordNet

One of the approaches included in the system exploits WordNet [6], and some simple rules. Each time a word is going to be processed during the double-propagation algorithm, the combination of the current word plus the next word is checked. If some conditions are satisfied then we treat both words as a single multiword term. The conditions are the following:

- If word n and word n+1 are nouns, and the combination is an entry in WordNet (or in Wikipedia, see below). E.g.: *battery life*
- If word n is an adjective and word n+1 is a noun, and the combination is an entry in WordNet. E.g.: *hot dog*, *happy hour*
- If word n is an adjective, word n+1 is a noun, and word n is a relational adjective in WordNet (lexical file 01). E.g.: *Thai food*, *Italian food*

6.2 Using Wikipedia

In order to improve the coverage of the WordNet approach, we also check if a combination of two consecutive nouns appears as a Wikipedia article title. Wikipedia articles refer to real word concepts and entities, so if a combination of words is a title of a Wikipedia article it is very likely that this word combination is also meaningful for the domain under analysis (e.g. *DVD player*, *USB port*, *goat cheese*, *pepperoni pizza*). However, since Wikipedia contains many entries that are titles of films, books, songs, etc., that would lead to the inclusion of erroneous multiword expressions, for example *good time*. For this reason we limit the lookup in Wikipedia titles just to combination of nouns, avoiding combinations of adjective + noun. This gives a good balance between extended coverage and inclusion of incorrect aspect terms.

[6] `http://wordnet.princeton.edu/`

SemEval Restaur.	Precision	Recall	F-score
SemEval Baseline	0.539	0.514	0.526
Our system (S)	0.576	0.649	0.610
Our system (W)	0.555	0.661	0.603
Our system (W+S)	0.551	0.662	0.601
SemEvaml-Best	0.853	0.827	0.840

Table 4: Result comparison for SemEval restaurant review dataset

6.3 Using simple patterns

In this work we have limited the length of the multiword terms to just bi-grams. But in some cases it is interesting to have word combinations of a bigger size. For that purpose we have included some configurable patterns to treat longer chains of words as a single aspect term. The patterns are very simple, being expressed with a simple syntax like *A of N*. It means that a known aspect term (represented by the uppercased A) followed by the word *of*, followed by a noun (represented by the uppercased N) must be processed as a single aspect term. Similar patterns would be *N of A*, *A with N*, *N with A*, etc. These patterns are useful to extract expressions like *chicken with onion*, or *glass of wine*.

7 Evaluation

To evaluate the quality of the resulting aspect term lists, we have used our method to annotate the SemEval 2014 datasets of task 4, *Aspect Based Sentiment Analysis* which provides two datasets, one for "restaurants" domain and another for "laptops" domain. An example of the format can be seen in the Figure 3. The datasets are composed by individual sentences. Each sentence contains annotated data about the aspect terms present in that sentence. The aspect terms are the span of characters inside the sentence that holds the mention to the aspect.

The SemEval task provides an evaluation script which evaluates standard precision, recall and F-score measures. Both datasets (restaurants and laptops) contain 3,000 sentences each. The restaurant dataset contains 3,693 labeled gold aspect term spans (1,212 different aspect terms), and the laptop dataset contains 2,358 labeled gold aspect term spans (955 different aspect terms). We use these gold aspect terms to evaluate the experiments.

The experiment using our approach consists of using the generated aspect term lists

```
<sentence id="270">
       <text>From the incredible food, to the warm atmosphere, to the
       friendly service, this downtown neighborhood spot doesn't miss a beat.
       </text>
       <aspectTerms>
           <aspectTerm term="food" polarity="positive" from="20" to="24"/>
           <aspectTerm term="atmosphere" polarity="positive" from="38" to="48"/>
           <aspectTerm term="service" polarity="positive" from="66" to="73"/>
       </aspectTerms>
</sentence>
```

Table 3: Example of SemEval 2014 Task 4 dataset sentence

SemEval Laptops	Precision	Recall	F-score
SemEval Baseline	0.401	0.381	0.391
Our system (S)	0.309	0.475	0.374
Our system (W)	0.327	0.508	0.398
Our system (W+S)	0.307	0.533	0.389
SemEval-Best	0.847	0.665	0.745

Table 5: Result comparison for SemEval laptop review dataset

(for restaurants and laptops) to annotate the sentences. The generated aspect term lists have been limited to the top 550 items. In this particular experiment, we have observed than using longer lists increases the recall, but decreases the precision due to the inclusion of more incorrect aspects terms. The annotation process is a simple lemma matching between the words in the dataset and the words in our generated lists.

We compare the results against the SemEval baseline which is also calculated by some scripts provided by the Semeval organizers. This baseline splits the dataset into *train* and *test* subsets, and uses all the labeled aspect terms in the *train* subset to build a dictionary of aspect terms. Then it simply uses that dictionary to label the test subset for evaluation. We also show the result of the best system submitted to SemEval (SemEval-Best in the table) for each domain. However the results are not comparable since our approach is unsupervised and just a first step to a more complex system and does not use any machine learning or other supervised techniques to annotate the data.

Tables 4 and 5 show the performance of our system with respect to the baselines in both datasets. "Our system (S)" stands for our system only using the SemEval provided data (as it is unsupervised it learns from the available texts for the task). (W) refers to the results when using our own dataset scraped from the Web. Finally (W+S) refers to the results using both SemEval and our Web dataset. On the restaurant dataset our system outperforms the baseline and it obtains quite similar results on the laptop dataset. Interestingly, the results are quite similar even if the learning datasets are very different in size. Probably this is because it only leverages more documents if they include new words that can be bootstrapped. If the overall distribution of words and relations does not change, the resulting aspect term list would be ranked very similarly.

Apart from the non-recognized aspect terms (i.e. not present in the generated list) another important source of errors is the multiword aspect term detection. In the SemEval training dataset, about the 25% of the gold aspect terms are multiword terms. In the restaurant dataset we find a large number of names of recipes and meals, composed by two, three or even more words. For example *hanger steak au poivre* or *thin crusted pizza* are labeled as single aspect terms. In the laptop domain multiword terms are also very important, due to the amount of technical expressions (i.e. hardware components like "RAM memory", software versions like "Windows 7" and product brands like "Samsumg screen"). These aspect terms cannot be present in our automatically acquired aspect term list because we limit the multiword length up to two words.

There are also errors coming from typos and variations in the word spelling (e.g. *ambience* and *ambiance*) that our system does not handle.

8 *Conclusions and future work*

Aspect term extraction (also known as *features* or *opinion targets*) is an important first step to perform fine grained automatic opinion mining. There are many approaches in

the literature aiming to automatically generate aspect terms for different domains. In this paper we propose a simple and unsupervised method to bootstrap and rank a list of domain aspect terms using a set of unlabeled domain texts. We use a double-propagation approach, and we model the obtained terms and their relations as a graph. Then we employ PageRank algorithm to score the obtained terms. We evaluate the approach in the SemEval 2014 Task 4 and our unsupervised system performs better than the supervised baseline. In our future work we will try to improve the way we deal with multiword terms and the propagation method to reduce the amount of erroneous aspect terms and generate a better ranking of the resulting terms.

Acknowledgements

This work has been partially funded by OpeNER[7] (FP7-ICT-2011-SME-DCL-296451) and SKaTer[8] (TIN2012-38584-C06-02).

References

Blair-Goldensohn, S. 2008. Building a sentiment summarizer for local service reviews. *WWW Workshop on NLP in the Information Explosion Era*.

Fei, Geli, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2012. A Dictionary-Based Approach to Identifying Aspects Im-plied by Adjectives for Opinion Mining. 2(December 2012):309–318.

Ganu, Gayatree, N Elhadad, and A Marian. 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content. *WebDB*, (WebDB):1–6.

Hai, Zhen, Kuiyu Chang, and Gao Cong. 2012. One seed to find them all: mining opinion features via association. *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 255–264.

Hu, Minqing and Bing Liu. 2004. Mining opinion features in customer reviews. *AAAI*.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pontiki, Maria, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Popescu, AM and Oren Etzioni. 2005. Extracting product features and opinions from reviews. *Natural language processing and text mining*, (October):339–346.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding Domain Sentiment Lexicon through Double Propagation. *IJCAI*.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, (July 2010).

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1533–1541. Association for Computational Linguistics.

Zhang, L, Bing Liu, SH Lim, and E O'Brien-Strain. 2010. Extracting and ranking product features in opinion documents. *Proceedings of the 23rd International Conference on Computational Linguistics*, (August):1462–1470.

Zhang, Lei and Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining. *Data Mining and Knowledge Discovery for Big Data*.

Zhuang, Li, F Jing, and XY Zhu. 2006. Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.

[7] http://www.opener-project.eu/
[8] http://nlp.lsi.upc.edu/skater/

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 129-136 recibido 14-04-14 revisado 15-07-14 aceptado 15-07-14

Boosting Terminology Extraction through Crosslingual Resources

Mejora de la extracción de terminología usando recursos translingües

Sergio Cajal, Horacio Rodríguez
Universitat Politècnica de Catalunya, TALP Research Center
Jordi Girona Salgado 1-3 edifici Omega D-316
Campus Nord 08034 Barcelona (Spain)
scajal@gmail.com, horacio@lsi.upc.edu

Resumen: La extracción de terminología es una tarea de procesamiento de la lengua sumamente importante y aplicable en numerosas áreas. La tarea se ha abordado desde múltiples perspectivas y utilizando técnicas diversas. También se han propuesto sistemas independientes de la lengua y del dominio. La contribución de este artículo se centra en las mejoras que los sistemas de extracción de terminología pueden lograr utilizando recursos translingües, y concretamente la Wikipedia y en el uso de una variante de PageRank para valorar los candidatos a término.
Palabras clave: Extracción de terminología. Procesamiento translingüe de la lengua. Wikipedia, PageRank

Abstract: Terminology Extraction is an important Natural Language Processing task with multiple applications in many areas. The task has been approached from different points of view using different techniques. Language and domain independent systems have been proposed as well. Our contribution in this paper focuses on the improvements on Terminology Extraction using crosslingual resources and specifically the Wikipedia and on the use of a variant of PageRank for scoring the candidate terms.
Keywords: Terminology Extraction, Wikipedia, crosslingual NLP, PageRank

1 Introduction

Terminology Extraction is an important Natural Language Processing, *NLP*, task with multiple applications in many areas. Domain terms are a useful mean for tuning both resources and *NLP* processors to domain specific tasks. The task is important and useful but it is also challenging. In (Krauthammer, Nenadic, 2004), it has been said that "terms identification has been recognized as the current bottleneck in text mining and therefore an important research topic in *NLP*".

Terms are usually defined as lexical units that designate concepts in a restricted domain. Term extraction (or detection) is difficult because there is no formal difference between a term and a non terminological unit of the language. Furthermore, the frontier between terminological and general units is not always clear and the belonging to a domain is more a fuzzy than a rigid function. (Hartmann, Szarvas and Gurevych, 2012) present the lexical units in a two dimensional space where *x* axe refers to *domainhood*, represented as a continuous, and *y* axe to *constituency* of the linguistic unit, i.e. single words and multiword expressions, *MWE*, (2-grams, 3-grams, etc.). Several types of *MWE* can be considered such as idioms, "kick the bucket", particle verbs, "fall off", collocations, "shake hands", Named Entities, "Los Angeles", compound nouns, "car park", some of which are compositional and other not. Obviously not all the *MWE* are terminological and not all the terms are *MWE*[1].

In this paper we prefer to refer to terms as term candidates (*TC*). As pointed out above, *TC*

[1] Many authors claim that most terms are MWE-From our experience we think that almost half of the TC extracted are single words.

ISSN 1135-5948

can be atomic lexical units or *MWE* composed by atomic units (usually named basic components of the term). There are some properties that must hold for a given *TC* in order to be considered a term: i) *unithood*, ii) *termhood* and iii) specialized usage. *Unithood* refers to the internal coherence of a unit: Only some sequences of POS tags can produce a valid term, N (e.g. "Hepatology" in the Medical domain), NN (e.g. "Blood test"), JN (e.g. "Nicotinic antagonist"), etc. and these combinations are highly language dependent), *termhood* to the degree a *TC* is related to a domain-specific concept and specialized usage (general language versus specialized domain). It is clear that measuring such properties is not an easy task. They can only be measured indirectly by means of other properties easier to define and measure like frequency (of the *TC* itself, its basic components or in relation to general domain corpus), association measures, syntactic context exploration, highlighting and/or structural properties, position in an ontology, etc.

We present in this paper a term ranker aimed to extract a list of *TC* sorted by *termhood*. Our claim is that the system is language and domain independent. In fact nothing in our approach depends on the language or the domain. The experiments and evaluation are carried out in two domains, *medicine* and *finance* and four languages: English, Spanish, Catalan, and Arabic.

Our approach is based on extracting for each domain the *TC* corresponding to all the languages simultaneously, in a way that the terms extracted for a language can reinforce the corresponding to the other languages. As unique knowledge sources we use the wikipedias of the involved languages.

Following this introduction, the paper is organized as follows. In section 2 we describe some recent work done in this area. Section 3 describes the methodology that we use to obtain new terms while section 4 describes the experiments carried out as well as its evaluation. Finally, in section 5, we present some conclusions and directions for future work.

2 Related work

Term extraction, *TE*, and related tasks (Term ranking, Named Entity Recognition, *MWE* extraction, lexicon and ontology building, multilingual lexical extraction, etc.) have been approached typically using linguistic knowledge, as in (Heidet al, 1996), or statistical strategies, such as ANA (Enguehard, Pantera, 1994), with results not fully satisfactory, see (Cabré, Estopà, Vivaldi, 2001) and (Pazienza. Pennacchiotti, Zanzotto, 2005). Also, *TE* systems often favor recall over precision resulting in a large number of *TC* that have to be manually checked and cleaned.

Some approaches combine both linguistic knowledge and Statistics, such as TermoStat (Drouin, 2003), or (Frantzi, Ananiadou and Tsujii, 2009), obtaining clear improvement. A common limitation of most extractors is that they do not use semantic knowledge, therefore their accuracy is limited. Notable exceptions are Metamap (Aronson, Lang, 2010) and YATE (Vivaldi, 2001).

Wikipedia[2], *WP*, is by far the largest encyclopedia in existence with more than 32 million articles contributed by thousands of volunteers. *WP* experiments an explosive growing. There are versions of *WP* in more than 300 languages although the coverage (number of articles and average size of each article) is very irregular. For the languages covered by the experiments reported here the size of the corresponding *WPs* are 4,481,977 pages in English, 1,091,299 in Spanish, 425,012 in Catalan, and 269,331 in Arabic. A lot of work has been performed for using this resource in a variety of ways. See (Medelyan et al, 2009) and (Gabrilovich, Markovitch, 2009) for excellent surveys.

WP has been, from the very beginning, an excellent source of terminological information. (Hartmann, Szarvas and Gurevych, 2012) present a good survey of main approaches, see also (Sabbah, Abuzir, 2005). Both the structure of *WP* articles (infoboxes, categories, redirect pages, input, output, and interlingual links, disambiguation pages, etc.) and their content have been used for *TE*. Figure 1 presents the bi-graph structure of *WP*. This bi-graph structure is far to be safe. Not always the category links denote belonging of the article to the category; the link can be used to many other purposes. The same problem occurs in the case of links between categories, not always these links denote hyperonymy/hyponymy and so the structure shown in the left of figure 1 is not a real taxonomy. Even worse is the case of inter-

[2] https://www.wikipedia.org/

page links where the semantics of the link is absolutely unknown.

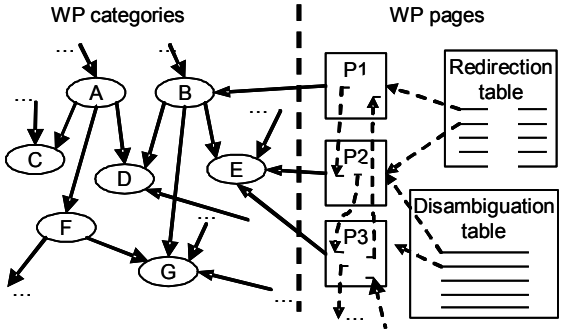


Figure 1: The graph structure of Wikipedia

Nakayama and colleagues, (Erdmann et al, 2008), and (Erdmann et al, 2009) face the problem of bilingual terminology extraction mainly using the interlingual links of *WP*, while (Sadat, 2011) uses, as well, the context of words and the Wikitionary[3]. Gurevych and collegues, (Wolf, Gurevych, 2010), (Niemann, Gurevych, 2011, map *WP* and WordNet[4], *WN*. Vivaldi and Rodríguez propose in (Vivaldi, Rodríguez, 2011) to use *WP* for extracting and evaluating term candidates in the medical domain, and in (Vivaldi, Rodríguez, 2012) propose to obtain lists of terms a multilingual/multidomain setting. (Alkhalifa, Rodríguez, 2010) use *WP* for enriching the Arabic WordNet with *NE*. The approaches more related to ours are those of Vivaldi and Rodríguez, both use WP categories and pages and their relations as knowledge sources, but there are clear differences: our use of interlingual links and the way of scoring candidates by means of the modified PageRank algorithm.

3 Our approach

The global architecture of our approach is displayed in Figure 2. As we can see it consists of 6 steps that are applied for each of the domains as detailed below. Let *d* be the domain considered (as we will see in Section 4 our experiments and evaluation have been carried out for medicine and finance). We will note WP^l the wikipedia for language *l* (*l* ranging on the four languages considered, i.e. en, sp, ca, ar).

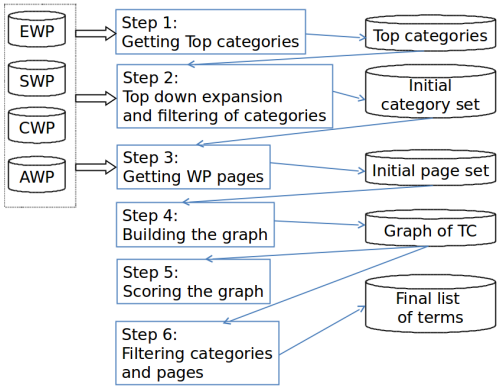


Figure 2: Architecture of our approach

As a preparatory step we downloaded the required *WP*s from the *WP* dumps site[5] and then we used the *JWLP*[6] toolbox, (Zesch, Müller, Gurevych, 2008) for obtaining a *MYSQL* representation of the *WP*s and the interlingual links. We then looked for the Top of the *WP* category graph (*topCat*), that for English *WP* corresponds to "Articles"[7]. Further on we enriched the *WP* category graph with the depth respect to *topCat* of all the categories. We have also downloaded the tables corresponding to the interlingual links. Although these links present problems of lack of reciprocity and inconsistency, see (De Melo, Weikum, 2010) for a method of facing these problems, we have made no attempt to face them and we have accepted all the links as correct.

In step 1 the top category of domain *d* is looked for in WP^{en}. Let $topCaDom^d$ be this category. Once located $topCaDom^d$ for English, the top categories for the other languages are obtained through the corresponding interlingual links. This is the only step requiring a small amount of human intervention.

In step 2, the initial set of categories is obtained for each language *l* by navigating top down, from the top category, through category/category links, the category graph of WP^l. Although ideally the *WP* category graph is a *DAG*, it is not really the case because two problems: i) the existence of cycles and ii) the presence of backward links.

Both problems have the same origin: the way of building the resource by lots of volunteers working independently. Many cycles

[3] http://www.wiktionary.org/

[4] http://wordnet.princeton.edu/

[5] http://dumps.wikimedia.org/

[6] http://www.ukp.tu-darmstadt.de/software/jwpl/

[7] In fact the real top category is "Contents", we have used "Articles" instead as topCat for avoiding that the shortest paths to the top traverse meta-categories.

occur in *WK*, an example, from the Spanish *WP,* is *Drogas → Drogas y Derecho → Narcotráfico → Drogas*. Detecting cycles and removing them is quite straightforward.

The second problem is more serious and difficult to face. When working with English *WP* we discovered that for the domain *Medicine* 90% of the whole *WP* category graph was collected as descendants of the domain top category. Consider the following example, from English *WP: Volcanology → Volcanoes → Volcanic islands → Iceland*. In this case going Top Down from the category *Volcanology* a lot of categories related to *Iceland*, but with no relation with *Volcanology* will be collected. For facing the second problem (backward links) we can take profit of the following information:

- The relative depth of each category *c* regarding $topCaDom^d$, i.e, the length of the shortest path from *c* to $topCaDom^d$.
- The absolute depth of c, computed in the preparatory step, i.e, the length of the shortest path from *c* to *topCat*.
- The absolute depths of *topCat* and $topCaDom^d$.
- The absolute depth of the parent of *c* in the dop down navigation.

We have experimented with several filtering mechanisms, from the very simplest one, pruning the current branch when the depth of *c* is lower than the depth of the parent of *c*, to others more sophisticated. Finally we decided to apply the following filtering: *c* is pruned, and not further expanded, if the relative depth of *c* is greater than the difference between the absolute depths of *topCat* and $topCaDom^d$ plus 1.

Applying this filtering mechanism resulted in reducing the set of involved categories (more than 900,000 without filtering) to a manageable number of 5,874 categories for English.

In step 3 we build the initial set of pages, collecting for each category in the set of initial categories the corresponding pages through the category/page links. The process is, so, quite straightforward. A simple filtering mechanism is performed for removing Named Entities and not content pages.

In step 4, from the two sets built in step 2 and 3 a graph representing the whole set of *TC* for the domain *d* and for all the languages is built. Figure 3 presents an excerpt of this graph. The nodes of the graph correspond to all the pages and categories selected in steps 2 and 3 for all the involved languages. The edges, which are directional, correspond to all the links considered (category → category, category → page, page → category, page → page and interlingual links).

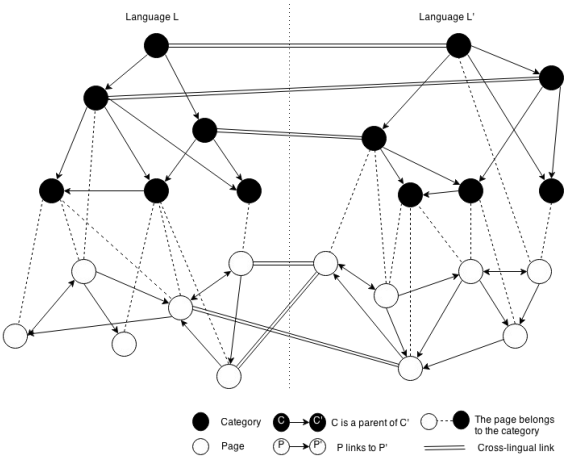


Figure 3: Graph representation of *TC*

In step 5 the nodes of the graph of *TC* are scored. For doing so we use an algorithm inspired in Topic-Sensitive PageRank, (Haveliwala. 2002), in turn based on the original PageRank algorithm, (Page, Brin, 1998).

The original PageRank algorithm is based on a scoring mechanism that allows for a given node upgrading its score accordingly with the scores of its incident nodes. So in this setting all the incident edges are equally weighted and the new score is only affected by the old one and the scores of the incident nodes. As is discussed en section 4, this setting does not work very well and we looked for some form of weighting of the edges, and not only of the nodes for computing the final score of a node.

In the case of nodes corresponding to pages there are three types of incident edges (for nodes corresponding to categories the formulas are similar):

- *il*: inlinks, links from other pages.
- *cp*: links from the categories the page belongs to.
- *ll*: langlinks, links from pages in other languages

The score of a page is computed by adding three weighted addends, one for each type of edge. The formula applied is the following:

$$PR(p) = F_{il} \sum_{il \in inlinks_p} \frac{PR(il)}{L(il)} + F_{cp} \sum_{c \in categories_p} \frac{PR(c)}{L(c)} + F_{ll} \sum_{ll \in langlinks_p} \frac{PR(ll)}{L(ll)}$$

where *PR(i)* is the PageRank score of node *i*, F_t are weights of edges of type *t* (*il, cp*, or *ll*), and L(n) are normalizing factors for pages or categories, computed as:

$$L(p) = F_{ol} \times |outlinks| + F_{pc} \times |cats| + F_{ll} \times |langlinks|$$

for pages and similarly for categories.

Finally, in step 6 the set of nodes corresponding to each language are sorted by descendent score giving the final result of the system. No distinction is made in this sorted sequence between *TC* corresponding to categories and these corresponding to pages.

4 *Experiments and evaluation*

4.1 Initial Settings

We performed some initial experiments for setting the parameters F_t defined in step 5. Finally we set F_{ll} and F_{cp} to 100 and the other parameters to 1. For evaluating these settings we limited ourselves to English and Spanish in the medical domain for which a golden repository of terms, *SNOMED*[8], is available. We consider four scenarios: i) *all_zeroes*, where no scoring procedure is used, ii) *all_ones*, where the standard *PageRank* algorithm is applied, iii) *no_langlinks*, where interlingual links weights are set to zero, i.e. the *TC* for each language are extracted independently, and, iv) *best*, where the setting described above was applied. The results are presented in Figures 4 , for English, and 5, for Spanish. All *PageRank* based scenarios clearly outperform the *all_zeroes* baseline. The differences between these scenarios are small for English but significant for Spanish where *best* outperforms clearly the others.

4.2 Experiments

We applied the procedure described in section 3 to the two domains and 4 languages using the setting of section 4.1. The results are presented in Table 1.

[8] http://www.ihtsdo.org/

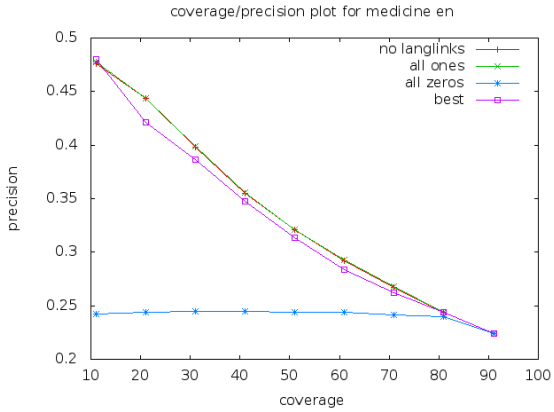


Figure 4: Initial experiments for English

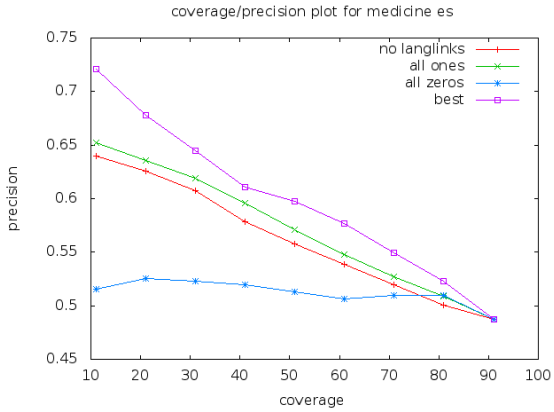


Figure 5: Initial experiments for Spanish

Language	Medicine	Finance
English	67,448	8,711
Spanish	8,872	1,310
Catalan	2,827	674
Arabic	7,318	1,557

Table 1: Overall results of our experiments

The figures in Table 1 are not very informative. Being our system a ranker what is important is accepting as true terms the best ranked until some threshold. We depict, so, in Figures 6 (for medicine) and 7 (for finance) the distribution of *TC* in a coverage/score plots[9]. Content of these Figures and Table 1 are somewhat complementary.

[9] Note that, contrary to Figures 4 and 5 where ordinates display precision, in this case ordinate display scores, i.e. PR values.

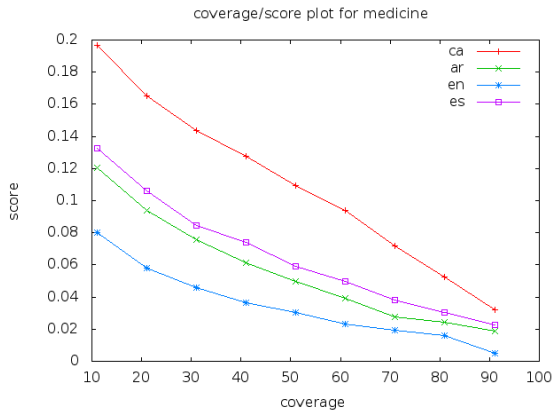


Figure 6: Results for medicine

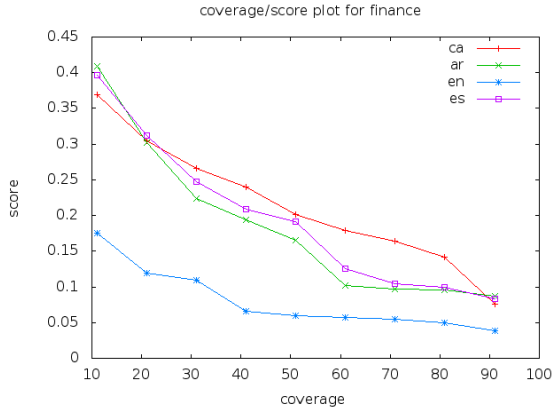


Figure 7: Results for finance

4.3 Evaluation

Evaluation of our results is not easy. For the pairs medicine/English and medicine/Spanish we can use as golden repository *SNOMED* and use as evaluation the results of the *best* curve in Figures 4 and 5. We have measured the correlation between precision in Figures 4 and 5 and score in Figure 6. Pearson's coefficient is 0.93 for Spanish and 0.98 for English, so we are pretty confident on our results for these two pairs. However, as pointed out in (Vivaldi, Rodríguez, 2012), *SNOMED* is far to be a reliable reference, for English only 62% of the correct *TC* were found in *SNOMED*. So the figures in Figures 4 and 5 can be considered a lower bound of the precision. For measuring a more accurate value we performed an additional manual validation[10] over the *TC* not found in *SNOMED* corresponding to the best 20% ranked ones. Figure 8 compares for this rank interval the precisions computed against *SNOMED* golden and those that combines it with the manual evaluation. At can be seen, results improvement is between 20 and 30 points.

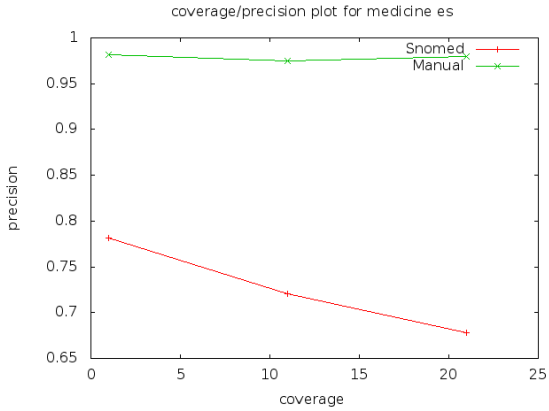


Figure 8: Comparison of SNOMED based and manual evaluation for Spanish

Obviously all these evaluations are in some cases partial and in other cases indirect. A point to be assessed is whether the evaluation results could be extrapolated to other domains and/or languages. For having some insights we computed the Pearson's correlation coefficient between the non-cumulated ranked scores for the different languages. Table 2 shows the results for *medicine*. A very similar result has been obtained for *finance*. The high values of these coefficients seem to support out hypothesis. The score distribution correlates well between all the languages for all the domains. At the beginning of this section we saw that for *medicine* and for the languages English and Spanish scores and precision correlated well too. So our guess is that the evaluation based on *SNOMED* for English and Spanish and the manual one for a segment of Spanish can be likely been extended to the other cases.

A comparison with other systems is not possible globally but we can perform some partial and indirect comparisons with the system closest to ours', (Vivaldi, Rodríguez, 2012). In this work, applied to Spanish and English, one of the domains included is *medicine* and *SNOMED* is used for evaluation. The main differences with ours' are that i) it is a term extractor, not a ranker and, ii) the evaluation is performed over terms belonging to *WordNet*. So the comparison has to be indirect. For the level of precision reported there, 0.2 for English, 0.4 for Spanish, the corresponding coverage in Figures 4 and 5 are 0.8 and 0.9. So, the number of terms we extract are 53,950 and

[10] Performed by the two authors independently, followed by a discussion on the cases with no agreement.

7,985 that clearly outperform largely the 21,073 and 4,083 reported there.

	en	es	ca	ar
en	1.0	0.996	0.990	0.992
es	0.996	1.0	0.995	0.994
ca	0.990	0.995	1.0	0.982
ar	0.992	0.994	0.982	1.0

Table 2: Correlations between non-cumulated ranked scores for the different languages

5 Conclusions and Future work

We have presented a terminology ranker, i.e. a system that provides a ranked list of terms for a given domain and language. The system is domain and language independent and uses as unique Knowledge Source the *Wikipedia* versions of the involved languages. The system proceeds in a cross-lingual way using for scoring a variant of the well-known *PageRank* algorithm.

We have applied the system to four languages and two domains. The evaluation, though not complete, and somehow indirect, and the comparison with a recent system closely related to ours', at least at the level of the source, shows excellent results clearly outperforming the subjects of ours comparisons.

Future work includes i) the application of the system to other domains and, possibly, to other languages and, ii) the improvement of the evaluation setting applying the system to domains for which terminology exists.

No attempt has been made to face the reciprocity and inconsistency of interlingual links. We plan in the near future to analyze these issues and to try to obtain aligned collections of multilingual terminologies.

The software and datasets described in this paper will be made publicly available in the near future through github.

Acknowledgements

The research described in this article has been partially funded by Spanish MINECO in the framework of project SKATER: Scenario Knowledge Acquisition by Textual Reading (TIN2012-38584-C06-01).

We are in debt with three anonymous reviewers whose advices and comments have contributed to a clear improvement of the paper.

References

Aronson, A., Lang, F., 2010. An overview of MetaMap: historical perspective and recent advances. *JAMIA 2010* 17:229-236.

Cabré, M.T., Estopà, R., Vivaldi, J., 2001. Automatic term detection. A review of current systems. *Recent Advances in Computational Terminology* 2:53-87.

Drouin, P., 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1):99-115.

Enguehard, C., Pantera, L., 1994. Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2(1):27-32.

Erdmann, M., Nakayama, K., Hara, T., Nishio, S., 2009. Improving the extraction of bilingual terminology from Wikipedia. *TOMCCAP* 5(4).

Erdmann, M., Nakayama, K., Hara, T., Nishio, S., 2008. An Approach for Extracting Bilingual Terminology from Wikipedia. *DASFAA 2008* 380-392.

Frantzi, K.T., Ananiadou, S., Tsujii, J., 2009. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. *Lecture Notes in Computer Science* 1513:585-604.

Gabrilovich, E., Markovitch, S., 2009. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 34:443-498.

Hartmann, S., Szarvas, G., Gurevych, I., 2012. Mining Multiword Terms from Wikipedia. M.T. Pazienza and A. Stellato: *Semi-Automatic Ontology Development: Processes and Resources* 226-258.

Haveliwala, T.H., 2002. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web (WWW '02)* 517-526.

Heid, U., Jauß, S., Krüger, K., Hohmann, A., 1996. Term extraction with standard tools for corpus exploration. Experience from German. *Proceedings of Terminology and Knowledge Engineering (TKE'96)*.

Alkhalifa, M., Rodríguez, H., 2010. Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia. *International Journal on Information and Communication Technologies* 3(3).

Krauthammer, M.I., Nenadic, G., 2004. Term identification in the biomedical literature. *Journal of Biomed Inform* 37(6):512-26.

Medelyan, O., Milne, D.N., Legg, C., Witten, I.H., 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716-754.

De Melo, G., Weikum, G., 2009. Untangling the Cross-Lingual Link Structure of Wikipedia, *48th Annual Meeting of the Association for Computational Linguistics*.

Niemann, E., Gurevych, I., 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet}. *Proceedings of the 9th International Conference on Computational Semantics* 205-214.

Page, L., Brin, S., 1998. The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh International Web Conference (WWW-98)*.

Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M., 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Soft Computing* 185:255-279.

Sabbah, Y.W., Abuzir, Y., 2005. Automatic Term Extraction Using Statistical Techniques- A Comparative In-Depth Study and Application. *Proceedings of ACIT'200.*

Sadat, F., 2011. Extracting the multilingual terminology from a web-based encyclopedia. *RCIS 2011* 1-5.

Vivaldi, J., 2001. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD Thesis, Universitat Politècnica de Catalunya.

Vivaldi, J., Rodríguez, H., 2008. Evaluation of terms and term extraction systems. A practical approach. *Terminology* 13(2):225-248. John Benjamins.

Vivaldi, J., Rodríguez, H., 2011. Using Wikipedia for term extraction in the biomedical domain: first experience. *Procesamiento del Lenguaje Natural* 45:251-254.

Vivaldi, J., Rodríguez, H., 2012. Using Wikipedia for Domain Terms Extraction. In Gornostay, T. (ed.) *Proceedings of CHAT 2012: The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources: co-located with TKE 2012*.

Wolf, E., Gurevych, I., 2010. Aligning Sense Inventories in Wikipedia and WordNet. *Proceedings of the First Workshop on Automated Knowledge Base Construction* 24-28.

Zesch, T., Müller, C., Gurevych, I., 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *LREC 2008: Proceedings of the Conference on Language Resources and Evaluation* 1646-1652.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 139-142 recibido 10-04-14 revisado 02-06-14 aceptado 10-06-14

Tratamiento inteligente de la información para ayuda a la toma de decisiones

Intelligent information processing to support decision-making

Sonia Vázquez, Elena Lloret, Fernando Peregrino, Yoan Gutiérrez, Javier Fernández, José Manuel Gómez
Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
{svazquez, elloret, fsperegrino, ygutierrez, javifm, jmgomez}@dlsi.ua.es

Resumen: Proyecto emergente centrado en el tratamiento inteligente de información procedente de diversas fuentes tales como micro-blogs, blogs, foros, portales especializados, etc. La finalidad es generar conocimiento a partir de la información semántica recuperada. Como resultado se podrán determinar las necesidades de los usuarios o mejorar la reputación de diferentes organizaciones. En este artículo se describen los problemas abordados, la hipótesis de trabajo, las tareas a realizar y los objetivos parciales alcanzados.
Palabras clave: Minería web, tratamiento de información textual, PLN

Abstract: This project is focused on intelligent information processing using different sources such as micro-blogs, blogs, forums, specialized websites, etc. The goal is to obtain new knowledge using semantic information. As a result we can determine user requirements or improve organizations reputation. This paper describes the problems faced, working hypothesis, tasks proposed and goals currently achieved.
Keywords: Web mining, textual information processing, NLP

1 Datos del proyecto

Proyecto dirigido por Sonia Vázquez, miembro del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. Financiado por la Universidad de Alicante (GRE12-44) dentro del programa de ayudas a proyectos emergentes. Inicio 01/09/2013 (duración 2 años).

Contacto
Email: svazquez@dlsi.ua.es
Teléfono: 965903400 ext. 2947
Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera San Vicente del Raspeig s/n,
03690, Alicante, España.

2 Introducción

Actualmente, las opiniones de los consumidores sobre diferentes tipos de productos y servicios están disponibles en Internet en diferentes lugares pudiendo ser expresadas a través de redes sociales, foros, portales especializados, blogs personales, etc. Mediante este tipo de información los usuarios que deseen adquirir un nuevo producto o servicio pueden consultar las críticas realizadas por diferentes usuarios acerca de ciertas características concretas. En muchas ocasiones, las críticas no son realizadas por expertos sino por usuarios reales que han probado el producto o el servicio y dan su opinión desde su punto de vista particular. En estos casos, las críticas pueden ser incompletas y centrarse únicamente en ciertos aspectos, de forma que el posible comprador o nuevo usuario recibe la información sesgada y debe buscar en diferentes lugares hasta conseguir una visión más completa de las características del producto. En ocasiones, la información que ofrecen algunos portales de Internet donde los usuarios pueden opinar acerca de diferentes productos viene acompañada de puntuaciones que indican el grado de utilidad de esa crítica, pudiendo seleccionar aquellas que tengan mayor puntuación para facilitar la toma de decisiones (Amazon, Ciao).

En los últimos años, se han realizado diversos estudios enfocados a la detección de la subjetividad en los textos llegando incluso a

ISSN 1135-5948

determinar la intensidad de dichas opiniones (disgusto, agrado, ironía, felicidad, etc). La información relativa a opiniones o críticas por parte de diferentes usuarios aparece dispersa en Internet por lo que es necesario establecer mecanismos que permitan una correcta búsqueda, recopilación y utilización de la misma. La tarea de descubrir conocimiento útil a través de información procedente de Internet es conocida como Web mining.

Dado el inminente interés generado por tratar la información subjetiva presente en la web se han desarrollado diversos recursos que permiten detectar el grado de afectividad presente en los textos. Estos recursos sirven como base de conocimiento para diferentes tipos de sistemas.

Manejar toda la información disponible y conseguir obtener una visión general de las opiniones de los usuarios es una tarea muy compleja y que requiere de diversas técnicas de PLN: detección de la informalidad, geolocalización, generación de resúmenes automáticos, resolución de la ambigüedad, etc.

3 Objetivos del proyecto

El objetivo principal de este proyecto es el tratamiento inteligente de la información textual procedente de diversas fuentes tales como micro-blogs, blogs, foros, portales especializados, etc. Mediante el uso de técnicas de PLN se extraerán de forma automática las principales características sobre un producto o un servicio y se recuperará información (opiniones de usuarios) relativa a estas características. Entre los problemas que se deben resolver se encuentran la detección de ironía o el sarcasmo, la ambigüedad, la geolocalización, la informalidad en los textos y la generación automática de resúmenes.

La extracción de características se realizará atendiendo a diferentes dominios de aplicación como por ejemplo el sector tecnológico, turístico, económico, etc. En cuanto al tema de la geolocalización se determinará la distribución geográfica de las páginas que hablan sobre un producto o un servicio para determinar en qué zonas ha tenido mejor acogida y en cuáles no. En este caso, se podrán restringir las búsquedas a ciertas zonas, como por ejemplo, la provincia de Alicante. La informalidad en los textos será tratada para transformar elementos de carácter informal en sus correspondientes implicaciones formales de manera que no se pierda el contenido semántico implícito en la descripción inicial. Además, las tareas de detección de opiniones y generación de resúmenes automáticos completarán el sistema para proporcionar unos resultados que generen nuevo conocimiento.

El interés de este proyecto viene determinado por la necesidad de disponer de información veraz de forma inmediata proveniente de diferentes fuentes y que abarque un amplio abanico de posibilidades. De esta forma tanto usuarios como entidades u organizaciones, tendrán el conocimiento necesario para poder tomar decisiones lo más acertadas posibles. De forma que un usuario pueda decantarse por la compra de un producto o servicio o una organización centre sus esfuerzos en mejorar ciertas partes de su imagen, sus productos o servicios.

4 Hipótesis de trabajo

Esta investigación se centra en la hipótesis de que la información procedente de las opiniones y críticas presentes en Internet puede ser utilizada para mejorar la reputación de organizaciones o determinar las necesidades de los usuarios para la creación de nuevos productos.

Debido al rápido crecimiento de la información en Internet los usuarios potenciales de un producto determinado tienen muchas dificultades a la hora de realizar una elección. Cualquier modificación en el producto, en las políticas de la empresa o en la atención al usuario puede variar de forma sustancial la opinión de los usuarios y posibles compradores. Poseer la información de las reacciones del público en general puede aportar beneficios a las diferentes partes implicadas tanto para la toma de decisiones por parte de los usuarios como para la mejora de los productos y servicios por parte de las organizaciones. Por ello, es necesario establecer qué necesidades se han visto cubiertas y cuáles quedan pendientes para poder desarrollar nuevos productos que satisfagan a la mayoría de usuarios. Mediante la utilización de técnicas de PLN se conseguirá generar conocimiento a partir de la información semántica, la localización, las opiniones, etc. De esta forma, se facilitará la toma de decisiones por parte de usuarios y organizaciones.

5 Tareas a desarrollar

Para la consecución del proyecto será necesario completar el conjunto de tareas y subtareas que se mencionan a continuación:

Análisis del problema

Se analizarán las distintas aproximaciones existentes para la extracción automática de características, detección del foco geográfico, informalidad en los textos, detección de opiniones y generación automática de resúmenes y conocimiento. Sobre esta base teórica se investigarán nuevas técnicas para la mejora de cada una de las actividades implicadas en el sistema.

Estudio de técnicas para la extracción automática de características

Se determinarán las características más relevantes relacionadas con un producto o servicio. Se estudiarán las aproximaciones actuales de extracción automática de características para desarrollar un sistema de dominio abierto que pueda ser aplicado a diferentes sectores (Dave, Lawrence, y Pennock, 2003), (Gamon et al., 2005).

Estudio de técnicas para la recuperación y extracción de información

Se realizará un proceso de recuperación y extracción de información cuya finalidad sea la de seleccionar de forma automática los textos relativos a esas características (Baeza-Yates y Ribeiro-Neto, 2011), (Manning, Raghavan, y Schütze, 2008). En este caso, se utilizarán diversas fuentes de información tales como micro-blogs, blogs, foros y portales especializados. Se estudiarán las técnicas actuales para recuperación y extracción de información tratando de adaptar dichas técnicas a nuestro ámbito de aplicación, resolviendo diferentes tipos de problemas como la detección de entidades entre otros (Kozareva, Vázquez, y Montoyo, 2007).

Estudio de técnicas para la detección de opiniones

Se realizará una clasificación basada en la opinión de los usuarios acerca de las diferentes características o servicios de los productos. Se realizará un estudio de las técnicas actuales adaptando las mejores aproximaciones a cada entorno específico. Técnicas basadas en la combinación de características semánticas y adjetivos (Hatzivassiloglou y Wiebe, 2000), métodos basados en conocimiento (Gutiérrez, Vázquez, y Montoyo, 2011), técnicas basadas en la extracción de términos relevantes que definan la polaridad (Zubaryeva y Savoy, 2010), métodos basados la proximidad semántica entre conceptos (Balahur y Montoyo, 2009).

Estudio de técnicas para la detección de la informalidad

Se tratará la informalidad de los textos sobre opiniones y críticas de una forma específica (Mosquera y Moreda, 2012). Se determinarán las aproximaciones más adecuadas para tratar textos provenientes de diversas fuentes como: micro-blogs, blogs, foros y portales especializados (Buscaldi y Rosso, 2008).

Estudio de técnicas para la geolocalización

Se realizará un estudio sobre diferentes técnicas de desambiguación (Vázquez, Montoyo, y Kozareva, 2007) y detección de entidades. Debido a que un mismo topónimo puede pertenecer a diferentes lugares es crucial establecer de qué lugar concreto se está hablando (Peregrino, Tomás, y Llopis, 2011).

Estudio de técnicas para la generación de resúmenes y nuevo conocimiento

El objetivo de esta tarea es la determinar qué información será la más relevante para un usuario y sintetizarla en forma de un pequeño texto. Debido a la gran cantidad de información existente los usuarios son incapaces de manejar de manera eficiente dicha información. Por tanto, a partir de toda la información recopilada y clasificada según el grado de aceptación de los usuarios se estudiarán nuevas técnicas para la generación de conocimiento que permitan sintetizar de forma coherente y veraz la gran cantidad de información relacionada con un producto o servicio concreto (Barzilay y McKeown, 2005), (Lloret y Palomar, 2013).

6 Situación actual del proyecto

Dentro de las tareas antes mencionadas, hasta el momento se ha desarrollado un recurso que relaciona diferentes bases de datos léxicas y ontologías junto con el grado de afectividad (Gutiérrez et al., 2011): SentiWordnet, WordNet, SUMO, WordNet Affect. Además, se ha mejorado un sistema de desambiguación basado en conocimiento para adecuarlo a las necesidades del proyecto. Y por último, se han estudiado diversas técnicas para mejorar la generación automática de resúmenes.

Bibliografía

Baeza-Yates, Ricardo A. y Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Balahur, Alexandra y Andrés Montoyo. 2009. A semantic relatedness approach to classifying opinion from web reviews. *Procesamiento del Lenguaje Natural*, 42.

Barzilay, Regina y Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328, Septiembre.

Buscaldi, Davide y Paolo Rosso. 2008. Geo-wordnet: Automatic georeferencing of wordnet. En *LREC*. European Language Resources Association.

Dave, Kushal, Steve Lawrence, y David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. En *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, páginas 519–528, New York, NY, USA. ACM.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver, y Eric Ringger. 2005. Pulse: Mining customer opinions from free text. En *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, páginas 121–132, Berlin, Heidelberg. Springer-Verlag.

Gutiérrez, Yoan, Antonio Fernández Orquín, Sonia Vázquez, y Andrés Montoyo. 2011. Enriching the integration of semantic resources based on wordnet. *Procesamiento del Lenguaje Natural*, 47:249–257.

Gutiérrez, Yoan, Sonia Vázquez, y Andrés Montoyo. 2011. Sentiment classification using semantic features extracted from wordnet-based resources. En *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, páginas 139–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hatzivassiloglou, Vasileios y Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. En *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, páginas 299–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kozareva, Zornitsa, Sonia Vázquez, y Andrés Montoyo. 2007. Multilingual name disambiguation with semantic information. En *TSD*, páginas 23–30.

Lloret, Elena y Manuel Palomar. 2013. Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering*, 19(2):147–186.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Mosquera, Alejandro y Paloma Moreda. 2012. The study of informality as a framework for evaluating the normalisation of web 2.0 texts. En *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, NLDB'12, páginas 241–246, Berlin, Heidelberg. Springer-Verlag.

Peregrino, Fernando S., David Tomás, y Fernando Llopis. 2011. Map-based filters for fuzzy entities in geographical information retrieval. En *NLDB*, páginas 270–273.

Vázquez, Sonia, Andrés Montoyo, y Zornitsa Kozareva. 2007. Word sense disambiguation using extended relevant domains resource. En *IC-AI*, páginas 823–828.

Zubaryeva, Olena y Jacques Savoy. 2010. Opinion detection by combining machine learning & linguistic tools.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 143-146 recibido 10-04-14 revisado 02-06-14 aceptado 02-06-14

Proyecto FIRST (Flexible Interactive Reading Support Tool): Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos

FIRST (Flexible Interactive Reading Support Tool) project: developing a tool for helping autistic people by document simplification

María-Teresa Martín Valdivia, Eugenio Martínez Cámara, Eduard Barbu, L. Alfonso Ureña López
SINAI - Universidad de Jaén
Campus Las Lagunillas s/n, 23071, Jaén
{maite,eduard,emcamara,laurena}@ujaen.es

Paloma Moreda, Elena Lloret
GPLSI-Universidad de Alicante
Campus de San Vicente del Raspeig, 03080, Alicante
{moreda,elloret}@dlsi.ua.es

Resumen: El Trastorno de Espectro Autista (TEA) es un trastorno que impide el correcto desarrollo de funciones cognitivas, habilidades sociales y comunicativas en las personas. Un porcentaje significativo de personas con autismo presentan además dificultades en la comprensión lectora. El proyecto europeo FIRST está orientado a desarrollar una herramienta multilingüe llamada Open Book que utiliza Tecnologías del Lenguaje Humano para identificar obstáculos que dificultan la comprensión lectora de un documento. La herramienta ayuda a cuidadores y personas con autismo transformando documentos escritos a un formato más sencillo mediante la eliminación de dichos obstáculos identificados en el texto. En este artículo se presenta el proyecto FIRST así como la herramienta desarrollada Open Book.
Palabras clave: Procesamiento de Lenguaje Natural, Simplificación de textos, Trastorno del Espectro Autista (TEA)

Abstract: Autism Spectrum Disorder (ASD) is a condition that impairs the proper development of people cognitive functions, social skills, and communicative abilities. A significant percentage of autistic people has inadequate reading comprehension skills. The European project FIRST is focused on developing a multilingual tool called Open Book that applies Human Language Technologies (HLT) to identify reading comprehension obstacles in a document. The tool helps ASD people and their carers by transforming written documents into an easier format after removing the reading obstacles identified. In this paper we present the FIRST project and the developed Open Book tool.
Keywords: Natural Language Processing, Text simplification, Autism Spectrum Disorder (ASD)

1 Introducción

El Trastorno del Espectro Autista (TEA) o simplemente autismo (en inglés, Autistic Spectrum Disorder - ASD) es un trastorno del desarrollo neurológico caracterizado por la alteración cualitativa de la comunicación y conductas repetitivas estereotipadas (Mesibov, Adam y Klinger, 1997). Las personas con TEA tienen un déficit en la comprensión del lenguaje, incluyendo una mala interpretación de los significados no literales y la dificultad para comprender instrucciones complejas. Las frases hechas, abstracciones, palabras poco comunes y ambiguas son a menudo fuente de confusión para estas personas.

Investigaciones recientes estiman que alrededor de 3 millones de personas en Europa presentan un diagnóstico de autismo (Barthélémy et al., 2008). Tanto los síntomas como las dificultades del TEA varían bastante de unos individuos a otros, si bien la mayoría presentan dificultad en mayor o menor media en los siguientes aspectos:

ISSN 1135-5948

- Dificultades en comunicación.
- Dificultades en interacción social.
- Dificultades en comprensión de cierta información.

Dentro del espectro autista se presentan distintos grados y manifestaciones del trastorno. En este sentido encontramos desde individuos totalmente aislados con un bajo nivel de interacción con las personas, una tendencia a la repetición de actividades motoras y una falta completa del desarrollo del lenguaje y comunicación alternativa, hasta personas con un lenguaje muy desarrollado y casi sin alteraciones aparentes. Es este último caso, los niveles cognitivos de estas personas pueden presentar habilidades a nivel de inteligencia que superen la media normal en un área específica del desarrollo.

El proyecto FIRST[1] (Flexible Interactive Reading Support Tool) financiado por la Unión Europea dentro del 7º programa marco (FP7-2007-2013- n° 287607) tiene como objetivo principal desarrollar, implementar y evaluar tecnologías para apoyar la creación de contenido accesible para usuarios con TEA. Su finalidad es facilitar la lectura y comprensión de textos para personas autistas, y de esta manera permitirles una mayor inclusión en la sociedad. El proyecto tiene una duración total de 3 años (empezó el 1 de octubre de 2011 y termina el 30 de septiembre de 2014) e incluye un equipo multidisciplinar de socios técnicos, científicos y clínicos distribuidos entre 4 países europeos (España, Reino Unido, Bélgica y Bulgaria).

Como producto final, el proyecto FIRST ha desarrollado la herramienta software Open Book. A través de una plataforma online, esta herramienta permite la simplificación de documentos escritos en tres idiomas distintos: inglés, español y búlgaro.

2 Objetivos del proyecto FIRST

En realidad, el proyecto persigue dos grandes objetivos: por un lado, existe un reto tecnológico donde se han aplicado y desarrollado tecnologías innovadoras en tecnologías del lenguaje capaces de tratar un texto cualquiera (un libro, una revista, documentos de la web…) para detectar y simplificar, siempre que sea posible, cualquier complejidad que incluya (conceptos difíciles, metáforas, texto figurado…) y obtener otro documento que sea más fácil de comprender.

Por otra parte, el segundo gran objetivo del proyecto FIRST consiste en la evaluación del impacto que la aplicación de dichas tecnologías tiene en las personas con autismo. Por este motivo, los socios del consorcio se distribuyen entre socios tecnológicos y socios relacionados con el ámbito médico que permiten la adecuada evaluación de la herramienta en la comunidad autista.

El proyecto se espera que tenga un impacto en la calidad de vida de las personas con autismo, así como mejorar su acceso a la educación y obtener mayores oportunidades de formación profesional, cultural y social, favoreciendo su inclusión social.

3 Resultado del proyecto FIRST: La herramienta Open Book

Como resultado final del proyecto FIRST, se ha implementado una herramienta software online, denominada Open Book. Mediante esta herramienta, basada en tecnologías del lenguaje, se pueden simplificar textos permitiendo su personalización y adaptación a cada usuario, de manera que, dependiendo de las dificultades que éste tenga, se podrán activar y desactivar las funcionalidades asociadas a la detección y resolución de los distintos tipos de obstáculos tratados. Adicionalmente, la herramienta incluye dos modos de operación: modo cuidador y modo usuario final. La principal diferencia entre ambos modos radica en que accediendo a través del modo cuidador se pueden editar los textos simplificados de forma automática, permitiendo la revisión, corrección y modificación de dichos documentos para una mejor adaptación a cada uno de los usuarios finales.

La tecnología desarrollada convertirá los documentos, a los que los usuarios desean acceder, en una forma personalizada para facilitar la comprensión de lectura. Este proceso de conversión incluye la detección automática de rasgos lingüísticos en el documento de entrada que pueden obstaculizar la comprensión, así como la eliminación automática de los obstáculos identificados, en la medida de lo posible, de tal forma que el significado del documento original no se vea afectado al realizar la transformación.

Las tecnologías del lenguaje se han aplicado para detectar y eliminar los obstáculos en la

[1] http://www.first-asd.eu/

comprensión causados por la complejidad estructural (por ejemplo, frases muy largas) y la ambigüedad en el significado de los documentos escritos (por ejemplo, presencia de palabras polisémicas de uso infrecuente).

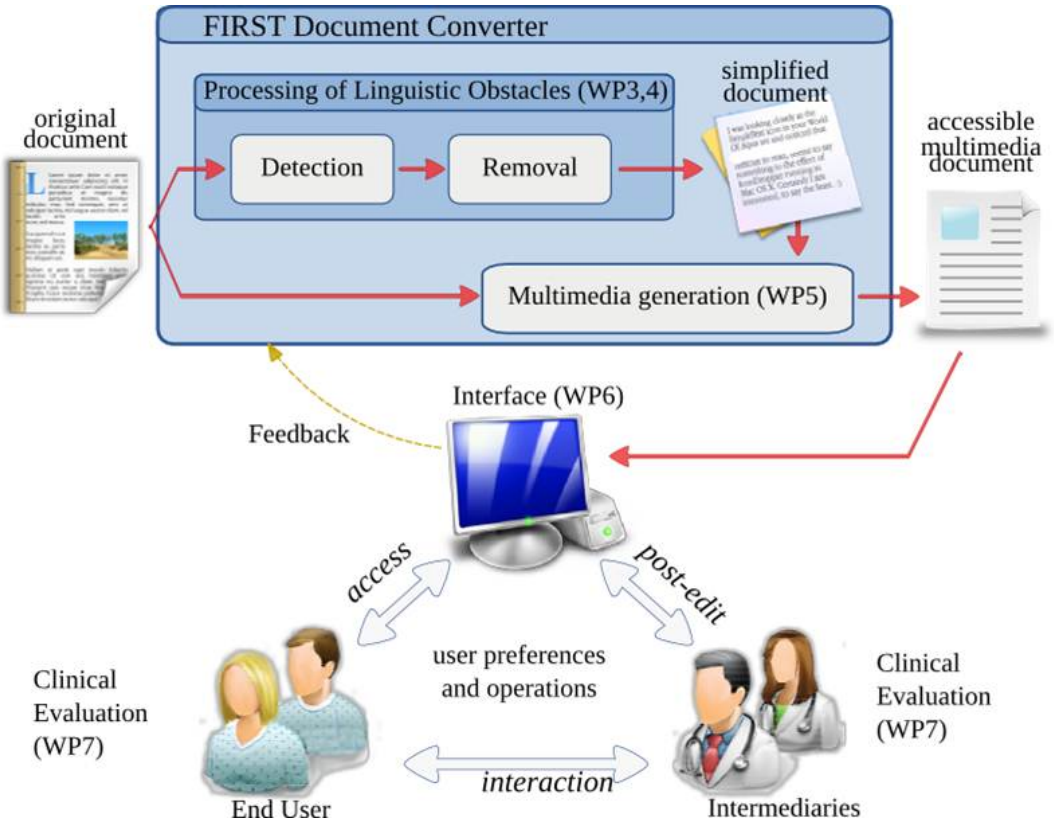


Figura 1: Arquitectura general de la herramienta Open Book

Por otra parte, dado que cada individuo con autismo es totalmente distinto a otro, la herramienta se ha diseñado para que pueda ser personalizada y ajustada, a través de su interfaz, para cada usuario en particular. De esta manera, por ejemplo si un individuo presenta una mayor dificultad en la comprensión de conceptos técnicos, es posible activar solamente la opción para que detecte y resuelva este tipo de obstáculos, mientras que, si por el contrario, para otro usuario, la inclusión de definiciones para conceptos técnicos no supone ninguna ventaja es posible desactivar dicha funcionalidad.

La Figura 1 muestra la arquitectura general del sistema que se ha desarrollado. En ella, se muestra que la herramienta puede ser utilizada tanto por los usuarios finales (personas con autismo) como por sus intermediarios (sus cuidadores). Éstos últimos podrán además, revisar y post-editar el documento simplificado para realizar pequeños ajustes y adaptaciones en el texto, en función del usuario final. Dado que se trabaja con usuarios reales, la herramienta se evaluará en un entorno clínico controlado bajo la supervisión de profesionales clínicos (socios del ámbito médico) que normalmente tratan con pacientes con TEA.

4 Funcionamiento de Open Book

Open Book tiene como objetivo asistir a las personas con autismo para acceder a la información mediante la reducción de ciertas barreras lingüísticas, permitiéndoles o ayudándoles a leer una amplia variedad de textos sin ayuda. Para ello, Open Book utiliza las tecnologías del lenguaje humano para adaptar el lenguaje empleado en textos escritos mediante la reducción de la complejidad estructural y lingüística, la supresión de la ambigüedad y la mejora de la legibilidad. Entre

las funcionalidades que se implementan, se destacan:

- Inclusión de definiciones y sinónimos para palabras y expresiones poco comunes, largas, técnicas y/o ambiguas.
- Sustitución de expresiones figuradas (por ejemplo; metáforas, lenguaje figurado o frases hechas) por lenguaje literal.
- Inclusión de imágenes para conceptos complejos o relacionados con emociones.
- Generación de resúmenes concisos.
- Desarrollo de herramientas de navegación para textos extensos, como por ejemplo índices o tablas de contenido.

5 Impacto esperado

El objetivo del proyecto FIRST es ayudar a las personas con autismo a leer documentos con mayor confianza y autonomía. Por tanto, se pretende ayudar a esta comunidad a romper algunas de las barreras lingüísticas a las que se enfrentan en la vida diaria, con el fin de incrementar su participación e inclusión en todos los aspectos de la sociedad, incluyendo educación, empleo, sanidad y actividades sociales.

Las tecnologías desarrolladas no sólo están destinadas a personas con autismo sino que podrán ser utilizadas por cualquier persona que tenga dificultades en la comprensión lectora, desde individuos con dislexia, parálisis cerebral o no nativos que estén aprendiendo un idioma, hasta personas con un bajo nivel de alfabetización o con problemas de aprendizaje. Esto es posible gracias a que parte de los obstáculos estructurales y lingüísticos tratados en el proyecto FIRST se pueden aplicar a otros ámbitos. Así pues, el fin último que se desea obtener es mejorar la calidad de vida de las personas en riesgo de exclusión social debido a su falta de comprensión lectora.

6 Lista de participantes

El consorcio del proyecto además de incluir varias universidades y centros de investigación también está formado por organizaciones involucradas en la educación y el cuidado de las personas con TEA, asegurando así que los resultados del proyecto se difundan amplia y eficazmente a las partes interesadas. Concretamente, el consorcio está formado por 9 socios: 5 socios tecnológicos y 4 socios clínicos. Estos socios están distribuidos entre 4 países europeos (España, Reino Unido, Bélgica y Bulgaria).

- Socios Científicos y Tecnológicos
  - University of Wolverhampton, Reino Unido
  - Universidad de Alicante, España
  - Universidad de Jaén, España
  - iWeb Technologies LTD, Reino Unido
  - Kodar OOD, Bulgaria
- Socios clínicos
  - Central and North West London NHS
  - Foundation Trust, Reino Unido
  - Parallel World Sdruzhenie, Bulgaria
  - Deletrea SL, España
  - Autism-Europe aisbl, Bélgica

Agradecimientos

La investigación que desarrolla este producto de software ha recibido financiación del Séptimo Programa Marco de la Comunidad Europea (FP7-2007-2013), en virtud del acuerdo de subvención n° 287607. También ha sido parcialmente financiada por el gobierno español a través del proYecto ATTOS (TIN2012-38536-C03-0), el gobierno regional de la Junta de Andalucía a través del proyecto AORESCU (TIC - 07684) y la Generalitat Valenciana, mediante la acción complementaria ACOMP/2013/067.

Bibliografía

Barthélémy, C., J. Fuentes, P. Howlin, Rutger J. V. der Gaag, 2008, Persons with Autism Spectrum Disorder: Identification, Understanding & Intervention.

Mesibov, G.B., L.W. Adams, L.G. Klinger. 1997. Autism: Understanding the disorder. New York, Plenum Press.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 147-150 recibido 11-04-14 revisado 05-06-14 aceptado 10-06-14

Open Idea: Plataforma inteligente para gestión de ideas innovadoras*

Open Idea: An intelligent platform for managing innovative ideas

Miguel Ángel Rodríguez-García, Rafael Valencia-García
Universidad de Murcia
Facultad de Informática
Campus de Espinardo, 30100, Murcia, España
miguelangel.rodriguez@um.es, valencia@um.es

Gema Alcaraz-Mármol
Universidad Católica San Antonio de Murcia
Campus de los Jerónimos, 135, 30107 Guadalupe, Murcia, España
galcaraz@ucam.edu

César Carralero
Quality Objects S.L.
C\ Santa Leonor 65. Edif C, 2 Izq Madrid, España
ccarralero@qualityobjects.com

Resumen: La finalidad del proyecto OPEN IDEA es el desarrollo de una herramienta que permita gestionar de manera eficiente las ideas innovadoras dentro de una organización, mediante el uso de tecnologías semánticas y del procesamiento del lenguaje natural. El objetivo central del sistema es fomentar el concepto de innovación abierta facilitando, durante todo el proceso de gestión de ideas, la interacción entre usuarios de la organización con las ideas innovadoras aportadas. Este proyecto está siendo desarrollado conjuntamente por la empresa QualityObjects y el grupo TECNOMOD de la Universidad de Murcia y ha sido financiado por el Ministerio de Industria, Energía y Turismo a través de la convocatoria de Avanza Competitividad I+D de 2012.
Palabras clave: anotación semántica, indexación semántica, innovación abierta, ontologías

Abstract: The main goal of the OPEN IDEA Project is the development of a platform which efficiently manages the innovative ideas within an organization by using semantic technologies and natural language processing. The main challenge of this system is to promote the concept of Open Innovation in the enterprise by making easier the interaction between the organization users and the innovative ideas proposed during the whole management process. This project is being jointly developed by the Quality Objects Enterprise and the TECNOMOD resarch group from the University of Murcia, and it is funded by the Ministry of Industry, Energy and Tourism (Research and Development programme Avanza Competitividad 2012).
Keywords: semantic tagging, semantic indexing, open innovation, ontologies

1 Introducción y objetivos del proyecto

La innovación abierta es un paradigma que asume que las empresas pueden y deben utilizar todos los flujos de información disponibles provenientes de canales externos o internos a la organización, para mejorar procesos de innovación internos y externos (Chesbrough, 2003). Para llevar a cabo este modelo de flujos de información es necesario una herramienta colaborativa que permita gestionarlos de manera eficiente.

En este sentido, la Web 2.0 proporciona el escenario colaborativo idóneo para la gestión de flujos de información internos y externos en una organización. Las redes sociales son grandes contenedores de información diseñadas para compartir datos fácilmente entre personas.

Otro aspecto importante a tener en cuenta es el modelado de la información. Éste es esencial para extraer el conocimiento necesario que

* Este trabajo ha sido financiado por el Ministerio de Industria, Energía y Turismo a través del proyecto OPEN IDEA (TSI-020603-2012-219)

proporcione información relevante a la organización, y así lograr un adecuado posicionamiento en el mercado. La utilización de tecnologías semánticas como esquemas conceptuales facilitan el modelado de estos flujos de información a través de ontologías, sobre las que se puede aplicar inferencias que ayuden a simplificar las tareas de gestión del ciclo de vida de las ideas. Por otro lado, las ideas se suelen expresar en lenguaje natural por lo que es necesario incorporar este tipo de tecnologías para poder extraer el conocimiento de estas fuentes de información.

El principal reto de este proyecto es el desarrollo de una plataforma que fomente la innovación abierta en las organizaciones. La plataforma utilizará las tecnologías de la Web Semántica para generar ese valor adicional que proporcione un factor diferenciador a la organización con respecto a sus competidores.

2 Estado actual del proyecto

Hasta el momento se han definido los dos módulos principales en los que se compone la plataforma (ver Figura 1): el entorno colaborativo y la plataforma semántica. Para el entorno colaborativo se ha realizado un estudio de redes sociales que permitan seleccionar la solución adaptable a las características del proyecto. En cuanto a la plataforma semántica, se ha implementado un servidor web que dispone de un conjunto de módulos que permiten explotar las tecnologías semánticas y de procesamiento del lenguaje natural mediante el desarrollo de diversos servicios Web. Esta plataforma sirve como sistema de ayuda a la decisión, permitiendo así una mejor evaluación y gestión de las ideas en la plataforma. Por ejemplo, entre los servicios que ofrece encontramos la identificación de ideas similares, las organizaciones más idóneas para participar en una idea, o bien la configuración de un equipo de trabajo para el desarrollo de esta idea en base a su experiencia y currículum.

Por otro lado, se han desarrollado dos ontologías que modelan el dominio de las Tecnologías de Información y Comunicación (TICs) y el dominio de gestión de las ideas. La primera ontología se utiliza en los procesos de anotación semántica, mientras que la segunda se utiliza para organizar la información relacionada con los perfiles semánticos en el repositorio de ontologías. Para ello, por cada entidad importante de información de la plataforma (ya sean ideas, proyectos, trabajadores u organizaciones) se crea un perfil semántico con su información.

A continuación se describe brevemente la arquitectura del sistema en el estado actual del proyecto.

2.1 Arquitectura de la plataforma OPEN IDEA

El sistema OPEN IDEA (ver Figura 1) se basa en dos componentes principales: la red social donde se encuentran todas las funciones relacionadas con la gestión de perfiles, documentos, flujos de información, etc., y la plataforma semántica que ofrece servicios de consulta de información y anotación semántica.

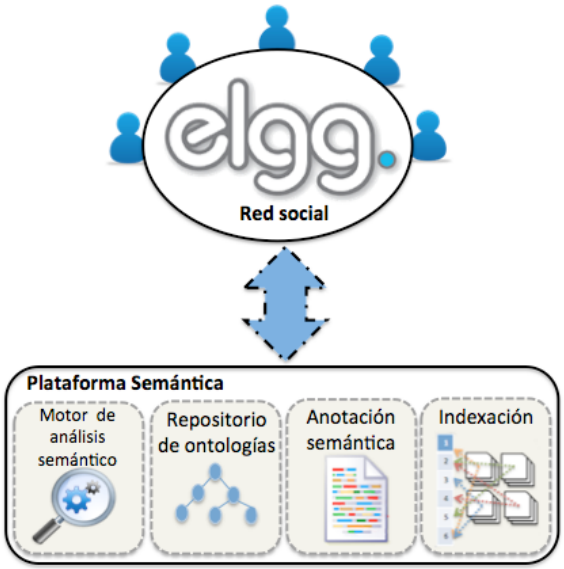


Figura 1: Arquitectura de OPEN IDEA

A continuación se describen cada uno de los módulos de la plataforma.

2.2 Red social

Elgg[1] es el motor de redes sociales que se ha seleccionado para proporcionar el entorno colaborativo necesario en el proyecto. Esta red social proporciona un marco sólido para definir varios tipos de redes sociales. Elgg está basado en una jerarquía bien definida de objetos que proporcionan interfaces sencillas, las cuales

[1] http://elgg.org

simplifican la extensibilidad y adaptación del motor de redes sociales a cualquier entorno colaborativo. Una vez instalado, el framework por defecto combina un conjunto de herramientas sociales como: blogs, gestión de ficheros y perfiles de usuarios, canales RSS, marcadores sociales, y gestión de redes sociales de uso personal y de grupo. Los aspectos más relevantes que facilitaron la selección de este framework como herramienta colaborativa fueron, en primer lugar, su sencilla y rápida instalación; y en segundo lugar, su arquitectura basada en plugins que hace sencilla su extensibilidad.

2.3 Plataforma Semántica

Esta plataforma tiene la función de proveer a la red social de los servicios inteligentes que apoyarán la toma de decisiones. Es por ello que su funcionamiento ha sido implementado como un servidor web. Este servidor web está formado por un conjunto de subsistemas que implementan las funciones ofrecidas por la plataforma semántica como servicios.

Los subsistemas que componen la plataforma semántica son: el buscador semántico, que ofrece funciones de búsqueda en lenguaje natural; el repositorio de ontologías que almacena los modelos ontológicos del dominio; el módulo de anotación semántica, encargado de anotar recursos semánticamente; y por último el módulo de indexación, que se encarga de generar índices semánticos que faciliten los procesos de búsqueda y la detección de perfiles similares. Los siguientes apartados se dedicarán a realizar un análisis de cada uno de los subsistemas que componen la plataforma semántica.

2.3.1 Anotación semántica

El objetivo de este subsistema es mantener y crear anotaciones semánticas basadas en una ontología del dominio. De esta forma, en el caso de estudio, el subsistema recibe como entrada una ontología que modela el dominio de Tecnologías de Información y Comunicación (TICs), además de descripciones en lenguaje natural sobre cualquier entidad definida en la ontología de gestión de las ideas.

El primer problema que se abordó durante el desarrollo de este subsistema fue la definición del dominio de las (TICs) a través de una ontología, y sobre todo cómo organizar la información utilizando los diferentes elementos (clases, propiedades, relaciones y axiomas) de los que dispone una ontología para definir de manera jerárquica este dominio. La solución a este problema se llevó a cabo mediante la selección de Wikipedia como fuente de información que facilitara la extracción de estas relaciones taxonómicas.

El funcionamiento del subsistema de anotación semántica está basado en el trabajo presentado en (Rodríguez-García et al., 2014) y se divide en dos fases principales: identificación y clasificación. La primera fase se encarga de identificar las expresiones lingüísticas más relevantes a través de cálculos estadísticos basados en la estructura sintáctica del texto. Durante la segunda fase el sistema intenta determinar, para cada expresión lingüística identificada, una relación con una instancia de una clase en la ontología definida. Esta fase se implementó mediante el framework GATE, el cual apoya todo el proceso de anotación semántica.

En la plataforma se anota semánticamente cada perfil semántico de las entidades importantes como por ejemplo ideas, proyectos, trabajadores y organizaciones.

2.3.2 Indexación

El subsistema de indexación utiliza las anotaciones semánticas para crear índices semánticos que optimicen el trabajo del buscador semántico, y así facilitar los procesos de búsqueda. El funcionamiento de este sistema se basa en el trabajo (Castells et al., 2007). Cada anotación semántica creada durante el proceso anterior se almacena en una base de datos relacional con un valor numérico, el cual representa cómo de relevante es la entidad ontológica con respecto a los perfiles semánticos anotados. Para calcular este peso se ha utilizado la fórmula Term Frequency – Inverse Document Frequency (TF-IDF) (Salton y McGill, 1986), medida numérica que expresa cuán relevante es una palabra para un documento en una colección.

El conjunto de anotaciones definidas para cada perfil semántico junto con sus pesos asignados constituyen un vector semántico calculado a través de la adaptación del modelo espacio vectorial presentado en el trabajo (Castells et al., 2007).

2.3.3 Motor de análisis semántico

El objetivo de este subsistema es proporcionar un servicio de búsqueda de perfiles semánticos basado en la identificación de posibles similitudes entre distintos perfiles y entidades de la plataforma. El subsistema recibe como entrada un recurso concreto como puede ser una idea, trabajador o una organización, y obtiene distintas entidades similares en base a los índices semánticos de los perfiles de cada entidad insertada en la plataforma.

Como se ha comentado anteriormente, cada entidad dispone de un vector creado por el módulo de indexación que representa los conceptos de los que trata esa entidad. Por lo tanto, se compara el vector de la entidad seleccionada con todos los vectores de las otras entidades del sistema. Para ello, se utiliza la función de similitud del coseno (Singhal, 2001) que permite, mediante cálculos vectoriales sencillos, obtener el grado de similitud entre cada par de vectores semánticos.

Además, este motor ofrece opciones que permitan filtrar las búsquedas en base al tipo de entidad buscada: ideas, proyectos, trabajadores, organización, mercado, tecnología, etc.

2.3.4 Repositorio de ontologías

El principal objetivo de este subsistema es el almacenamiento de ontologías. Por un lado, se tiene la ontología de las TICs y por otro cada perfil semántico definido por cada entidad existente en la red social. Este repositorio se ha implementado utilizando Virtuoso[2], un repositorio semántico donde se almacena la información en formato Resource Description Framework (RDF).

El objetivo del repositorio es recolectar información sobre instancias de la ontología de gestión de ideas. Esta ontología se implementó en OWL. Entre las clases y entidades principales de esta ontología podemos destacar: idea, propuesta, proyecto, trabajador, organización, grupo, mercado y tecnología. Estas clases se encuentran relacionadas entre sí mediante diferentes relaciones que, en conjunto, modelan el dominio de la gestión de la innovación en una organización.

3 Trabajo futuro

Actualmente el entorno colaborativo se ha adaptado a los requisitos del proyecto, desarrollando nuevos plugins que integran en la red social nuevas entidades relacionadas con la gestión de innovación empresarial. También se ha desarrollado la plataforma semántica, que estará constituida por un servidor web que proporcione acceso a las tecnologías semánticas a través de servicios web. Las siguientes líneas de trabajo estarán dedicadas a depurar el funcionamiento de los módulos en la plataforma semántica. La mayor parte de esta tarea de depuración se centrará en el proceso de validación exhaustiva del sistema, además de optimización y mejora de la presentación de resultados de las consultas semánticas.

Se pretende también refinar el motor de análisis semántico, para que contemple diferentes opciones de filtrado que permitan obtener resultados más precisos en función de las necesidades del usuario.

Bibliografía

Castells, P., Fernandez, M. y Vallet, D. 2007. An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering*, IEEE Transactions on, 19:261-272.

Chesbrough, H. W. 2003. Open Innovation: The new imperative for creating and profiting from technology. Boston: Harvard Business School Press. ISBN.

Rodriguez-García, M. A., Valencia-García, R., Garcia-Sánchez F. y Samper-Zapater, J. J. 2014. Ontology-based annotation and retrieval of services in the cloud, *Knowledge-Based Systems*, 56:15-25. January.

Salton G. y McGill, M. J. 1986. Introduction to modern information retrieval.

Singhal, A. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24(4):35-43.

[2] http://virtuoso.openlinksw.com

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 151-154 recibido 12-04-14 revisado 02-06-14 aceptado 02-06-14

ATTOS: Análisis de Tendencias y Temáticas a través de Opiniones y Sentimientos

ATTOS: Trend Analysis and Thematic through Opinions and Sentiments

L. Alfonso Ureña López*, Rafael Muñoz Guillena, José A. Troyano Jiménez***, Mª Teresa Martín Valdivia***

*SINAI - Universidad de Jaén. Campus Las Lagunillas s/n, 23071, Jaén
**GPLSI - Universidad de Alicante. San Vicente del Raspeig, s/n, 03690, Alicante
***ITALICA - Universidad de Sevilla. Av. Reina Mercedes, s/n. 41012 - Sevilla
laurena@ujaen.es, rafael@dlsi.ua.es, troyano@us.es, maite@ujaen.es

Resumen: El proyecto ATTOS centra su actividad en el estudio y desarrollo de técnicas de análisis de opiniones, enfocado a proporcionar toda la información necesaria para que una empresa o una institución pueda tomar decisiones estratégicas en función a la imagen que la sociedad tiene sobre esa empresa, producto o servicio. El objetivo último del proyecto es la interpretación automática de estas opiniones, posibilitando así su posterior explotación. Para ello se estudian parámetros tales como la intensidad de la opinión, ubicación geográfica y perfil de usuario, entre otros factores, para facilitar la toma de decisiones. El objetivo general del proyecto se centra en el estudio, desarrollo y experimentación de técnicas, recursos y sistemas basados en Tecnologías del Lenguaje Humano (TLH), para conformar una plataforma de monitorización de la Web 2.0 que genere información sobre tendencias de opinión relacionadas con un tema.
Palabras clave: Análisis de Opiniones y Sentimientos, Tecnologías del Lenguaje Humano, Recuperación de Información, Clasificación de Opiniones, Procesamiento de Lenguaje Natural

Abstract: The ATTOS project will be focused on the study and development of Sentiment Analysis techniques. Thanks to such techniques and resources, companies, but also institutions will be better understood which is the public opinion on them and thus will be able to develop their strategies according to their purposes. The final aim of the project is the automatic interpretation of such opinions according to different variables: opinion, intensity, geographical area, user profile, to support the decision process. The main objective of the project is the study, development and evaluation of techniques, resources and systems based on Human Language Technologies to build up a monitoring platform of the Web 2.0 that generates information on opinion trends related with a topic.
Keywords: Opinion and Sentiments Analysis, Human Language Technology, Information Retrieval, Opinion Classification, Natural Language Processing

1 Introducción

La interacción actual de los usuarios de la Sociedad de la Información es muy participativa. Los usuarios expresan sus puntos de vista, opiniones que llegan de forma inmediata al resto de usuarios a través de la Web 2.0 (foros, blogs, microblogs, redes sociales, etc.). Como consecuencia, la cantidad de información de carácter subjetivo y lenguaje informal se ha multiplicado en los últimos años, surgiendo así nuevos retos para las Tecnologías del Lenguaje Humano (TLH), como son el tratamiento de distintos registros de uso con diferentes grados de informalidad, el estudio de distintas actitudes subjetivas o el multilingüismo.

Las actuales herramientas para las TLH no son directamente aplicables a estos nuevos usos y medios de comunicación o son, simplemente, inadecuadas, por lo que se hace esencial adaptar

ISSN 1135-5948

y crear nuevos recursos, métodos y herramientas para su tratamiento.

El objetivo global del proyecto es, por tanto, el estudio, desarrollo y experimentación de recursos, diferentes técnicas y sistemas basados en TLH para el desarrollo y la comprensión de las expresiones subjetivas y el lenguaje informal en diversos dominios de aplicación, así como el desarrollo de una plataforma online de monitorización de diversos tipos de objetos de acuerdo a la información subjetiva e informal extraída de varias fuentes online de información. En este nuevo escenario, los sistemas deben incorporar recursos, herramientas y sistemas que descubrirán la subjetividad de la información en todos sus contextos (espacial, temporal y emocional) analizando la dimensión multilingüe, y su aplicación en diversos dominios.

2 Objetivos

Los objetivos generales del proyecto ATTOS son[1]:

- Creación, adaptación y mejora de recursos, técnicas y herramientas que modelan el lenguaje subjetivo e informal generado en diversas fuentes de información (blogs, microblogs, redes sociales, etc.). Tratamiento del lenguaje emocional, la multilingualidad y la aplicación a entornos concretos.
- Desarrollo de subsistemas inteligentes de procesamiento (recuperación, tratamiento, comprensión y descubrimiento) de la información adaptados a las nuevas formas de comunicación con capacidad de interpretar y valorar el contexto del mensaje.
- Integración de los recursos, herramientas y sistemas desarrollados para el análisis de la subjetividad en una plataforma web de monitorización, cuya validez se demostrará sobre varios escenarios de uso concreto de distintos ámbitos (turismo, política, empresarial, comercio electrónico, etc.). Evaluación de la plataforma. Promoción de las líneas de investigación del proyecto mediante la participación y organización de actividades en campañas, congresos, talleres, seminarios y redes temáticas

Para la consecución del objetivo global y el desarrollo óptimo de las diferentes líneas de actuación del proyecto, se propuso la coordinación de tres subproyectos complementarios cuyos objetivos específicos abarcan los objetivos globales planteados, y cuya reunificación proporcionará el valor añadido que se buscaba en la coordinación.

El subproyecto ATTOS -Análisis de Tendencias y Temáticas a través de Opiniones y Sentimientos- que lleva a cabo el equipo de investigación de la Universidad de Jaén, tiene como objetivo central la construcción de una plataforma de procesamiento inteligente que integre las técnicas desarrolladas por todos los equipos de este proyecto para la explotación de la información subjetiva y que sea fácilmente adaptable a diversos dominios de aplicación. Así el subproyecto SOTTA -Semantic Opinion Techniques for Tendencies Analysis- que lleva a cabo el equipo de la Universidad de Alicante, tiene como objetivo principal el desarrollo de una herramienta de análisis de tendencias en función a perfiles de usuarios, incorporando técnicas que permitan identificar y resolver la presencia de metáforas, ironía y sarcasmo en textos subjetivos. Y finalmente el subproyecto ACOGEUS -Análisis de COntenidos GEnerados por USuarios- a cargo del grupo de la Universidad de Sevilla, cuyo objetivo es la identificación de fuentes online con información subjetiva y recuperación de dicha información, creando recursos propios para los dominios a abordar, así como el desarrollo de técnicas que permitan identificar diversos registros del lenguaje (ofensivo, violento, etc.).

3 Propuesta

El objetivo general del proyecto se centra en el estudio, desarrollo y experimentación de diferentes técnicas y sistemas basados en Tecnologías del Lenguaje Humano (TLH) para el desarrollo de una plataforma de tratamiento de información subjetiva y lenguaje informal, afrontando los actuales retos de la comunicación digital. En este nuevo escenario, los sistemas deben incorporar capacidades de razonamiento que descubrirán la subjetividad de la información desde diversas dimensiones: multilingüe, espacial, temporal y emocional.

[1] Sitio web: http://attos.ujaen.es

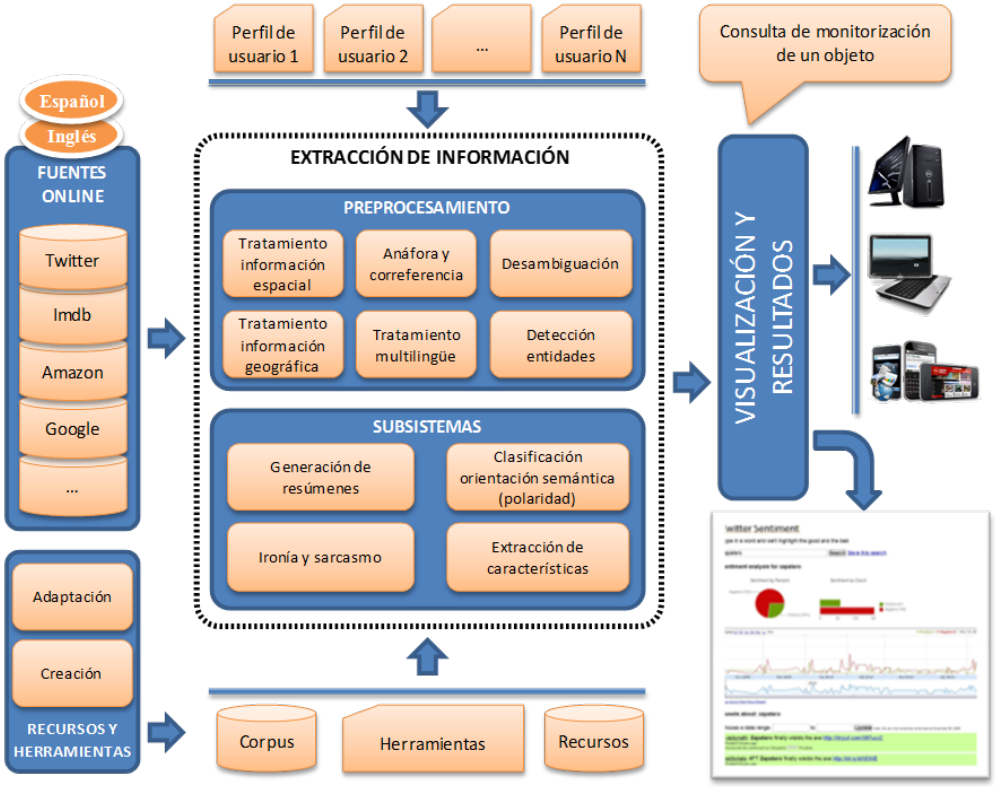


Figura 1: Arquitectura general del sistema

La figura 1 muestra la manera en la que se pueden integrar distintos componentes para construir un sistema capaz de procesar distintas fuentes online y extraer indicadores de utilidad mediante la aplicación de distintas tecnologías del lenguaje humano.

El diseño de los módulos del plan de trabajo propuesto se corresponde con las líneas de actuación marcadas en los objetivos del proyecto.

- En el módulo 1 se gestiona el proyecto y se diseñan mecanismos de coordinación que permitan una comunicación fluida y una colaboración eficiente entre los distintos miembros del proyecto.
- El módulo 2 se centra en el desarrollo y adaptación de recursos, herramientas y métodos de TLH para el modelado, análisis y tratamiento de información subjetiva e informal.
- En el módulo 3 se desarrollan los sistemas de detección y tratamiento de la información subjetiva y su tratamiento, su especialización en diversos dominios de aplicación y el desarrollo de una plataforma online de visualización y presentación de resultados.
- El módulo 4 contempla las actividades necesarias para la evaluación de la utilidad de la plataforma tanto interna como externa, así como la promoción, coordinación y participación en diferentes foros de evaluación.
- Finalmente, mediante el módulo 5, se creará un plan estratégico para diseminar los resultados tanto científicamente como mediáticamente para lograr la mayor difusión posible y facilitar la transferencia de tecnología a la empresa.

Respecto al enfoque científico, este proyecto supone un reto en el modo de abordar nuevos registros del lenguaje, como es la información digital subjetiva y el lenguaje informal. El problema actual es afrontar el tratamiento de una creciente cantidad de información en los nuevos registros que la Web 2.0 contiene: información textual en formatos muy variados expresada en muchas ocasiones de manera espontánea sin la precisión, formalidad ni corrección de los textos normativos. Desde la perspectiva computacional, requiere un replanteamiento de los métodos y técnicas de adquisición automática de conocimiento para tratar nuevas unidades y características, además de aquellas que son tradicionalmente aceptadas.

4 Resultados

En el tiempo en el que el proyecto lleva en ejecución, los trabajos realizados se han

materializado en diferentes contribuciones como publicaciones en revistas, congresos, organización de eventos o participación en evaluaciones competitivas. En esta sección comentaremos brevemente algunos de estos resultados que constituyen una muestra significativa de los avances que se están consiguiendo en el proyecto.

En al ámbito de las redes sociales, y en concreto en Twitter, en (Cotelo et al., 2014) se definió un método para obtener de forma automática consultas adaptativas a partir de un conjunto de *hashtags* semilla. Este método es especialmente interesante para poder capturar tweets relacionados con una temática, contemplando de forma automática los términos que puedan ir apareciendo en el transcurso de los diálogos colaborativos que permite esta red social.

En el contexto del análisis de opiniones, en (Cruz et al., 2013) se presenta un sistema de extracción adaptable a dominio que permite identificar opiniones en textos escritos por usuarios extrayendo la característica sobre la que se opina (p.e. el precio) y la valoración correspondiente (positiva o negativa). También se desarrolló un algoritmo de polaridad en 6 niveles mediante la modificación de un algoritmo de ranking (RA-SR) mediante la utilización de bigramas un puntuador de skipgrams (Fernández et al, 2013). En (Molina-González et al., 2013) se presenta una lista de palabras indicadoras de opinión en español de dominio general, así como una metodología para la adaptación de lexicones de palabras de opinión a un dominio concreto. También se han obtenido unos primeros resultados en la clasificación de la polaridad en redes sociales. En (Montejo-Ráez et al., 2013) se presenta un sistema de clasificación de la polaridad sobre *tweets*, cuya mayor aportación es el método de desambiguación utilizado, el cual utiliza información del contexto para mejorar la exactitud de la desambiguación, y la inclusión de términos relacionados para el cálculo de la polaridad del *tweet*.

En la detección de la subjetividad se desarrolló un método a nivel de oraciones basado en la desambiguación subjetiva del sentido de las palabras. Para ello se extiende un método de desambiguación semántica basado en agrupamiento de sentidos para determinar cuándo las palabras dentro de la oración están siendo utilizadas de forma subjetiva u objetiva (Ortega et al. 2013).

Agradecimientos

El proyecto ATTOS está financiado por el Ministerio de Economía y Competitividad con número de referencia TIN2012-38536-C03-01, TIN2012-38536-C03-02 y TIN2012-38536-C03-03. Con el apoyo de la Red Temática TIMM: Tratamiento de Información Multimodal y Multilingüe. (TIN2011-13070-E).

Bibliografía

Cotelo, J.M., Cruz, F.L. Troyano, J.A. 2014. Dynamic topic-related tweet retrieval. *JASIST*. 65(3): 513-523.

Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., Vallejo, C.G. 2013. 'Long autonomy or long delay?' The importance of domain in opinion mining. *Expert Syst. Appl*. 40(8): 3174-3184.

Fernández, J., Gómez, J.M.; Martínez, P., Montoyo, A, Muñoz, R. 2013 Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. TASS 2013: Taller de Análisis de Sentimientos en la SEPLN / Workshop on Sentiment Analysis at SEPLN Madrid, Spain, SEPLN.

Molina-González, M. Dolores, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Perea-Ortega, Jose M. 2013. Semantic Orientation for Polarity Classification in Spanish Reviews. *Expert Systems with Applications*. 40(18):7250-7257.

Montejo-Ráez, Arturo, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Ureña-López, L. Alfonso. 2014. A Knowledge-Based Approach for Polarity Classification in Twitter. *JASIST*. 65(2):414-425.

Ortega, R.; Fonseca, A.; Gutierrez, Y.; Montoyo, A.2013 Improving Subjectivity Detection using Unsupervised Subjectivity Word Sense Disambiguation. Revista Procesamiento del Lenguaje Natural, 51.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 155-158 recibido 15-04-14 revisado 15-07-14 aceptado 15-07-14

NewsReader project

Proyecto NewsReader

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Begoña Altuna
Zuhaitz Beloki, Egoitz Laparra, Maddalen López de Lacalle
German Rigau, Aitor Soroa, Ruben Urizar
IXA NLP Group, University of the Basque Country UPV/EHU
Manuel Lardizabal Pasealekua 1, 20018 Donostia
german.rigau@ehu.es

Resumen: El proyecto europeo NewsReader desarrolla tecnología avanzada para procesar flujos contínuos de noticias diarias en 4 idiomas, extrayendo lo que pasó, cuándo, dónde y quién estuvo involucrado. NewsReader lee grandes cantidades de noticias procedentes de miles de fuentes. Se comparan los resultados a través de las fuentes para complementar la información y determinar en qué están de acuerdo. Además, se fusionak noticias actuales con noticias previas, creando una historia a largo plazo en lugar de eventos separados. El resultado se acumula a lo largo del tiempo, produciendo una inmensa base de conocimiento que puede ser visualizada usando nuevas técnicas que permiten un acceso a la información más exhaustivo.
Palabras clave: Flujos de noticias, Extracción de eventos cross-lingual, historias

Abstract: The European project NewsReader develops advanced technology to process daily news streams in 4 languages, extracting *what* happened, *when* and *where* it happened and *who* was involved. NewsReader reads massive amounts of news coming from thousands of sources. It compares the results across sources to complement information and determine where the different sources disagree. Furthermore, it merges current news with previous news, creating a long-term history rather than separate events. The result is cumulated over time, producing an extremely large knowledge base that is visualized using new techniques to provide more comprehensive access.
Keywords: news streams, cross-lingual event extraction, history lines

1 Introduction

Professionals in any sector need to have access to accurate and complete knowledge to take well-informed decisions. Decision-makers are involved in a constant race to stay informed and to respond adequately to any changes, developments and news. However, the volume of news and documents provided by major information brokers has reached a level where state-of-the-art tools no longer provide a solution to these challenges.

The NewsReader project[1] (Vossen et al., 2014) analyzes news articles in 4 languages (English, Dutch, Italian and Spanish) to extract *what* happened, *where* and *when* it happened, and *who* was involved. NewsReader will reconstruct and visualize coherent storylines in which new events are related to past events. The system will not forget any detail, will keep track of all the facts and will even know when and how different sources told stories differently. The project will be tested on economic-financial news.

2 Objectives

One of the main goals of NewsReader is to extract events from multilingual news and to organize these events into coherent narrative storylines. The project will extract cross-lingual event information, participants taking part in the event, and additional time and location constraints. NewsReader will also detect the factuality of the events and their provenance. In addition, NewsReader will merge the news of today with previously stored information, creating a long-term his-

[1]NewsReader is funded by the European Union as project ICT- 316404. It is a collaboration of 3 European research groups and 3 companies: LexisNexis, ScraperWiki and Synerscope. The project started on January 2013 and will last 3 years. For more information see: `http://www.newsreader-project.eu/`

ISSN 1135-5948

tory rather than storing separate events. The final output will be stored in the KnowledgeStore that supports formal reasoning and inferencing.

The project foresees an estimating flow of 2 million news items per day and the complex linguistic analysis of those documents needs to be done in a reasonable time frame. The project faces thus an important challenge also regarding the scalability of the linguistic text processing.

In the same way, the amount of data produced in NewsReader is extremely large and complex. The content of the KnowledgeStore has to be offered to professional decision-makers in an effective way. NewsReader will develop innovative visualization techniques for events, their internal structure and their relations to other events that will graphically and adequately display the content of the KnowledgeStore. The visualizations of these storylines are expected to be more efficient and provide a more natural summarization of the changing world with more explanatory power.

3 Work Plan

The research activities conducted within the NewsReader project strongly rely on the cross-lingual detection of events, which are considered as the core information unit underlying news. The research focuses on four challenging aspects: event detection (addressed in WP04), event processing (addressed in WP05), storage and reasoning over events (addressed in WP06), and scaling to large textual streams (addressed in WP2). IXA group[2] is leading both WP2 and WP4.

The overall approach for processing data follows a sequence of steps, covered by the different work packages. The industrial partners define and collect relevant data sources, which are used as input by the system. The textual sources defined in WP01 (User Requirements) come in various formats.

The pieces of news are first processed through a language processing pipeline to detect event mentions, their participants and their location and time. This processing is document-based and the results are stored in the Natural Language Processing format (NAF, (Fokkens et al., 2014)). NAF is a sequel of the KYOTO Annotation Framework (KAF, (Bosma et al., 2009)) and is compliant with the Linguistic Annotation Format, LAF (Ide, Romary, and Éric Villemonte de La Clergerie, 2003).

Next, the event mentions within and across documents are compared to decide whether they refer to the same event. To represent these instances we use the Simple Event Model, SEM (Van Hage et al., 2011), which is an RDF-compliant model for representing events. Coreference can be applied to entities and to events and it can involve mentions within the same document and across documents. If different event mentions refer to the same event, duplication, complementing information and inconsistencies have to be detected. These comprise participants, place and time relations. If they make reference to different events, it is also necessary to determine the relation between them such as temporal or causal relations.

The final output, represented as NAF and SEM, is stored in the KnowledgeStore. The KnowledgeStore has different components for different types of data. It allows to store in its three interconnected layers all the typologies of content that have to be processed and produced when dealing with unstructured content and structured knowledge. The KnowledgeStore acts as a "history-recorder" which keeps track of the changes in the world as told in the media. It represents the information in RDF and supports reasoning over the data.

The next sections will describe the event detection and scalability tasks in more detail.

4 Event Detection

NewsReader uses an open and modular architecture for event detection. The system uses NAF as the layered annotation format, and separate modules have been developed to add new interpretation layers using the output of previous layers. Text-processing requires basic and generic NLP steps such as tokenization, lemmatization, part-of-speech tagging, parsing, word sense disambiguation, named-entity and semantic role recognition, etc. for all the languages within the project. Named entities are linked as much as possible to external sources such as DBpedia entity identifiers. We are also developing new techniques and resources to achieve interoperable semantic interpretation for English, Dutch, Spanish and Italian thanks to

[2] http://ixa.si.ehu.es/Ixa

the Predicate Matrix (López de Lacalle, Laparra, and Rigau, 2014).

Semantic interpretation involves the detection of event mentions and those named entities that play a role in these events, including time and location relations. This implies covering all expressions and meanings that can refer to events, their participating named entities, place and time relations. It also means to resolve coreference relations for these named entities and relations between different event mentions. As a result of this process, the text is enriched with semantic concepts and identifiers that can be used to access lexical resources and ontologies. For each unique event, we will also derive its factuality score based on the textual properties and its provenance.

NewsReader provides an abstraction layer for large-scale distributed computations, separating the *what* from the *how* of computation and isolating NLP developers from the details of concurrent programming. Section 4.1 explains the modules developed to perform event detection. Section 4.2 presents the implemented scaling infrastructure for advanced NLP processing.

4.1 NLP pipeline

We have defined a linguistic processing pipeline to automatically detect and model events. The NLP pipeline consists of basic and generic NLP processing steps, such as tokenization, lemmatization, part-of-speech tagging, word sense disambiguation and named-entity recognition. It also includes more sophisticated modules that deal with nominal coreference, nominal and verbal semantic role recognition, time recognition and interpretation, opinion detection, factuality detection, event classification and provenance identification.

Each task is executed by one independent module, which allows custom pipelines for text processing. We have developed a set of NLP tools which we refer to as the IXA pipeline (Agerri, Bermudez, and Rigau, 2014) for English and Spanish. The IXA pipeline currently provides the following linguistic annotations: Sentence segmentation, Tokenization, Part of Speech (POS) tagging, Lemmatization, Named Entity Recognition and Classification (NER), Syntactic Parsing and Coreference Resolution. This basic pipeline has been enhanced by adding new modules for word sense disambiguation, named-entity disambiguation, semantic role labeling, recognition of temporal expressions, factuality recognition, opinion mining and event coreference resolution.

The interoperability among the modules is achieved by using NAF as a common format for representing linguistic information. All the modules of the pipeline are adapted to read and write NAF, adding new layers to the NAF representation. The output can be streamed to the next module or it can be stored in the KnowledgeStore.

4.2 Scalability

The processing of news and documents provided by LexisNexis (one of the industrial partners of NewsReader and a large international news broker), has become a major challenge in the project. We have thus defined a new distributed architecture and technology for scaling up text analysis to keep pace with the rate of the current growth of news streams and collections. Scalable NLP processing requires parallel processing of textual data. The parallelization can be effectively performed at several levels, from deploying copies of the same LP among servers to the reimplementation of the core algorithms of each module using multi-threading, parallel computing. This last type of fine-grained parallelization is clearly out of the scope of the present work, as it is unreasonable to expect it to reimplement all the modules needed to perform such a complex task as mining events. We rather aim to process huge amounts of textual data by defining and implementing an architecture for NLP which allows the parallel processing of documents.

With this aim, we have created one Virtual Machine (VM) per language and pipeline so that a full processing chain in one language can be run on a single VM. This approach (Artola, Beloki, and Soroa, 2014) allows us to scale horizontally (or scale out) as a solution to the problem of dealing with massive quantities of data. We thus scale out our solution for NLP by deploying all the NLP modules into VMs and making as many copies of the VMs as necessary to process an initial batch of documents on time.

The modules are managed using the Storm framework for streaming computing[3]. Storm is an open source, general-

[3] http://storm.incubator.apache.org/

purpose, distributed, scalable and partially fault-tolerant platform for developing and running distributed programs that process continuous streams of data. Storm allows to set scalable clusters with high availability using commodity hardware and minimizes latency by supporting local memory reads and avoiding disk I/O bottlenecks.

Inside the VMs, each LP module is wrapped as a node inside the Storm topology. When a new document arrives, the processing node calls an external command sending the document to the standard input stream. The output of the LP module is received from the standard output stream and passed to the next node in the topology. Each module thus receives a NAF document with the (partially annotated) document and adds new annotations onto it. The tuples in our Storm topology comprise two elements, a document identifier and the document itself, encoded as a string with the XML serialization of the NAF document.

This setting has allowed the project to process more than 100.000 documents from the financial and economic domains using 8 copies of the VMs distributed among the project partners. As a result from the linguistic processing, more than 3 million events have been extracted.

5 Concluding Remarks

In this paper, we outlined the main objectives and methodology of the NewsReader project. We designed and implemented a complex platform for processing large volumes of news in different languages and storing the result in a KnowledgeStore that supports the dynamic growth and reasoning over data. The project shows that it is possible to develop reasoning technologies on top of the data that is generated from raw text.

Acknowledgments

This work has been supported by the EC within the 7th framework programme under grant agreement nr. FP7-IST-316040.

References

Agerri, Rodrigo, Josu Bermudez, and German Rigau. 2014. IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*

Artola, Xabier, Zuhaitz Beloki, and Aitor Soroa. 2014. A stream computing approach towards scalable NLP. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014).*

Bosma, Wauter, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation.*

Fokkens, Antske, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *To appear in Proceedings of 10th Joint ACL/ISO Workshop on Interoperable Semantic Annotation (ISA-10).*

Ide, Nancy, Laurent Romary, and Éric Villemonte de La Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS).* Association for Computational Linguistics.

López de Lacalle, Maddalen, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*

Van Hage, W.R., V. Malaisé, G.K.D. De Vries, G. Schreiber, and M.W. van Someren. 2011. Abstracting and reasoning over ship trajectories and web data with the simple event model (SEM). *Multimedia Tools and Applications*, pages 1–23.

Vossen, Piek, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving, and Willem Van Hage. 2014. Newsreader: recording history from daily news streams. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 159-162 recibido 15-04-14 revisado 30-06-14 aceptado 02-07-14

Análisis Semántico de la Opinión de los Ciudadanos en Redes Sociales en la Ciudad del Futuro

Opinion Mining in Social Networks using Semantic Analytics in the City of the Future

Julio Villena-Román
Adrián Luna-Cobos
Daedalus, S.A.
Av. de la Albufera 321
28031 Madrid, España
{jvillena, aluna}@daedalus.es

José Carlos González-Cristóbal
Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, España
jgonzalez@dit.upm.es

Resumen: En este artículo se presenta un sistema automático de almacenamiento, análisis y visualización de información semántica extraída de mensajes de Twitter, diseñado para proporcionar a las administraciones públicas una herramienta para analizar de una manera sencilla y rápida los patrones de comportamiento de los ciudadanos, su opinión acerca de los servicios públicos, la percepción de la ciudad, los eventos de interés, etc. Además, puede usarse como sistema de alerta temprana, mejorando la rapidez de actuación de los servicios de emergencia.
Palabras clave: Análisis semántico, redes sociales, ciudadano, opinión, temática, clasificación, ontología, eventos, alertas, big data, consola de la ciudad.

Abstract: In this paper, a real-time analysis system to automatically record, analyze and visualize high level aggregated information of Twitter messages is described, designed to provide public administrations with a powerful tool to easily understand what the citizen behaviour trends are, their opinion about city services, their perception of the city, events of interest, etc. Moreover, it can used as a primary alert system to improve emergency services.
Keywords: Semantic analytics, social networks, citizen, opinion, topics, classification, ontology, events, alerts, big data, city console.

1 Introducción

El objetivo final de las decisiones de las administraciones públicas es el bienestar ciudadano. Sin embargo, no siempre es fácil para los gestores identificar rápidamente los asuntos más importantes que afrontan sus ciudadanos y priorizarlos según la importancia real que los propios ciudadanos les asignan. El ciudadano se trata desde un punto de vista dual: como el principal usuario de los servicios que presta la ciudad, pero también, como un sensor proactivo, capaz de generar grandes cantidades de datos, por ejemplo en redes sociales, con información útil de su grado de satisfacción sobre su entorno. Por ello, el análisis de la opinión ciudadana es un factor clave dentro de la ciudad del futuro para identificar los problemas de los ciudadanos. Sin embargo, toda esta información no es realmente útil a no ser que sea automáticamente procesada y anotada semánticamente para distinguir la información relevante y lograr un mayor nivel de abstracción, y es aquí donde las tecnologías lingüísticas juegan un papel clave.

Este artículo presenta un sistema para el análisis en tiempo real de información en Twitter (no se aborda expresamente ninguna otra fuente de información). El sistema permite recopilar y almacenar los mensajes, analizarlos semánticamente y visualizar información

[1] Este trabajo ha sido financiado parcialmente por el proyecto Ciudad2020: *Hacia un nuevo modelo de ciudad inteligente sostenible* (INNPRONTA IPT-20111006), cuyo objetivo es el diseño de la Ciudad del Futuro, persiguiendo mejoras en áreas como la eficiencia energética, sostenibilidad medioambiental, movilidad y transporte, comportamiento humano e Internet de las cosas.

agregada de alto nivel. Aunque existen diversos trabajos que tratan el análisis semántico en redes sociales (TwitterSentiment, Twendz, SocialMention, etc.), no se conoce la existencia de un sistema que integre un análisis semántico completo y con capacidades de tiempo real, almacenamiento y capacidad de agregación estadística orientado a las ciudades inteligentes. Destaca un trabajo en esta línea (C. Musto et. al., 2014) pero centrado en análisis de cohesión social y sentido de pertenencia a la comunidad.

El objetivo último es proporcionar a los administradores públicos una herramienta potente para entender de una manera rápida y eficiente las tendencias de comportamiento, la opinión acerca de los servicios que ofrecen, eventos que tengan lugar en su ciudad, etc. y, además, proveer de un sistema de alerta temprana que consiga mejorar la eficiencia de los servicios de emergencia.

2 *Arquitectura del Sistema*

El sistema está formado por cuatro bloques principales, mostrados en la Figura 1.

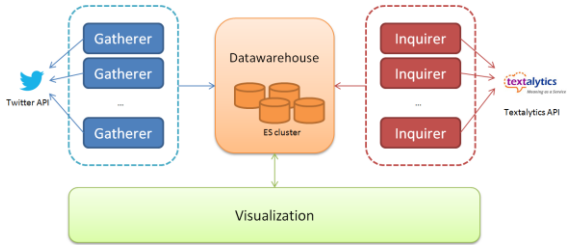


Figura 1: Arquitectura del sistema

El componente central es el ***datawarehouse***, el repositorio de información principal, capaz de almacenar el gran volumen de datos a los que hace frente el sistema además de proporcionar funcionalidad avanzada de búsqueda. Este componente se basa en Elasticsearch (Elasticsearch, 2014), motor de búsqueda en tiempo real, flexible y potente, de código abierto y distribuido. Su buena escalabilidad en escenarios con gran cantidad de datos fue el factor decisivo en la selección de esta tecnología.

El segundo componente lo forman un conjunto de procesos ***recolectores*** que implementan el acceso a los documentos vía consultas a las API de Twitter. Estos recolectores pueden ser configurados para filtrar tweets según una lista de identificadores de usuario, listas de palabras clave a seguir cómo términos o hashtags y localizaciones geográficas a las que restringir la búsqueda.

Un tercer componente, formado por un conjunto de procesos ***consumidores***, tiene como tarea anotar los mensajes de Twitter utilizando las APIs de Textalytics[2].

Se han diseñado dos modelos de clasificación temática (usando la API de clasificación de textos) específicos para este proyecto: *SocialMedia* y *CitizenSensor*, descritos más adelante. También se utiliza la API de extracción de *topics* para anotar entidades nombradas, conceptos, expresiones monetarias, URI, etc. Con la API de análisis de sentimiento se extrae la polaridad del mensaje, así como indicaciones acerca de su subjetividad o si expresa ironía. Por último, se utiliza *user demographics* para obtener información del tipo, sexo y edad del autor del tweet.

El proceso más pesado computacionalmente es la anotación semántica del texto y por lo tanto constituye el cuello de botella del sistema. Sin embargo, los procesos consumidores anotan los mensajes que aún no han sido procesados en orden descendente respecto a la fecha de indexación, de tal manera que la información más reciente siempre es la que está disponible primero. Esta característica es clave para poder reaccionar de forma temprana a alertas. Si el ratio de entrada de mensajes que los recolectores indexan en el sistema es mayor que lo que los consumidores son capaces de anotar, no será posible acceder a toda la información semántica de los mensajes en tiempo real, pero una vez que esta situación se revierta y el sistema consiga anotar a una velocidad mayor que los nuevos documentos, seguirá anotando los que quedaron sin procesar.

Por último, se ha definido un ***sistema de visualización*** para explotar los datos generados.

3 *Etiquetado semántico*

Se ha invertido un gran esfuerzo en la tarea de etiquetado semántico para este escenario particular: fragmentos cortos de texto, con capitalización inadecuada, faltas de ortografía, emoticonos, abreviaturas, etc. Los procesos consumidores proporcionan múltiples niveles de análisis según se describe a continuación.

En este primer despliegue se analizan exclusivamente tweets en español. Como las herramientas de procesamiento lingüístico

[2] http://textalytics.com

utilizadas en el sistema están ya disponibles para otros idiomas (inglés, francés, italiano, portugués y catalán), el sistema podría ser fácilmente extendido en estos aspectos. Sin embargo, sí habría que llevar un trabajo específico para migrar las ontologías de clasificación automática desarrolladas específicamente para este proyecto.

3.1 Clasificación automática

El algoritmo de clasificación de texto utilizado combina una clasificación estadística con un filtrado basado en reglas, que permite obtener un nivel de precisión bastante altos. Por ejemplo, evaluando el corpus Reuters-21578, se obtienen precisiones de más del 80% (Villena-Román et al., 2011). En concreto, se han diseñado dos ontologías específicas para este caso de uso.

El modelo de *SocialMedia* define los temas generales de clasificación, que ha sido desarrollado favoreciendo su precisión cuando se evalúan textos que proceden de redes sociales, respecto a los modelos generales ya disponibles y que se han usado satisfactoriamente en otros ámbitos.

Por otro lado, *CitizenSensor* se orienta a características propias del ciudadano como sensor de eventos de la ciudad, tratando de clasificar aspectos tales como su ubicación, eventos que ocurren en la ciudad o posibles catástrofes o alertas. Estos modelos se han desarrollado en base a reglas dónde se definen términos (o patrones) obligatorios, prohibidos, relevantes e irrelevantes para cada categoría.

3.2 Extracción de entidades

Este proceso se lleva a cabo combinando varias técnicas de procesamiento de lenguaje natural para obtener análisis morfosintáctico y semántico del texto y a través de estas características, identifican distintos tipos de elementos significativos.

Actualmente el sistema identifica (con flexión, variantes y sinónimos) distintos tipos de elementos: entidades nombradas (personas, organizaciones, lugares, etc.), conceptos (palabras clave relevantes para el texto tratado), expresiones temporales, expresiones monetarias y URIs.

Este análisis se apoya en recursos lingüísticos propios y reglas heurísticas. Evaluaciones internas sitúan al sistema en niveles de precisión y cobertura similares a otros sistemas de extracción de entidades.

3.3 Análisis de sentimiento

El análisis de sentimiento se realiza en otro nivel de análisis semántico, para determinar si el texto expresa un sentimiento positivo, neutral o negativo.

Este análisis se compone de varios procesos (Villena-Román et al., 2012): primero se evalúa el sentimiento local de cada frase y posteriormente se identifica la relación entre las distintas frases dando lugar a un sentimiento global. Además, empleando el análisis morfosintáctico, se detecta también la polaridad a nivel de a entidades y conceptos. El sistema ha sido evaluado en diversos foros obteniendo valores de medida-F superiores a 40%.

3.4 Características demográficas

Este módulo de análisis extrae características demográficas relativas al usuario que ha generado el texto analizado. Utilizando técnicas de extracción de información y algoritmos de clasificación, se estiman parámetros tales como el tipo de usuario (persona u organización), el sexo del usuario (hombre, mujer o desconocido) y su rango de edad (<15, 15-25, 25-35, 35-65 y >65 años).

Para realizar esta estimación, se utiliza la información del usuario en Twitter, el nombre asociado a su cuenta y la descripción de su perfil. El modelo se basa en n-gramas y ha sido desarrollado utilizando Weka.

3.5 Ejemplo de etiquetado

La Figura 2 muestra un mensaje anotado por el sistema, con salida JSON. Se pueden observar las distintas categorías asignadas para el modelo *CitizenSensor* (etiqueta "sensor"). El sistema identifica la ubicación del usuario (una vía pública), además de dos posibles alertas: aviso meteorológico por viento e incidencia por congestión de tráfico. Además, según el modelo *SocialMedia* (etiqueta "topic") se clasifica el mensaje dentro de la categoría de "medio ambiente". Posteriormente se muestran las entidades y conceptos detectados en el texto ("Gran Vía" y "viento", respectivamente), el análisis de sentimiento (se trata de un mensaje objetivo, no irónico y con polaridad negativa) y el análisis del usuario (el autor es una mujer de edad entre los 25 y los 35 años).

```
{
 "text":"el viento ha roto una rama y hay un
atascazo increible en toda la gran vía...",
 "tag_list":[
   {"type":"sensor", "value":"011002
Ubicación - Exteriores - Vías públicas"},
   {"type":"sensor", "value":"070700 Alertas
meteorológicas - Viento"},
   {"type":"sensor", "value":"080100
Incidencia - Congestión de tráfico"},
   {"type":"topic", "value":"06 medio
ambiente, meteorología y energía"},
   {"type":"entity", "value":"Gran Vía"},
   {"type":"concept", "value":"viento"},
   {"type":"sentiment", "value":"N"},
   {"type":"subjectivity", "value":"OBJ"},
   {"type":"irony", "value":"NONIRONIC"},
   {"type":"user_type", "value":"PERSON"},
   {"type":"user_gender", "value":"FEMALE"},
   {"type":"user_age", "value":"25-35"}
  ]
}
```


Figura 2: Ejemplo de anotación del sistema

4 *Módulo de visualización*

El módulo de visualización ofrece una interfaz web, que permite ejecutar consultas complejas de manera estructurada y presenta información de alto nivel, agregada y resumida.

La consola se define mediante elementos denominados *widgets*, configurados en una plantilla específica para los diferentes casos de uso del sistema y adaptada a cada necesidad. Para el desarrollo de los diferentes elementos se han utilizado librerías JavaScript existentes para la creación de gráficos[3], para la representación de mapas[4], y componentes propios.

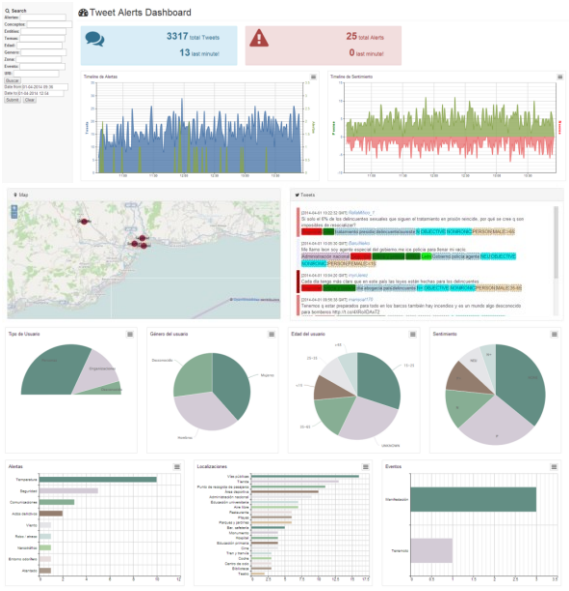


Figura 3: Consola de visualización

[3] http://www.highcharts.com
[4] http://openlayers.org

5 *Conclusiones y trabajos futuros*

Actualmente el sistema está en fase beta, acabando la puesta a punto de los diferentes módulos, y estará listo para ser desplegado en distintos escenarios a corto plazo. Las evaluaciones (informales) preliminares de la precisión de los diferentes módulos muestran que los resultados son totalmente válidos para cumplir con los objetivos de diseño del sistema.

Analizando el aspecto tecnológico, las capacidades de almacenamiento del sistema permiten, no sólo analizar los datos en tiempo real, sino también permiten aplicar algoritmos de minería de datos sobre los datos almacenados para, de esta manera, entender mejor las particularidades de la población, mediante técnicas de perfilado y *clustering* para identificar distintos grupos de ciudadanos que se encuentran en la ciudad, comparar singularidades entre los grupos detectados, etc.

Además se está investigando para explorar en el análisis de movilidad en la ciudad (cómo, cuándo y por qué los ciudadanos se mueven de un lugar a otro), la detección de los temas más relevantes a nivel de barrio o zona, y realizar un análisis de reputación o personalidad de marca.

Bibliografía

Musto, C., G. Semeraro, P. Lops, M. Gemmis, F. Narducci, L. Bordoni, M. Annunziato, C. Meloni, F.F. Orsucci, G. Paoloni. 2014. Developing a Semantic Content Analyzer for L'Aquila Social Urban Network. *In Proceedings of the 5th Italian Information Retrieval Workshop (IIR), Rome, Italy*.

Elasticsearch.org. Open Source Distributed Real Time Search & Analytics. 2014. *http://www.elasticsearch.org*

Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press.*

Villena-Román, J., S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J.C. González-Cristóbal. 2012. DAEDALUS at *RepLab 2012: Polarity Classification and Filtering on Twitter Data. CLEF 2012 Labs and Workshop Notebook Papers.*

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 163-166 recibido 15-04-14 revisado 02-06-14 aceptado 10-06-14

TrendMiner: Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams[1]

TrendMiner: Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams

Paloma Martínez, Isabel Segura
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es,
isegura@inf.uc3m.es

Thierry Declerck
Language
Technology Lab
DFKI
Saarbrucken, Germany
thierry.declerck@dfki.de

José L. Martínez
DAEDALUS- DATA,
DECISION and
LANGUAGE S.A.
jmartinez@daedalus.es

Resumen: El reciente crecimiento masivo de medios on-line y el incremento de los contenidos generados por los usuarios (por ejemplo, weblogs, Twitter, Facebook) plantea retos en el acceso e interpretación de datos multilingües de manera eficiente, rápida y asequible. El objetivo del proyecto TredMiner es desarrollar métodos innovadores, portables, de código abierto y que funcionen en tiempo real para generación de resúmenes y minería cross-lingüe de medios sociales a gran escala. Los resultados se están validando en tres casos de uso: soporte a la decisión en el dominio financiero (con analistas, empresarios, reguladores y economistas), monitorización y análisis político (con periodistas, economistas y políticos) y monitorización de medios sociales sobre salud con el fin de detectar información sobre efectos adversos a medicamentos.
Palabras clave: tecnologías del lenguaje en medios sociales, salud y finanzas, generación automática de resúmenes.

Abstract: The recent massive growth in online media and the rise of user-authored content (e.g weblogs, Twitter, Facebook) has led to challenges of how to access and interpret the strongly multilingual data, in a timely, efficient, and affordable manner. The goal of this project is to deliver innovative, portable open-source real-time methods for cross-lingual mining and summarization of large-scale stream media. Results are validated in three high-profile case studies: financial decision support (with analysts, traders, regulators, and economists), political analysis and monitoring (with politicians, economists, and political journalists) and monitoring patient postings in the health domain to detect adverse drug reactions.
Keywords: language technologies in health social media, financial analysis in social media, summarization, social media streams

1 Descripción General

TrendMiner (http://www.trendminer-project.eu/) es un proyecto europeo dedicado al análisis de medios sociales en distintos idiomas que comenzó en el año 2012 con los socios DFKI (Alemania) coordinador del proyecto, Universidad de Sheffield (UK), Ontotext AD, (Bulgaria), Universidad de Southampton (UK), Internet Memory Research (Francia), Eurokleis (Italy), Sora Ogris & Hofinger (Austria). En el 2014 se ha ampliado el consorcio con los siguientes socios: Grupo LaBDA de la Universidad Carlos III de Madrid (España), Nyelvtudomanyi Intezet, Magyar Tudomanyos Akademia (Hungría), Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polonia) y la empresa DAEDALUS-DATA, DECISIONS AND LANGUAGE, S.A. (España).

[1] FP7-ICT287863

2 Objetivos

El proyecto Trendminer se plantea los siguientes retos científicos:

Modelado de conocimiento y extracción de información multilingüe basada en ontologías. Se trata de desarrollar métodos de extracción de información novedosos que sean capaces de analizar streams de medios sociales caracterizados por ser cortos, ruidosos, coloquiales y contextualizados. El objetivo es identificar tendencias y sentimiento en múltiples lenguajes así como de extraer entidades y eventos relevantes y almacenarlos en una base de conocimiento. Los participantes DFKI y Universidad de Sheffield trabajan en esta línea generando ontologías ampliadas con elementos de opinión que proporcionan el conocimiento extralingüístico necesario en los métodos de extracción.

Es precisamente en esta línea donde el trabajo del Grupo LaBDA está desarrollando un recurso ontológico para representar información sobre fármacos así como sus indicaciones y efectos adversos que se incorporará al repositorio de ontologías del proyecto. Esta ontología se probará en tareas de extracción de información en blogs relacionados con salud donde los pacientes informan sobre sus tratamientos y problemas con su medicación (como por ejemplo www.forumclinic.org y www.enfemenimo.com)

Modelos basados en aprendizaje automático para minería de tendencias en medios sociales. Se trabaja en desarrollar enfoques de aprendizaje automático para la identificación de mensajes o posts importantes y para la extracción de fragmentos de texto a partir de grandes volúmenes de texto generados por los usuarios en medios sociales como Twitter. No existen datos de entrenamiento para el aprendizaje supervisado y crearlos consume muchos recursos. Por ello, se investiga en diversas formas para hacer uso de datos en forma de movimiento de precios del mercado y resultados de encuestas para su utilización en los casos de uso a modo de supervisión ligera para inferir la importancia de distintas características de los textos.

En esta línea se han llevado a cabo trabajos para la predicción de voto a partir de Twitter y la predicción de la tasa de desempleo también a partir de Twitter, ver los trabajos (Lampos, Preotiuc-Pietro & Cohn, 2013).

Generación de resúmenes cross-lingüe a partir de medios sociales. El objetivo es definir nuevos enfoques para la generación automática de resúmenes en una línea temporal para observar la evolución de los eventos. En esta línea la Universidad de Sheffield ha desarrollado trabajos como los presentados en (Rout, et al., 2013)

Plataforma para la recolección, análisis y almacenamiento de colecciones de medios sociales en tiempo real. Desarrollada por Ontotext. El objetivo es facilitar la recogida y la minería de conocimiento a partir de medios sociales (tales como Twitter, Facebook, blogs sobre salud y periódicos, etc.). Como principales retos están el procesamiento en tiempo real de grandes volúmenes de posts, la agregación y el almacenamiento así como el procesamiento distribuido y basado en la nube según arquitecturas elásticas para minería de textos. Esto hace posible que las empresas puedan utilizar los datos sin un coste prohibitivo y sin necesidad de invertir en infraestructuras privadas o centros de datos.

Minería de tendencias multilingüe y generación de resúmenes en el caso de uso de ayuda a la toma de decisiones en finanzas. El objetivo es poner en práctica en un desarrollo real a través de la plataforma del socio Ontotext las técnicas desarrolladas en el proyecto en el dominio financiero proporcionando métodos en tiempo real para inversores, analistas, consejeros, agentes de bolsa, en particular para analizar la influencia en los precios en el mercado.

Extensiones de lenguajes y dominios: Con la entrada de nuevos socios en el proyecto se buscaba ampliar los idiomas y los dominios de trabajo. En el caso del español se ha propuesto incrementar la cobertura de las ontologías y las herramientas en el dominio de la salud, en particular, para los fármacos, efectos terapéuticos y reacciones adversas. El objetivo es aplicar las técnicas utilizadas para la extracción de información en textos científicos biomédicos a la extracción en posts y contribuciones en medios sociales tales como *Forumclinic* y *Portales Médicos*. Sin embargo ninguno de estos blogs y foros es tan sofisticado como *Patientslikeme* (http://www.patientslikeme.com/), una plataforma on-line que integra datos de pacientes en inglés.

La figura 1 muestra la arquitectura general del proyecto.

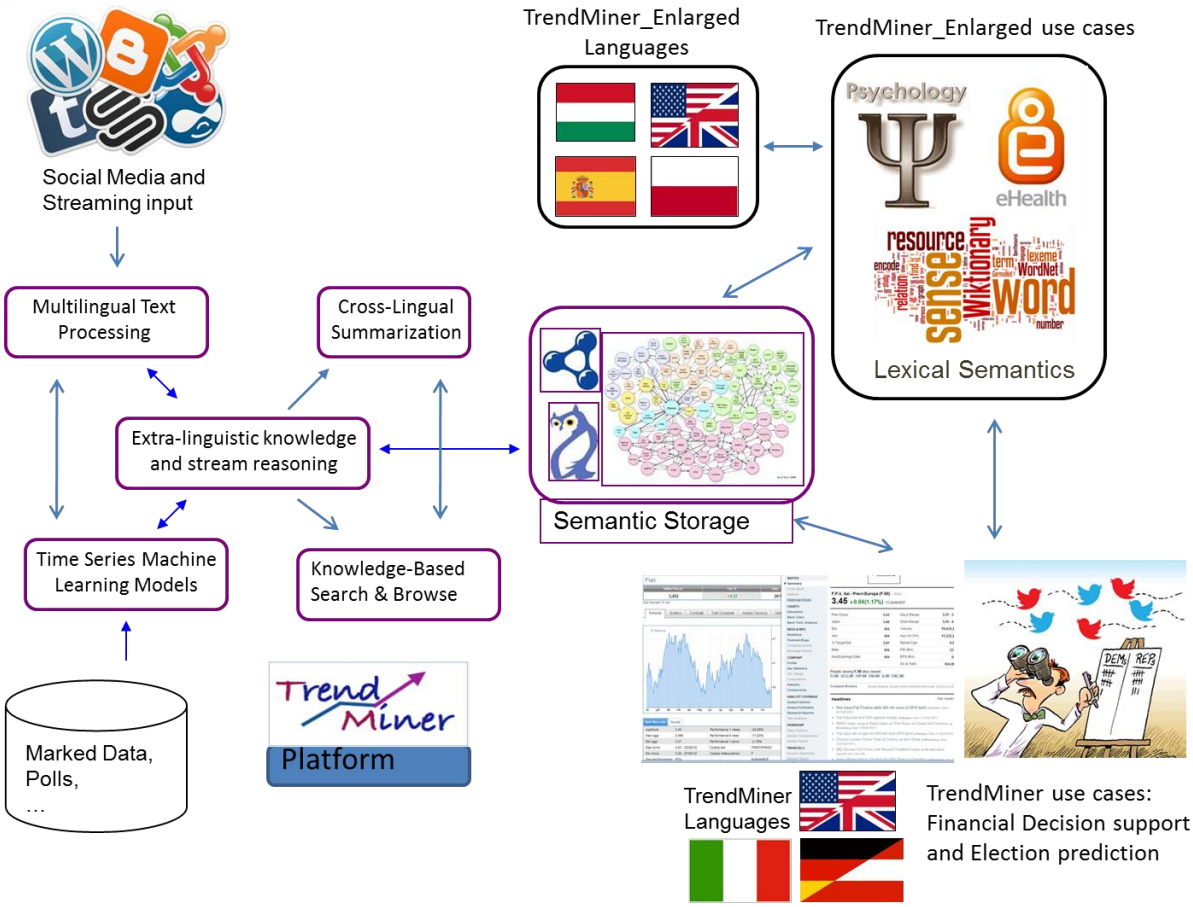


Figura 1: Arquitectura general del proyecto TrendMiner

3 *Situación actual*

En relación con el trabajo de los socios españoles, en la actualidad se trabaja en el dominio de salud en el análisis de consumo de antidepresivos en relación con distintos parámetros como la legislación actual en materia de trabajo, la tasa de desempleo, etc. Se están recogiendo tweets relacionados con diversos eventos usando diversas keywords relacionadas con fármacos antidepresivos, estados emocionales y términos relacionados con la economía con el fin de relacionarlos con el consumo de fármacos

Por otro lado, también se trabaja en la detección de efectos adversos de fármacos en medios sociales usando tecnología del lenguaje para analizar el contenido de los posts (reconocimiento de entidades y relaciones semánticas), (Segura-Bedmar, Revert y Martínez, 2014). Se trata de analizar si los pacientes reportan sobre estos efectos en blogs y foros. El objetivo es estudiar si estos medios pueden ser una fuente de conocimiento adicional a los sistemas de notificación que tienen las agencias europeas de medicamentos y productos sanitarios para que los pacientes y personal sanitario informen sobre sospechas de reacciones adversas que tienen poco uso. En (Segura-Bedmar, Peña-González y Martínez, 2014) se describe un recurso desarrollado para almacenar los fármacos y sus efectos (indicaciones y efectos adversos) relacionados que se integra en el prototipo bajo desarrollo. Hasta la fecha se disponía de información de fármacos y efectos pero de manera aislada.

En esta línea son varias las dificultades para extraer las menciones de fármacos y efectos adversos a partir de los comentarios de los usuarios. Además de abordar los problemas específicos de los medios sociales (como son el uso de abreviaturas, slangs, emoticonos, faltas ortográficas, etc.) son necesarios recursos que no existen en la actualidad. Por ejemplo, no se dispone en español de un diccionario orientado a pacientes como el que existe en inglés,

llamado Consumer Health Vocabulary[2]. Los pacientes no se refieren a sus problemas usando la terminología propia de los profesionales sanitarios

Por último también es interesante analizar la evolución de nuevos fármacos en el mercado analizando cómo evolucionan los comentarios de los pacientes considerando también las opiniones que manifiestan al respecto. Esto podría incluso servir de ayuda a la detección de nuevos efectos que no han sido descubiertos en los ensayos clínicos, una tarea importante en fármaco vigilancia. En el caso español, la Agencia Española de Medicamentos publica anualmente una lista con los fármacos de reciente aprobación a los que hay que hacer un especial seguimiento.

En la actualidad se dispone de un prototipo que analiza comentarios de pacientes en español disponible en http://163.117.129.57:8090/gate/ basado en un pipeline de procesos construido sobre GATE[3] que incorpora una API del procesador lingüístico Textalytics[4], software de DAEDALUS. Para el análisis de opinión se trabajará también con Sentimentalytics[5] del mismo socio.

En relación con el escenario financiero DAEDALUS está trabajando en el caso de uso de Responsabilidad Social Corporativa (RSC), para ayudar a las empresas a obtener una visión clara sobre su reputación on-line entre la gente. Para este propósito se ha trabajado en un prototipo que recoge tweets y sitios de noticias on-line accedidos a través de RSS para empresas del IBEX35. Esta información se analiza de acuerdo a un modelo de reputación predefinido similar a los existentes Merco©[6] y RepTrack©[7] El modelo incluye 7 dimensiones: estrategia y liderazgo, innovación y flexibilidad, integridad oferta responsabilidad social, situación financiera y trabajo. Además, cada una de estas categorías se divide en diferentes características. Dado un texto, un proceso de detección de entidades reconoce si se menciona una de las empresas del IBEX35 y si es así, un proceso de clasificación de textos es capaz de establecer cuáles de las dimensiones mencionadas se cubre. Finalmente, un proceso de análisis de sentimiento puede distinguir entre textos positivos y negativos con el fin de conocer si la gente tiene una idea buena o mala de una empresa en una determinada dimensión.

[2] http://www.consumerhealthvocab.org/
[3] http://gate.ac.uk/
[4] http://textalytics.com
[5] https://sentimentalytics.com/
[6] http://www.merco.info
[7] http://www.reputationinstitute.com

Bibliografía

Moreno, J., Declerck, T., Martínez Fernández, J.L. & Martínez, P. 2013. Prueba de Concepto de Expansión de Consultas basada en Ontologías de Dominio Financiero. Procesamiento del Lenguaje Natural, 51, 109-117.

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from Social Media. In Proc 51st Annual Meeting of the Association for Computational Linguistics, 993-1003

Rout, D., Preotiuc-Pietro, D., Bontcheva, K., & Cohn, T. Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties. 24th ACM Conference on Hypertext and Social Media, HT, May 2013, Paris.

Segura-Bedmar, I., Revert, R. & Martínez, P. 2014. Detecting drugs and adverse events from Spanish social media streams. En Proceedings of the 5th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2014).

Segura-Bedmar, I., Peña-González, S., & Martínez, P. (2014). Extracting drug indications and adverse drug reactions from Spanish health social media, Proceedings of the BioNLP 2014, June, 2014,

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 167-170 recibido 16-04-14 revisado 30-06-14 aceptado 02-07-14

Utilización de las Tecnologías del Habla y de los Mundos Virtuales para el Desarrollo de Aplicaciones Educativas*

Using Language Technologies and Virtual Worlds to Develop Educative Applications

D. Griol, A. Sanchis, J.M. Molina
Dpto. de Informática
Universidad Carlos III de Madrid
Av. de la Universidad, 30
28911 - Leganés (Spain)
{david.griol, araceli.sanchis, josemanuel.molina}@uc3m.es

Zoraida Callejas
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Granada
C/ Periodista Daniel Saucedo Aranda s/n
18071 - Granada (Spain)
zoraida@ugr.es

Resumen: Los continuos avances en el desarrollo de tecnologías de la información han dado lugar actualmente a la posibilidad de acceder a contenidos educativos en la red desde prácticamente cualquier lugar, cualquier momento y de forma casi instantánea. Sin embargo, la accesibilidad no suele considerarse como criterio principal en el diseño de aplicaciones educativas, especialmente para facilitar su utilización por parte de personas con discapacidad. Diferentes tecnologías han surgido recientemente para fomentar la accesibilidad a las nuevas tecnologías y dispositivos móviles, favoreciendo una comunicación más natural con los sistemas educativos. En este artículo se describe un Proyecto de Innovación Docente en el que se propone el uso innovador de los Sistemas Multiagente, los Sistemas de Diálogo y los Mundos Virtuales para el desarrollo de una plataforma educativa.
Palabras clave: Sistemas de Diálogo Hablado, Sistemas Multimodales, Interacción Oral, Educación, E-learning.

Abstract: Continuous advances in the development of information technologies have currently led to the possibility of accessing learning contents from anywhere, at anytime and almost instantaneously. However, accessibility is not always the main objective in the design of educative applications, specifically to facilitate their use by disabled people. Different technologies have recently emerged to foster the accessibility of computers and new mobile devices favouring a more natural communication between the student and the developed educative systems. This paper describes an Educational Innovation Project focused on the application of Multiagent Systems, Spoken Dialog Systems, and Virtual Worlds to develop an educative platform.
Keywords: Spoken Dialog Systems, Multimodal Systems, Spoken Interaction, Education, E-learning.

1 Introducción

Dentro de la visión de la Inteligencia Ambiental (Ambient Intelligence, AmI) (Nakashima, Aghajan, y Augusto, 2010), los usuarios estaremos rodeados de interfaces intuitivas e inteligentes embebidas en toda clase de objetos y en un entorno capaz de reconocer y responder ante nuestras acciones y los cambios que puedan sin obstaculizar y de forma continua y no visible. De este modo, se parte de la idea fundamental de que la tecnología debe estar diseñada para los usuarios y no los usuarios adaptarse a ella.

Para asegurar esta interacción natural e inteligente, es necesario proporcionar una comunicación eficaz, fácil, segura y transparente entre el usuario y el sistema. Por este motivo, los sistemas de diálogo (Pieraccini, 2012), que conjugan funcionalidades propias de las

* Trabajo parcialmente financiado por los proyectos TRA2011-29454-C03-03, MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02 y CAM CONTEXTS (S2009/TIC-1485). Desarrollado en el marco del Proyecto "Aplicación de nuevas metodologías docentes para un mejor aprovechamiento de las clases prácticas" (11ª Convocatoria de Apoyo a Experiencias de Innovación Docente en Estudios de Grado y Postgrado en la UC3M).

ISSN 1135-5948

Tecnologías del Habla y el Procesamiento del Lenguaje, pueden utilizarse para incorporar capacidades comunicativas inteligentes en los sistemas multiagente (multiagent systems, MAS) (Wooldridge, 2002).

La combinación de estas nuevas modalidades de comunicación con el aprendizaje social permite a los estudiantes interactuar con profesores y compañeros (tanto reales como simulados) durante las actividades y conseguir aprendizajes más significativos (Bishop, 2009). De este modo, las redes sociales han surgido como solución para integrar los agentes conversacionales en las comunidades 2.0. Dada la popularidad de estas tecnologías, se han producido durante la última década enormes avances en el desarrollo de redes sociales cada vez más complejas en las que poder llevar a cabo iniciativas relacionadas con el e-learning. Entre ellas destacamos los mundos sociales virtuales, que son entornos gráficos simulados por ordenador "cohabitados" por los usuarios a través de avatares (Boyd y Ellison, 2007).

Al mismo tiempo que se desarrollan estas iniciativas, las instituciones educativas en España estamos haciendo frente durante los últimos años a los desafíos que conlleva la implantación del nuevo marco del Espacio Europeo de Educación Superior (EEES). De hecho, el EESS supone una docencia fundamentalmente orientada al logro de competencias, de forma que la función del profesor sea facilitar y guiar al alumno para que pueda acceder intelectualmente a los contenidos y prácticas profesionales correspondientes a su titulación.

Para lograr este objetivo, es necesario el diseño de metodologías docentes más participativas y reflexivas en las que el alumno alcance el máximo desarrollo académico y personal de la forma más autónoma posible. En este espacio, el profesor deja de ser un transmisor de conocimientos para convertirse en un profesional que crea y organiza ambientes de aprendizaje complejos, implicando a los alumnos en su propio proceso de aprendizaje a través de las estrategias y actividades adecuadas. Éste es justamente uno de los aspectos fundamentales que debe enfatizarse al desarrollar cursos de aprendizaje on-line en los que el estudiante participa de forma autónoma en la mayor parte del proceso de aprendizaje.

A este respecto, en la Universidad Carlos III de Madrid, se ha definido que la evaluación sea continua y se base en el esfuerzo y la implicación activa del estudiante en las clases, la elaboración de trabajos y la realización de ejercicios y prácticas. De este modo, debe realizarse un seguimiento continuo del alumno mediante actividades que sirvan tanto para fomentar su participación y adquisición de conocimientos, como iniciativas que les permitan conocer sus progresos y evaluar su propio proceso de aprendizaje. En este marco, según Roda et al., los sistemas de aprendizaje virtual surgen como solución para: i) acelerar el proceso de aprendizaje, ii) facilitar el acceso al mismo, iii) personalizar el aprendizaje, y iv) proporcionar un entorno educativo más completo (Roda, Angehrn, y Nabeth, 2001).

El principal objetivo del proyecto que se describe es el desarrollo de una plataforma educativa, basada en el uso de agentes inteligentes, que permita al alumno la realización de cuestionarios de autoevaluación que incorporan preguntas seleccionadas automáticamente en función de los conceptos a enfatizar. El alumno podrá proporcionar sus respuestas utilizando la voz en lenguaje natural y recibir una respuesta del sistema también en lenguaje natural. Por último, la plataforma permitirá además la interacción y consulta de los contenidos desarrollados a través de mundos virtuales 3D como Second Life (`secondlife.com`) o los generados mediante la plataforma gratuita Open Simulator (`opensimulator.org`).

2 Plataforma educativa a desarrollar

Los objetivos principales definidos para la plataforma educativa a desarrollar son los siguientes. En primer lugar, el desarrollo de material docente multimedia innovador que facilite los procesos de desarrollo y evaluación de cursos online. En el proyecto nos alejamos de las herramientas multimedia docentes tradicionales que suelen basarse en la creación de material estático en un único formato.

En segundo lugar, nuestra propuesta se fundamenta en el desarrollo de un espacio virtual innovador, que haga de las asignaturas un espacio más flexible, participativo y atractivo. Para ello, llevamos

a la práctica de nuestra docencia los últimos avances realizados por los miembros del proyecto en materia investigadora. Concretamente, nuestro objetivo es la creación de una plataforma con agentes autónomos que funcione como espacio virtual en el que los alumnos puedan interactuar para presentarles casos y problemas que deben resolver, permitiendo esta interacción evaluar además su aprendizaje.

En tercer lugar, aplicamos los criterios de internacionalización de la docencia mediante la generación de contenidos tanto en inglés como en castellano, así como el posterior procesamiento teniendo en cuenta el idioma en el que el alumno haya suministrado las respuestas.

La arquitectura de la plataforma estará compuesta de una serie de módulos distribuidos en agentes que se encargan de facilitar la comunicación con el estudiante, la generación y selección de los contenidos, el análisis de las respuestas proporcionadas por el mismo, la generación de la respuesta adecuada teniendo en cuenta este análisis y la comunicación de la misma al estudiante. La plataforma educativa se fundamenta en tres tecnologías fundamentales: el desarrollo de sistemas multiagente, la interacción oral con el usuario y la capacidad de sociabilización y de realismo ofrecida por los mundos virtuales.

Los agentes inteligentes son entidades computacionales autónomas que en conjunto pueden resolver problemas computacionales complejos por medio del uso de modelos de razonamiento, aprendizaje y negociación. Podemos destacar tres características fundamentales presentes en la mayoría de los agentes inteligentes educativos: comunicación, inteligencia y autonomía (Wooldridge, 2002).

Estos agentes pueden comunicarse con el usuario, con otros agentes y con otros programas. Con el usuario se comunican con un interfaz amigable, mediante el que éste puede personalizar sus preferencias. Para el desarrollo de nuestra plataforma, hemos primado la comunicación oral con el estudiante. Con ello, perseguimos dos objetivos principales:

- facilitar el acceso a la aplicación en el mayor número de entornos y dispositivos posible, no limitando este acceso únicamente a aquellos que posibilitan el uso de los interfaces tradicionales;
- garantizar el acceso a la herramienta a estudiantes con discapacidades motoras que imposibiliten el uso de estos interfaces tradicionales, como es el caso del teclado o el ratón.

El grado de inteligencia varía mucho de unos agentes a otros, que suelen incorporar módulos con tecnologías procedentes de la Inteligencia Artificial (Litman y Silliman, 2004). Los más sencillos se limitan a recoger preferencias del usuario, quien debe personalizarlos. En nuestro caso, utilizamos la Tecnologías del Habla y de Procesamiento de Lenguaje Natural para posibilitar el análisis automático de las respuestas de los estudiantes. Para cada uno de los bloques temáticos que componen un curso, la herramienta seleccionará de la base de datos las preguntas correspondientes a los apartados que se desee evaluar en el idioma seleccionado por el estudiante.

Por último, en cuanto a la autonomía, un agente no sólo debe ser capaz de hacer sugerencias al usuario sino de actuar proactivamente. Los agentes de la plataforma poseerán la autonomía necesaria para seleccionar cuáles son los contenidos más apropiados que deben preguntarse al alumno y cuál es la respuesta que le debe proporcionar como resultado de la interacción.

Por último, el proyecto se apoya en la utilización de los mundos virtuales para posibilitar la interacción con la plataforma en entornos educativos inmersivos. Tradicionalmente, los mundos virtuales se han estructurado a priori predefiniendo las tareas realizables por los usuarios. En la actualidad, la interacción social posee un papel clave en estos entornos y los usuarios pueden determinar sus experiencias en el mundo virtual siguiendo sus propias decisiones. De este modo, los mundos virtuales se han transformado en verdaderas redes sociales útiles para la interacción entre personas de diferentes lugares. Tanto Second Life como los mundos virtuales desarrollados mediante Open Simulator presentan varias utilidades diseñadas específicamente para su uso educativo. Probablemente la más relevante es Sloodle (Simulation Linked

Object Oriented Dynamic Learning Environment, `sloodle.org/moodle/`), un proyecto de código abierto que integra el mundo virtual con el gestor de aprendizaje Moodle (Modular Object-Oriented Dynamic Learning Environment, `moodle.org/`). De este modo, se posibilitará el acceso a los contenidos de la plataforma de forma multimodal a través de la web mediante Moodle y también a través del mundo virtual mediante las herramientas facilitadas por Sloodle.

Para explotar estas tecnologías, el entorno educativo virtual totalmente está basado en tres tecnologías principales: la elaboración de formularios mediante las herramientas proporcionadas por Sloodle y con la posibilidad de interacción oral para proporcionar la respuesta a las preguntas incluidas en los mismos, un metabot que proporcione de forma automática información académica (Griol et al., 2012), y la construcción de objetos 3D que posibiliten al alumno una mayor compresión de los conceptos con un mayor nivel de abstracción.

Para posibilitar que el alumno responda a las preguntas mediante la voz, es necesario disponer de un reconocedor automático del habla, que obtenga la frase(s) que con mayor probabilidad corresponden con la señal de voz recibida. Seguidamente, el módulo de comprensión del lenguaje obtiene la interpretación semántica de las frases recibidas, utilizando para ello de gramáticas apropiadas para el análisis de los contenidos de cada unas de las preguntas. Mediante la obtención del significado y su comparación con la respuesta correcta elaborada por el personal docente del curso (contenida en la base de datos del sistema y accesible a través del módulo de consulta a la misma), el módulo de Análisis de la Respuesta calculará el porcentaje de éxito del estudiante y las recomendaciones que se le deben proporcionar.

3 Conclusiones y trabajo futuro

A lo largo del presente artículo se ha descrito el desarrollo de una plataforma para facilitar el aprendizaje autónomo y la autoevaluación de conocimientos en cursos on-line. La plataforma a desarrollar incluye funcionalidades que facilitan la generación de contenidos, el análisis de las respuestas de los estudiantes, la generación de una respuesta adecuada teniendo en cuenta el resultado de este análisis y la interacción con la herramienta de la forma más natural y personalizada al estudiante, utilizando para ello agentes conversacionales e interacción en mundos virtuales. Tal y como se ha descrito, estas nuevas tecnologías y entornos ofrecen una amplia gama de posibilidades educativas que los convierten en escenarios propicios para el aprendizaje, en los que los alumnos puedan además explorar, conocer a otros residentes, socializar, participar en actividades individuales y grupales, así como participar en la creación del entorno.

Bibliografía

Bishop, J. 2009. Enhancing the understanding of genres of web-based communities: The role of the ecological cognition framework. *International Journal of Web-Based Communities*, 5(1):4–17.

Boyd, D. y N. Ellison. 2007. Social network sites, definition, history and scholarship. *Journal of Computer Mediated Communication*, 13(1):210–230.

Griol, D., J.M. Molina, A. Sanchis, y Z. Callejas. 2012. A Proposal to Create Learning Environments in Virtual Worlds Integrating Advanced Educative Resources. *UCS Journal*, 18:2516–2541.

Litman, D.J. y S. Silliman. 2004. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. En *Proc. of Human Language Technology Conference*, páginas 5–8.

Nakashima, H., H. Aghajan, y J.C. Augusto. 2010. *Handbook of Ambient Intelligence and Smart Environments*. Springer.

Pieraccini, R. 2012. *The Voice in the Machine: Building Computers That Understand Speech*. MIT Press.

Roda, C., A. Angehrn, y T. Nabeth. 2001. Conversational Agents for Advanced Learning: Applications and Research. En *Proc. of BotShow Conference*, páginas 8–13.

Wooldridge, M. 2002. *An Introduction to MultiAgent Systems*. John Wiley & Sons.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 171-174 recibido 13-04-14 revisado 07-06-14 aceptado 10-06-14

Establishing a Linguistic Olympiad in Spain, Year 1

Estableciendo una olimpiada lingüística en España, año 1

Antonio Toral[†], Guillermo Latour[⋆], Stanislav Gurevich[△], Mikel Forcada[‡], Gema Ramírez-Sánchez[◇]

[†] NCLT, School of Computing, Dublin City University, Ireland

[⋆] IES Gran Vía. Alacant, Spain

[△] IOL Board and Liceo Anichkov, Saint Petersburg, Russia

[‡] Dep. Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

[◇] Prompsit Language Engineering, S.L., Elx, Spain

oleinfo@abumatran.eu

Resumen: Presentamos la OLE, una actividad de divulgación del proyecto europeo Abu-MaTran cuyo objetivo es establecer una olimpiada lingüística en España. Damos una introducción a las olimpiadas lingüísticas, presentamos nuestra motivación, nuestros objetivos y el plan de implementación. Pensamos que nuestro trabajo puede ser útil para otros países que quieran empezar una olimpiada lingüística.
Palabras clave: divulgación, lingüística, olimpiada

Abstract: We present OLE, a dissemination activity of the EU Abu-MaTran project which aims to establish a linguistic Olympiad in Spain. We introduce the Linguistic Olympiads, our rationale and objectives for setting up OLE as well as our implementation plan. We foresee our work to be useful for other countries looking to start a Linguistic Olympiad.
Keywords: dissemination, linguistics, olympiad

1 Introduction

The Linguistics Olympiad is a competition in which second level students are challenged to solve problems in the field of linguistics. It should be noted that the students do not need to have any prior knowledge of languages or even linguistics, as the puzzles are expected to be addressed with problem solving techniques such as logic and lateral thinking.

Most problems present samples of text in a language that the student does not know and involve finding certain patterns arising in these samples (and probably identifying exceptions too). In this regard, the problems of the Linguistic Olympiad can be considered as closely related to the field of computational linguistics (Littell et al., 2013).

The Linguistic Olympiads originated in the 1960s, and were initially run in eastern European countries such as Russia and Bulgaria. Recently the Olympiads have spread to other areas such as Asia, western Europe and North America (Radev, Levin, and Payne, 2008). Moreover, since 2003 there is an annual International Linguistics Olympiad[1] (IOL), which is attended by the winners of the national competitions. The last edition of the IOL was held in Manchester (United Kingdom) with 138 contestants in 35 teams, representing 26 countries. Until now, Spain has not taken part in the IOL (except for an ad-hoc team[2] in the 2007 edition).

This paper discusses the establishment of a sustainable Linguistics Olympiad in Spain (OLE).[3] OLE is a current dissemination activity of Abu-MaTran,[4] a research project in the field of machine translation funded under Marie Curie's Industry-Academia Partnerships and Pathways (IAPP) action.[5] IAPP projects focus in transfer of knowledge between academia and industry.

The rest of the paper is organised as follows. In the following section we cover the

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

[2] Ad-hoc teams are teams from a country where no national competition exists that compete in the international Olympiad.

[3] OLE stands for Olimpiada Lingüística d'Espanya (Catalan), Olimpiada Lingüística de España (Spanish), Espainako Olinpiada Linguistikoa (Basque) and Olimpíada Lingüística de España (Galician).

[4] http://www.abumatran.eu/

[5] http://ec.europa.eu/research/mariecurieactions/about-mca/actions/iapp/index_en.htm

ISSN 1135-5948

rationale and the objectives of the OLE. Subsequently we outline the overall plan to implement the activity. This is followed by an account of the first edition of the olympiad. We then report on the exercise types used. Finally we draw conclusions and plan our future lines of work.

2 Rationale and Objectives

There are two main reasons that encouraged us to carry out this project:

- As mentioned before, there has not been a Linguistics Olympiad in Spain so far. Linguistic Olympiads promote problem solving skills such as logic, lateral thinking, etc. which we deem of paramount importance for students. The fact that the performance by Spanish students in the latest PISA test on problem solving was significantly lower than OCDE's average (477 points vs 500)[6] seems to support our hypothesis that more emphasis on problem solving is needed.

- The Abu-MaTran consortium is in a good position to run this project as one of the partners, Dublin City University (DCU), has extensive experience in the area. In fact, DCU organises the All Ireland Linguistics Olympiad[7] annually since 2009.

  As previously mentioned, the Abu-MaTran project focuses on intersectoral transfer of knowledge. In the case of the OLE, this implies the transfer of organisational knowledge from DCU's experience to Prompsit, the industrial partner of the consortium, based in Spain.

The main objectives of the OLE can be summarised as follows:

- Foster the acquisition and mastering of problem solving techniques by the participant students.

- Develop the interest of students in the area of linguistics and in the knowledge of new languages.

- Get students acquainted with the area of computational linguistics and related disciplines such as linguistics, translation and computer science. In this regard, OLE is contributing to the area of computational linguistics in the long term by bringing this area to the next generation of researchers.

3 Plan and Implementation

The aim of the Abu-MaTran project with respect to OLE is to set up a sustainable Olympiad in Spain over the course of the project (January 2013 – December 2016).

In this respect, the plan is to run three annual Olympiads during the second (2014), third (2015) and final year (2016) of the project as follows:

- First edition (2014). Pilot task, targetting the area of Alacant/Alicante province (1,917,012 inhabitants, 5,816 km^2).

- Second edition (2015). The area is extended to the Valencian Community (5,111,706 inhabitants, 23,255 km^2)

- Third edition (2016). The area is extended to Spain (46,815,916 inhabitants, 505,992 km^2). Starting with 2016 we hope to proceed with that level further on.

This iterative approach allows us to adopt initially the organisative model of AILO for the two first years, since our target area is similar (Valencian Community) or smaller (Alacant/Alicante) to that of AILO, Ireland (6,378,000 inhabitants, 84,421 km^2). During these first two years we are gaining experience that will allow us to adapt the model as necessary in order to tackle successfully and efficiently our whole target area, Spain, substantially bigger than that of our initial organisative model.

4 First Edition

We now discuss in more detail the first edition of OLE, carried out from September 2013 to July 2014. The main phases have been the following:

- September 2013. Development of the website of OLE,[8] as well as its corporate image (logo, font, colours, etc) and the relevant materials (e.g. brochure).

[6] http://www.mecd.gob.es/inee/Ultimos_informes/PISA-2012-resolucion-de-problemas.html

[7] http://www.cngl.ie/ailo/

[8] http://ole.abumatran.eu

- October–December 2013. Registration period for interested schools. Two sets of training excercises were prepared and sent to registered schools.
- January 2014. First round, carried out in each school. The top 75 students qualify for the second round.
- March 2014. Second round, carried out at Miguel Hernandez University (Elx), where Prompsit is based. The top 4 students qualify for the IOL final.
- June 2014. Training session for students that have qualified.
- July 2014. The top 4 students from the second round represent the OLE at the final of the IOL (Beijing, China).

The reception of OLE has been already very satisfactory in its first year. 20 schools registered to take part in OLE's first edition. From these schools, over 400 students took part in the first round.

5 Exercises

We report on the exercises used in the first edition of OLE.[9] Exercises can be classified in different types, according to the linguistic phenomena they are about. Exercises can also be classified according to their level of difficulty.

Individual tests (training, first round and second round) in OLE are made up of a number of exercises (6 in all our tests). We have designed the tests so that they have exercises of different levels of difficulty (and accordingly different punctuation) and the exercises belong to different types. It should be mentioned that it is almost impossible to solve all of the problems in the time provided (3 hours). Therefore, time management is a relevant skill to tackle the tests. It is up to the student to manage their time to maximise their chances to obtain high marks.

Apart from these tests, which are tackled by students individually, we have also designed a group test (for the second round). The group test consisted of one exercise, considerably more difficult than the exercises in the individual tests. In the group test, students are expected to apply techniques of work in group (e.g. divide the problem in simpler tasks).

[9] http://ole.abumatran.eu/?page_id=8

Most of the exercises used in the training and the two rounds are previous years' exercises from other chapters of the IOL (Ireland and Russia). They were adapted and translated accordingly (from English and Russian into Catalan and Spanish). In this respect, we should note that the collaboration with other chapters of the IOL was very beneficial, as it allowed us to use exercises of good quality and appropriate difficulty.

For illustrative purposes, we provide details of the languages and topics of the exercises used in the first and second rounds of OLE's first edition (see Table 1). Furthermore, an exercise is shown in Appendix A.

6 Conclusions and Future Work

We have presented OLE, a dissemination activity of the Abu-MaTran project which aims to establish a sustainable linguistic Olympiad in Spain. We have introduced the Linguistic Olympiads, our rationale and objectives for setting up OLE as well as our plan for implementing it. We have then reported on the first edition of OLE, corresponding to the academic year 2013–2014. Finally, we have given a detailed account of the exercise types and the tests of the Olympiad.

The outcome of the first year of the Olympiad is very positive, with over 400 students taking part. As a side effect, we foresee our current work to be useful for other countries looking to start a Linguistic Olympiad.

Looking into the near future, we will face challenges due to the plan to extend the area covered by the Olympiad. In this regard we are looking at how to adapt our organisative model to be able to run the Olympiad while staying within our limits (budget and manpower). On-line tests (used e.g. by the North American chapter of the Olympiad) and a distributed Olympiad (as done e.g. by the Russian chapter), among others, are models that we consider exploring.

As another corresponding activity we may be looking forward to introducing special courses for high school students as well as organising vacational training camps (as it has been done for many years in Bulgaria, Russia, Estonia and some other countries) where students could get preliminary acquaintance with the basics of linguistics, a discipline which is to a very subtle extent included into traditional school educational patterns.

Language	Topic
Invented	Numbers, order of number positions
La-Mi	Syllable alternation
Unua	Translation, declination and order of sentence constituents
Amharic	Translation, morphology
Japanese	Adjectives, declination
Panyabi	Translation, tones
Latvian	Conjugation of verbs
Aroma	Numbers, morphology
Mundari	Translation, morphology
Tokhari	Translation, imperative verbs
English	Numbers, numbering system and rhyme
Turkish	Translation, morphology

Table 1: Language and topic of each of the exercises of the first and second rounds of OLE's first edition

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

We would like to thank organisers of the All Ireland Linguistic Olympiad (Cara Green, Harold Somers and Laura Grehan) for their advice in setting up a Linguistic Olympiad and for allowing us to use their exercises.

Similarly, we would like to thank the organisation (Polina Pleshak) and authors of problems of the Russian chapter of the Linguistic Olympiad for allowing us to use their exercises, assisting us with their translation and, most of all, for attending OLE's second round and helping us with its organisation.

We also thank institutions for having given their support for OLE's first edition, namely the Department of Software and Computing Systems of the University of Alicante, Fundación Quorum and the Vicerrectorado de Investigación of the Miguel Hernandez University.

References

Littell, Patrick, Lori Levin, Jason Eisner, and Dragomir Radev. 2013. Introducing Computational Concepts in a Linguistics Olympiad. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 18–26, Sofia, Bulgaria, August. 8 pages.

Radev, Dragomir R., Lori S. Levin, and Thomas E. Payne. 2008. The North American Computational Linguistics Olympiad (NACLO). In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, TeachCL '08, pages 87–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Sample Exercise: Complicated Numeration System

There isn't any language that uses such a complicated numeration system, but let's imagine that there is a language with numbers such as:
1 nut
2 dok
3 tris
4 kwat
5 kwin
6 ses
7 sep
8 ok
9 nou
10 des
100 hun
1000 mil
11 desnut
14 deskwat
20 dokdes
32 doktrisdes
47 sepkwatdes
185 hunkwinokdes
237 dokhunseptrisdes
1234 dokhunkwattrisdesmil
4567 kwinhunsepsesdeskwatmil

Task. Write the following numbers:
78, 874, 3210, 215

Demostraciones y Artículos de la Industria

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 177-180 recibido 08-04-14 revisado 07-06-14 aceptado 17-06-14

ADRSpanishTool: una herramienta para la detección de efectos adversos e indicaciones

ADRSpanishTool: a tool for extracting adverse drug reactions and indications

Santiago de la Peña*, Isabel Segura-Bedmar*, Paloma Martínez*, José Luis Martínez**
*Universidad Carlos III de Madrid, Av. Universidad, 30, 28911, Madrid
**Daedalus SA, Avda. de la Albufera 321, 28031, Madrid
spena@pa.uc3m.es, {isegura,pmf}@inf.uc3m.es, jmartinez@daedalus.es

Resumen: Presentamos una herramienta basada en coocurrencias de fármaco-efecto para la detección de reacciones adversas e indicaciones en comentarios de usuarios procedentes de un foro médico en español. Además, se describe la construcción automática de la primera base de datos en español sobre indicaciones y efectos adversos de fármacos.
Palabras clave: Extracción de Información, Medios Sociales

Abstract: We present a tool based on co-occurrences of drug-effect pairs to detect adverse drug reactions and drug indications from user messages that were collected from an online Spanish health forum. In addition, we also describe the automatic construction of the first Spanish database for drug indications and adverse drug reactions.
Keywords: Information Extraction, Social Media

1 Introducción

El objetivo principal de la Farmacovigilancia consiste en estudiar el uso y los efectos de los medicamentos en los pacientes. Además, se encarga de generar alertas sobre posibles reacciones adversas a un medicamento, intentando intenta evaluar el riesgo, para finalmente informar a los profesionales sanitarios y a los pacientes con el fin de evitar estas reacciones adversas. En los últimos años, la Farmacovigilancia ha ganado cada vez más interés debido al creciente número de reacciones adversas (Bond y Raehl, 2006) y el coste asociado de estas reacciones (van Der Hooft et al., 2006).

Hoy en dia, los principales organismos regulatorios de medicamentos, como son la "Food and Drug Administration"(FDA) en EE.UU o la Agencia de Medicina Europea (EMA), trabajan activamente en el diseño de políticas e iniciativas dirigidas a facilitar la notificación de reacciones adversas a medicamentos por los profesionales sanitarios y por los pacientes. Sin embargo, varios estudios han demostrado que la notificación de estas reacciones adversas es aún poco frecuente debido a que muchos profesionales de la salud no tienen tiempo suficiente para utilizar los sistemas de notificación de reacciones adversas (Bates et al., 2003; van Der Hooft et al., 2006; McClellan, 2007) . Además, los profesionales sanitarios sólo informan de aquellas reacciones de las que tienen certeza absoluta de su existencia(Herxheimer, Crombag, y Alves, 2010). A diferencia de los informes de los profesionales de la salud, los informes de los pacientes a menudo proporcionan información más detallada y explícita sobre las reacciones adversas. Sin embargo, la tasa de reacciones notificadas por los pacientes es todavía muy baja, probablemente debido a que muchos pacientes todavía no son conscientes de la existencia de estos sistemas de notificación.

Nuestra hipótesis principal es que la información que los pacientes escriben en distintos medios sociales, como por ejemplo los foros sobre temas de salud, pueden ser un recurso complementario para los sistemas de notificación de reacciones adversas.

En los últimos años, se han desarrollado varios sistemas dedicados a la detección de reacciones adversas en medios sociales (Leaman et al., 2010; Nikfarjam y Gonzalez, 2011), sin embargo ninguno de estos trabajos ha tratado el problema en español.

ISSN 1135-5948

En este artículo, presentamos una herramienta para la detección automática de efectos adversos e indicaciones de fármacos en mensajes de usuarios de ForumClinic[1], una plataforma social donde los pacientes intercambian información sobre sus enfermedades y sus tratamientos.

2 *Descripción de la herramienta ADRSpanishTool*

La infraestructura para desarrollar y desplegar los componentes de la herramienta que se ha usado ha sido GATE[2]. Para procesar los mensajes, hemos utilizado la herramienta de Textalytics[3], que sigue un enfoque basado en diccionario para identificar los fármacos y sus efectos. Para la construcción del diccionario, utilizamos los siguientes recursos: CIMA[4] y MedDRA[5]. CIMA es una base de datos online administrada por la Agencia Español de Medicamento y productos sanitarios (AEMPS) que contiene información sobre todos los fármacos aprobados en España. MedDRA es un recurso multilingüe que proporciona información sobre eventos adversos de medicamentos. El diccionario contiene un total de 5.800 fármacos y 13.246 efectos adversos con 48,632 sinónimos distintos. Además, se han creado varios gazetteers para aumentar la cobertura en la detección de fármacos y efectos proporcionada por el diccionario. En concreto, hemos desarrollados varios crawlers para buscar y descargar páginas relacionadas con fármacos de sitios webs como MedLinePlus[6] y Vademecum[7]. Mediante el uso de expresiones regulares aplicadas sobre las páginas descargadas, conseguimos obtener de forma automática una lista de fármacos y de efectos. El sistema ATC[8], para la clasificación de fármacos, también fue utilizado y volcado en un gazetteer para detectar nombre de grupos de fármacos.

GATE proporciona una herramienta de anotación de patrones denominada JAPE (Java Annotation Patterns Engine). Es una versión del Common Pattern Specification Language (CPSL) que mediante distintas fases, cada una de las cuales conteniendo unas reglas patrón/acción, constituyen una cascada de transductores de estados finitos que actúan sobre las anotaciones. En nuestro sistema se usa esta herramienta para filtrar las anotaciones procedentes del diccionario de Textalytics y separar las procedentes de los Gazetteers.

Aunque hay varias bases de datos con información sobre fármacos y sus efectos, como SIDER[9] o MedEffect[10], ninguna está disponible en español. Además, estos recursos no incluyen indicaciones. Así, hemos construido de manera automática la primera base de datos disponible en español, *SpanishDrugEffectBD*, con información sobre fármacos, sus indicaciones y sus reacciones adversas. Esta base de datos se puede usar para identificar de manera automática indicaciones y reacciones adversas en textos. La figura 1 muestra el esquema de la base de datos. El primer paso fue poblar la base de datos con los fármacos y efectos de nuestro diccionario. Los ingredientes activos se almacenan en la tabla *Drug*, mientras que sus sinónimos y nombres comerciales en la tabla *DrugSynset*. Igualmente, se almacenan los conceptos obtenidos de MedDRA en la tabla *Effect* y sus sinónimos en la tabla *EffectSynset*.

Para obtener las relaciones entre los fármacos y sus efectos, desarrollamos varios crawlers para descargar los apartados sobre indicaciones y reacciones adversas de prospectos contenidos en las siguientes webs: MedLinePlus, Prospectos.Net[11] y Prospectos.org[12]. Una vez descargados estos apartados, fueron procesados usando la herramienta Textalytics para etiquetar los fármacos y sus efectos. Los efectos descritos en el apartado de indicaciones de un fármaco fueron almacenados como relaciones de tipo indicación. De forma similar, los efectos descritos en los apartados de reacciones adversas fueron almacenados como relaciones de tipo reacción adversa.

La herramienta utiliza un enfoque basado en coocurrencia de entidades para extraer las relaciones. Además, mediante la consulta a la base de datos se comprueba si una relación corresponde a una indicación o a un efecto

[1] http://www.forumclinic.org/
[2] http://gate.ac.uk/
[3] https://textalytics.com/
[4] http://www.aemps.gob.es/cima/
[5] http://www.meddra.org/
[6] http://www.nlm.nih.gov/medlineplus/spanish/
[7] www.vademecum.es
[8] http://www.whocc.no/atc/
[9] http://sideeffects.embl.de/
[10] http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php
[11] http://www.prospectos.net/
[12] http://prospectos.org/

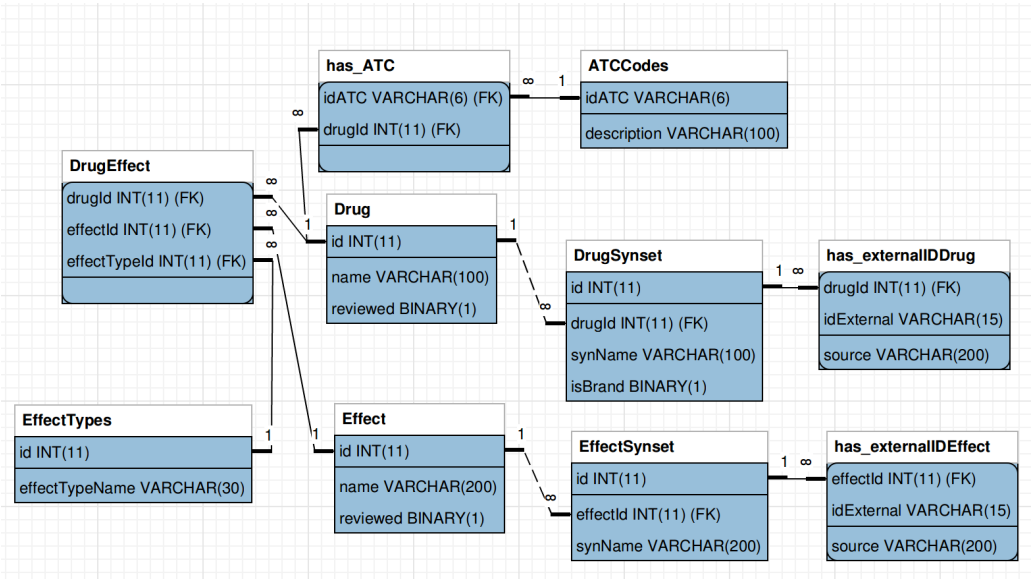


Figura 1: Esquema de la base de datos SpanishDrugEffectBD

adverso. Creamos una aplicación de Gate[13] en la que integramos el módulo de Textalytics, los gazetteers y la extracción de relaciones fármaco-efecto usando la base de datos. En la figura 2 puede verse una representación de la aplicación.

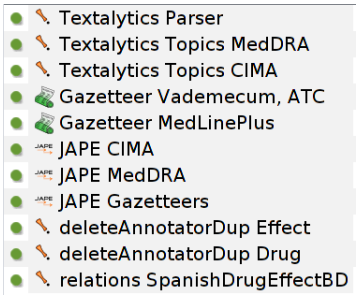


Figura 2: Aplicación completa para anotación de relaciones fármaco-efecto en GATE

El lector puede encontrar una descripción más detallada de los recursos y procesos utilizados por la herramienta en el trabajo (Segura-Bedmar, Revert, y Martínez, 2014).

Mediante la tecnología Jetty[14] implementada en el GATE se desplegó un servidor web para ejecutar la aplicación y poder visualizar los resultados. En la figura 3 se puede ver un ejemplo de salida de la aplicación.

La herramienta ha sido evaluada sobre el corpus SpanishADR[15], el primer corpus en Español formado por 400 comentarios. Estos comentarios fueron obtenidos de ForumClinic y anotados por dos anotadores. El corpus contiene un total de 188 fármacos y 545 efectos. Además, se han anotado 61 indicaciones y 103 reacciones adversas. La herramienta muestra una precision del 86 % y una cobertura del 81 % en la detección de fármacos. La detección de efectos es una tarea más difícil debido al gran número de expresiones idiomáticas que los pacientes utilizan para describir sus experiencias con los medicamentos. La herramienta sólo consigue un 63 % de precision y un 51 % en la detección de entidades de tipo efecto. Respecto a los resultados en la extracción de relaciones fármaco-efecto, la herramienta consigue una buena precisión de un 83 % pero con muy baja cobertura 15 %.

3 *Conclusión y Trabajo Futuro*

En este artículo presentamos la primera herramienta dedicada a la detección de indicaciones y reacciones adversas en mensajes de usuario obtenidos de un foro sobre salud en español. Una demo de la herramienta está disponible en el sitio web: http://163.117.129.57:8090/gate/.

También se describe en el trabajo la creación de una base de datos de indicaciones y reacciones adversas obtenidos de forma automática de prospecto sobre fármacos. Esta es la primera base de datos disponible en español. Aunque su uso no mejora los resultados por su limitada cobertura, pensamos que puede ser un recurso valioso para futuros

[13] http://gate.ac.uk/
[14] http://www.eclipse.org/jetty/
[15] http://labda.inf.uc3m.es/SpanishADRCorpus

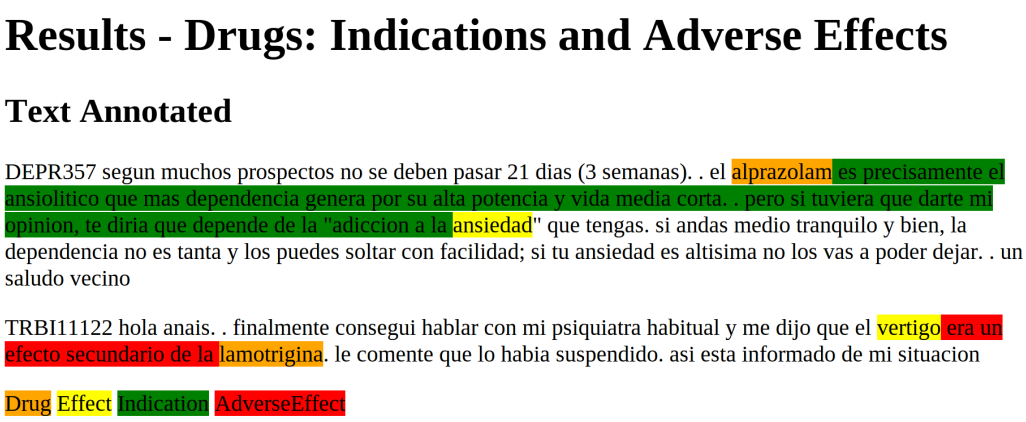


Figura 3: Ejemplo de la salida producida por la aplicación

desarrollos.

Este trabajo se enmarca dentro del proyecto europeo TrendMiner[16], siendo uno de sus principales objetivos incrementar la cobertura multilingüe de herramientas y ontologías en el dominio médico.

En un futuro nos planteamos aumentar el tamaño del corpus para poder aplicar técnicas de aprendizaje automático. También nos gustaría crear un diccionario con vocabulario y expresiones idiomáticas que los pacientes utilizan a la hora de describir sus tratamientos y efectos. Este tipo de recursos puede ser muy útil a la hora de desarrollar sistemas de detección de efectos adversos. También nos planteamos transformar la base de datos en una ontología y poblarla con más conceptos y relaciones.

Agradecimientos

Este trabajo se ha financiado por el proyecto europeo TrendMiner [FP7-ICT287863], por el proyecto MULTIMEDICA [TIN2010-20644-C03-01], y por la Red de InvestigaciónMA2VICMR [S2009/TIC-1542].

Bibliografía

Bates, DW., RS. Evans, H. Murff, PD. Stetson, L. Pizziferri, y G. Hripcsak. 2003. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115–128.

Bond, CA. y CL. Raehl. 2006. Adverse drug reactions in United States hospitals. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 26(5):601–608.

[16]http://www.trendminer-project.eu/

Herxheimer, A., MR. Crombag, y TL. Alves. 2010. Direct patient reporting of adverse drug reactions. a twelve-country survey & literature review. *Health Action International (HAI)(Europe). Amsterdam.*

Leaman, R., L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, y G. Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. En *Proceedings of the 2010 workshop on biomedical natural language processing*, páginas 117–125. Association for Computational Linguistics.

McClellan, M. 2007. Drug safety reform at the fda—pendulum swing or systematic improvement? *New England Journal of Medicine*, 356(17):1700–1702.

Nikfarjam, A. y GH. Gonzalez. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. En *AMIA Annual Symposium Proceedings*, volumen 2011, página 1019. American Medical Informatics Association.

Segura-Bedmar, I., R. Revert, y P. Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. En *Proceedings of the 5th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2014)*.

van Der Hooft, CS., MCJM. Sturkenboom, K. van Grootheest, HJ. Kingma, y BHCh. Stricker. 2006. Adverse drug reaction-related hospitalisations. *Drug Safety*, 29(2):161–168.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 181-184 recibido 10-04-14 revisado 07-07-14 aceptado 07-07-14

ViZPar: A GUI for ZPar with Manual Feature Selection

ViZPar: una GUI para ZPar con Selección Manual de Features

Isabel Ortiz[1], Miguel Ballesteros[2] and Yue Zhang[3]
[1]Pompeu Fabra University, Barcelona, Spain
[2]Natural Language Processing Group, Pompeu Fabra University, Barcelona, Spain
[3]Singapore University of Technology and Design
[1]iortiztornay@gmail.com
[2]miguel.ballesteros@upf.edu, [3]yue_zhang@sutd.edu.sg

Resumen: Los analizadores de dependencias y constituyentes se utilizan masivamente en la comunidad de Procesamiento de Lenguaje Natural. ZPar implementa versiones precisas y eficientes de algoritmos shift-reduce para parsing. En este artículo se presenta ViZPar, que es una interfaz gráfica de usuario para ZPar incluyendo visualización de árboles y selección automática de features. Durante la sesión de demostración se ejecutará ViZPar, para dependencias y constituyentes, y explicaremos las funcionalidades del sistema.
Palabras clave: Análisis de dependencias, Análisis de constituyentes, ZPar

Abstract: Phrase-structure and dependency parsers are used massively in the Natural Language Processing community. ZPar implements fast and accurate versions of shift-reduce dependency and phrase-structure parsing algorithms. We present ViZPar, a tool that enhances the usability of ZPar, including parameter selection and output visualization. Moreover, ViZPar allows manual feature selection which makes the tool very useful for people interested in obtaining the best parser through feature engineering, provided that the feature templates included in ZPar are optimized for English and Chinese. During the demo session, we will run ViZPar for the dependency and the phrase-structure versions and we will explain the potentialities of such a system.
Keywords: Dependency parsing, Phrase-Structure parsing, ZPar

1 Introduction

Natural language researchers and application developers apply dependency and constituency parsing frequently, however the parsers normally require careful tuning, complex optimization and the usage of complex commands that hinder their use.

ZPar[1] is a state-of-the-art parser implemented in C++ focused on efficiency. It provides a dependency parser (Zhang and Nivre, 2011; Zhang and Nivre, 2012) and a phrase-structure parser (Zhang and Clark, 2011a; Zhu et al., 2013), both implemented by shift-reduce parsing algorithms (Nivre, 2003; Nivre, 2008; Sagae and Lavie, 2005). ZPar gives competitive accuracies and very fast parsing speeds on both tasks. However, ZPar requires deep knowledge in command-line interfaces and programming skills (especially if the user is interested in performing feature engineering), which leaves its use to researchers that are able to understand the intricacies of such a system. As a result it is relatively more difficult to use by corpus linguists and researchers who need to apply syntactic analysis, but are not familiar with parsing research.

The usability limitation applies to other parsers also, including the Collins parser (Collins, 1999) or MaltParser[2] (Nivre et al., 2007), as it is a common practice for statistical parsers to use command-line interfaces. On the other hand, there has been a call for enhanced usability of parsers,[3] and visualization tools and application wrappers; they have made a non negligible impact on the parsing research field.

In order to have a system that tries to en-

[1]http://sourceforge.net/projects/zpar/

[2]MaltParser has been one of the most widely used parsers, since there are existing parallel tools, including visualization tools.

[3]See Section 4.

ISSN 1135-5948

hance the usability of ZPar, we present ViZPar,[4] which is a tool implemented in Java that provides a graphical user interface of ZPar, in its dependency and phrase-structure versions. ViZPar also allows manual feature engineering given the ZPar feature templates and provides automatic evaluation and comparison tools with the gold-standard.

2 ZPar

ZPar is a statistical syntactic analyzer that performs tokenization/segmentation, POS-tagging, dependency parsing and constituent parsing functionalities. ZPar is language-independent but contains optimized versions for the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) and the Chinese Treebank (Xue et al., 2004). Currently, in its out of the box version, it gives highly competitive accuracies on both English (Zhu et al., 2013) and Chinese (Zhang et al., 2013) benchmarks.

ZPar is implemented using the shift-reduce parsing mechanism (Yamada and Matsumoto, 2003; Nivre, 2008; Sagae and Lavie, 2005). It leverages a global discriminative training and beam-search framework (Zhang and Clark, 2011b; Zhang and Nivre, 2012) to improve parsing accuracies while maintaining linear time search efficiency. As a result ZPar processes over 50 sentences per second for both constituent parsing and dependency parsing on standard hardware. It is implemented in C++, and runs on Linux and MacOS. It provides command-line interfaces only, which makes it relatively less useful for researchers on corpus linguistics than for statistical parsing researchers.

3 ViZPar: a Vizualization tool for ZPar

ViZPar is a graphical user interface of ZPar implemented in Java. In its current version it supports a GUI for training and using ZPar, including the visualization of dependency and constituent outputs, evaluation and comparison with gold-standard treebanks, manual configuration and feature selection.

3.1 Java Wrapping of ZPar

The ZPar package includes a bash script that compiles the C++ code, trains a parsing model, runs the parser over the development set and finds the results of the best iteration. ViZPar provides a graphical-user interface for running the ZPar process, which wraps the entire process with a graphical user interface.

A user of ViZPar needs to download ZPar, which contains the C++ source code and the Python and bash scripts needed. On initialization, ViZPar asks for the ZPar directory, which is needed to train models and run the parser. After that, the user should proceed as follows:

- Selection of parsing mode: the user selects whether he/she wants to run a dependency parser or a constituent parser.
- Selection of mode: the user may select an existing parsing model or train a new one by also setting the number of iterations.
- Selection of the test set for parsing and evaluation.

Finally, when the process is finished the system allows to visualize the output by using graphical visualization of the syntactic trees. This feature is explained in the following section.

3.2 Tree Visualization

In order to visualize the output and the gold standard trees of the dependency and phrase structure versions we implemented two different solutions. For the dependency parsing version, we reused the source code of MaltEval (Nilsson and Nivre, 2008) for tree visualization, which includes all its functionalities, such as zooming or digging into node information, and for the constituent parsing version, we implemented a completely new tree visualizer.

Figure 1 shows the ViZPar graphical user interface when it has already parsed some dependency trees with a model trained over an English treebank. The dependency tree shown at the top of the picture is the gold standard and the one shown below is the output provided by the ZPar model. In the same way, Figure 2 shows ViZPar GUI in the case of the phrase-structure parser, which allows to traverse the tree and check the outcome quality by comparison with the gold-standard.

[4] ViZPar stands for graphical visualization tool for ZPar.

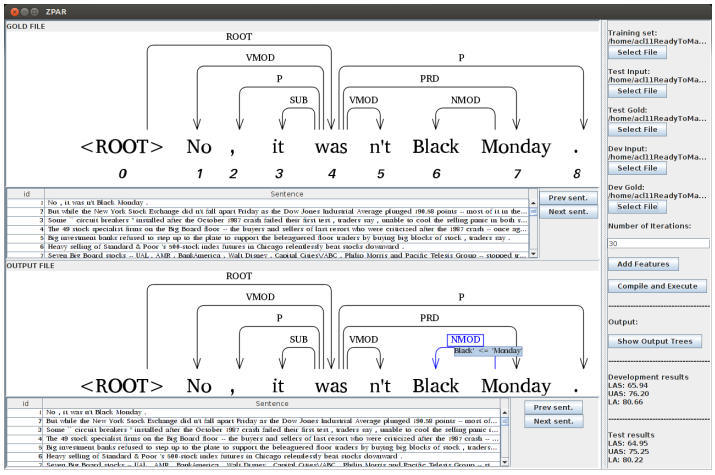


Figure 1: ViZPar in 'dependency' mode.

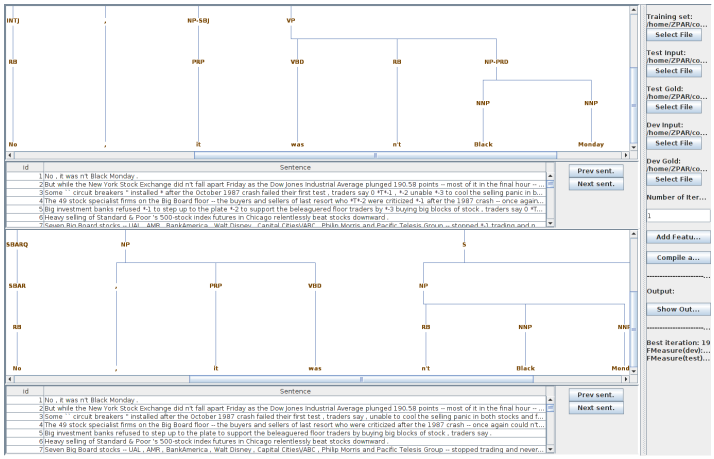


Figure 2: ViZPar in 'constituent' mode.

3.3 Feature Selection

ZPar provides rich feature templates, depicted by Zhang and Nivre (2011; 2012) for dependency parsing and by Zhu et al. (2013) for constituent parsing. However the features are handcrafted in the source code; this fact means that if the users would like to update the set of features for a new language or treebank, they would have to change the source code, compile it and run it again, provided that the user knows where the feature templates are and how to encode them. In ViZPar we provide a framework that allows the selection of the different features, and changes the source code automatically, it also makes the compilation and outputs the parser ready to generate a new model.

Our algorithm is implemented by scanning through the ZPar source code, detecting lines on feature definition, which follow regular patterns, and listing the features in a dialog box. The algorithm changes the ZPar source code according to the user selection by commenting out features that are deselected.

The manual feature selection tool might also provide the opportunity of running automatic and manual feature selection experiments as in MaltOptimizer (Ballesteros and Nivre, 2012).

4 Related Work

There has been recent research on visualization in the NLP community. In the parsing area we can find systems, such as MaltEval (Nilsson and Nivre, 2008), which allows the comparison of the output with a gold standard and also includes statistical significance tests. The Mate Tools (Bohnet, Langjahr, and Wanner, 2000) provide a framework for generating rule-based transduction and visualization of dependency structures. Icarus (Gartner et al., 2013) is a search tool and visualizer of dependency treebanks. Finally,

MaltDiver (Ballesteros and Carlini, 2013) visualizes the transitions performed by MaltParser.

5 Conclusions

In this paper we have presented ViZPar which is a Java graphical user interface of ZPar. We have shown its main functionalities, that are: (1) run ZPar in a user friendly environment, (2) dependency and constituent tree visualization and (3) manual feature engineering. ViZPar can be downloaded from `http://taln.upf.edu/system/files/resources_files/ViZPar.zip`

References

Ballesteros, Miguel and Roberto Carlini. 2013. MaltDiver: A Transition-Based Parser Visualizer. In *Proceedings of the System Demonstration Session of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*.

Ballesteros, Miguel and Joakim Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Bohnet, Bernd, Andreas Langjahr, and Leo Wanner. 2000. A development environment for an mtt-based sentence generator. In *Proceedings of the First International Natural Language Generation Conference*.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Gartner, Markus, Gregor Thiele, Wolfgang Seeker, Anders Bjorkelund, and Jonas Kuhn. 2013. Icarus – an extensible graphical search tool for dependency treebanks. In *ACL-Demos*, August.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Nilsson, Jens and Joakim Nivre. 2008. Malteval: an evaluation and visualization tool for dependency parsing. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Nivre, J. 2003. An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryiğit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13:95–135.

Nivre, Joakim. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34:513–553.

Sagae, Kenji and Alon Lavie. 2005. A classifier-based parser with linear runtime complexity. In *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT)*, pages 125–132.

Xue, Naiwen, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2004. The Penn Chinese Treebank: Phase structure annotation of a large corpus. *Journal of Natural Language Engineering*, 11:207–238.

Yamada, Hiroyasu and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 195–206.

Zhang, Meishan, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *ACL*.

Zhang, Yue and Stephen Clark. 2011a. Shift-reduce ccg parsing. In *ACL*.

Zhang, Yue and Stephen Clark. 2011b. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Zhang, Yue and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *ACL (Short Papers)*, pages 188–193.

Zhang, Yue and Joakim Nivre. 2012. Analyzing the effect of global learning and beam-search on transition-based dependency parsing. In *COLING (Posters)*.

Zhu, Muhua, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *ACL*.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 185-188 recibido 16-04-14 revisado 30-06-14 aceptado 02-07-14

Desarrollo de portales de voz municipales interactivos y adaptados al usuario*

Development of Interactive and User-Centered Voice Portals to Provide Municipal Information

David Griol, María García-Jiménez, José Manuel Molina, Araceli Sanchis
Universidad Carlos III de Madrid
Av. de la Universidad, 30
28911 - Leganés (Spain)
{david.griol, josemanuel.molina, araceli.sanchis}@uc3m.es, 100025080@alumnos.uc3m.es

Resumen: El objetivo principal de este artículo es mostrar la aplicación práctica de los sistemas de diálogo y del estándar VoiceXML para el desarrollo de portales de voz de información municipal. A través del portal desarrollado, los usuarios pueden interactuar telefónicamente para consultar información sobre el Ayuntamiento y la ciudad, realizar diversas gestiones y trámites, completar encuestas, acceder al buzón del ciudadano y ser transferido a la centralita del Ayuntamiento. Estas funcionalidades, conjuntamente con los mecanismos introducidos para su adaptación al usuario y a la evolución del diálogo, hacen que la aplicación desarrollada pueda considerarse como un avance considerable con respecto al desarrollo actual de los portales de voz en España. El artículo describe los servicios proporcionados por el portal, arquitectura y tecnologías utilizadas.
Palabras clave: Portal de Voz, Sistemas de Diálogo, VoiceXML, Adaptación al Usuario.

Abstract: In this paper, we describe a Voice Portal designed to provide municipal information by phone. It includes the set of modules required to automatically recognize users' utterances, understand their meaning, decide the following response and generate a speech response. The different functionalities include to consult information about the City Council, access city information, carry out several steps and procedures, complete surveys, access citizen's mailbox to leave messages for suggestions and complaints, and be transferred to the City Council to be attended by a teleoperator. The voice portal is, therefore, pioneer in offering an extensive and comprehensive range of user-centered services accessible through speech, creating a new communication channel which is useful, efficient, and easy to use. The paper describes the application software, architecture, and infrastructures required for its operation 24 hours a day.
Keywords: Voice Portals, Spoken Dialog Systems, VoiceXML, User Adaptation.

1 Introducción

Gracias a que la voz es un medio natural e intuitivo para interactuar y comunicarse, las aplicaciones basadas en sistemas de diálogo (Pieraccini, 2012) se han convertido en una de las opciones para facilitar la interacción con dispositivos electrónicos. Estos programas informáticos tienen como principal finalidad interactuar con los usuarios oralmente o de forma multimodal para proporcionarles información o un determinado servicio de forma automática. El número de dominios de aplicación actuales de estos sistemas es enorme.

Para la implementación de los sistemas de diálogo que proporcionan información en la web, el World Wide Web Consortium (W3C) propone el uso del estándar VoiceXML (Will, 2012). Este lenguaje de programación permite la interacción persona-máquina integrando funcionalidades como la síntesis de texto-a-voz, reproducción de audio, reconocimiento del habla y de tonos DTMF

* Trabajo parcialmente financiado por los proyectos TRA2011-29454-C03-03, MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02 y CAM CONTEXTS (S2009/TIC-1485).

ISSN 1135-5948

(Dual-Tone Multi-Frequency), grabación de voz, control de flujo de diálogo y funciones de telefonía. VoiceXML es además una tecnología independiente de la plataforma que permite la portabilidad y transferencia de datos entre aplicaciones heterogéneas. Su utilización de manera conjunta con otros estándares y lenguajes de programación proporciona una base sólida para el desarrollo de sistemas de diálogo.

En este trabajo proponemos la utilización del estándar VoiceXML para el desarrollo de portales de voz que proporcionen información municipal. En este contexto, la Ley 11/2007 sobre el acceso electrónico de los ciudadanos a los Servicios Públicos incide en el fomento de múltiples canales de acceso a la información como una de las principales obligaciones de los Ayuntamientos, y reconoce explícitamente el derecho de los ciudadanos a relacionarse con las Administraciones Públicas por medios electrónicos.

No obstante el desarrollo de portales de voz municipales en España es actualmente muy limitado. De hecho, las pocas aplicaciones existentes actualmente ofrecen únicamente el acceso a una grabación para la elección de una determinada área o sección entre un conjunto reducido, a la que el usuario es redirigido para ser atendido personalmente por un operador del Ayuntamiento (por ejemplo, el portal desarrollado para el Cabildo de Gran Canaria).

Adicionalmente, existen otros tipos de acceso oral a la información municipal. Por ejemplo, en la página Web del Ayuntamiento de Santander se puede descargar una aplicación que permite navegar oralmente por las páginas que integran el portal del Ayuntamiento, además de poder escuchar el contenido de estas páginas web con una voz sintetizada. Otros ejemplos del uso de voz sintetizada son el portal web del Ayuntamiento de Zaragoza y del Ayuntamiento de Alicante.

De este modo, el portal de voz que se describe en este artículo es pionero en ofrecer una extensa y completa oferta de servicios municipales accesibles mediante el habla a través de un número de teléfono, así como en cuanto a las posibilidades ofrecidas para la adaptación del servicio proporcionado, tal y como se detalla en la siguiente sección.

2 Portal de voz desarrollado

El desarrollo de la aplicación se ha resuelto utilizando una arquitectura cliente-servidor cuyo esquema se puede observar en la Figura 1. La interacción con el sistema comienza cuando el usuario inicia una llamada, bien mediante la línea telefónica o bien mediante cualquier cliente de VoIP (por ejemplo, Skype). La interacción con el usuario y la provisión de las diferentes funcionalidades se lleva a cabo gracias a la disposición de un servidor VoiceXML y de servidores web.

En el servidor VoiceXML (para nuestro proyecto, la plataforma Voxeo Evolution[1]), el intérprete de VoiceXML se encarga por un lado de responder las llamadas de los usuarios y, por otro, de interpretar los documentos VoiceXML para ofrecer el servicio al usuario. También es el encargado de solicitar los recursos necesarios para la ejecución de la aplicación, de seguir la lógica del servicio y de mantener el estado de sesión de los usuarios actuando en consecuencia. Voxeo permite crear una aplicación VoiceXML y acceder a ella a través de distintos medios, ya que proporciona un número de teléfono local (según país y provincia en el caso de España) y un número Skype para llamadas desde la aplicación. Para el desarrollo del portal de voz se ha utilizado las tecnologías Prophecy Premium ASR y TTS[2].

En cuanto al servidor web utilizado, éste alberga las diferentes páginas con información dinámica programadas utilizando el lenguaje PHP, así como las bases de datos MySQL que contienen la información estática de la aplicación. Se ha considerado información estática a aquella que no cambia con el paso del tiempo, o al menos no lo hace en un tiempo considerable. Este tipo de información se recopila de páginas web, principalmente de la página web del Ayuntamiento de Alcorcón, y se almacena, perfectamente clasificada, en la base de datos de la aplicación. Cuando el usuario solicita este tipo de información, el sistema accede a ella en la base de datos y la devuelve encapsulada en un fichero VoiceXML. Ejemplos de este tipo de información son la historia de Alcorcón, los accesos a la ciudad, los datos de contacto de un hotel de la ciudad o de la oficina de empleo del municipio.

[1] evolution.voxeo.com

[2] help.voxeo.com/go/help/evolution.platforms.chooseplat.eu

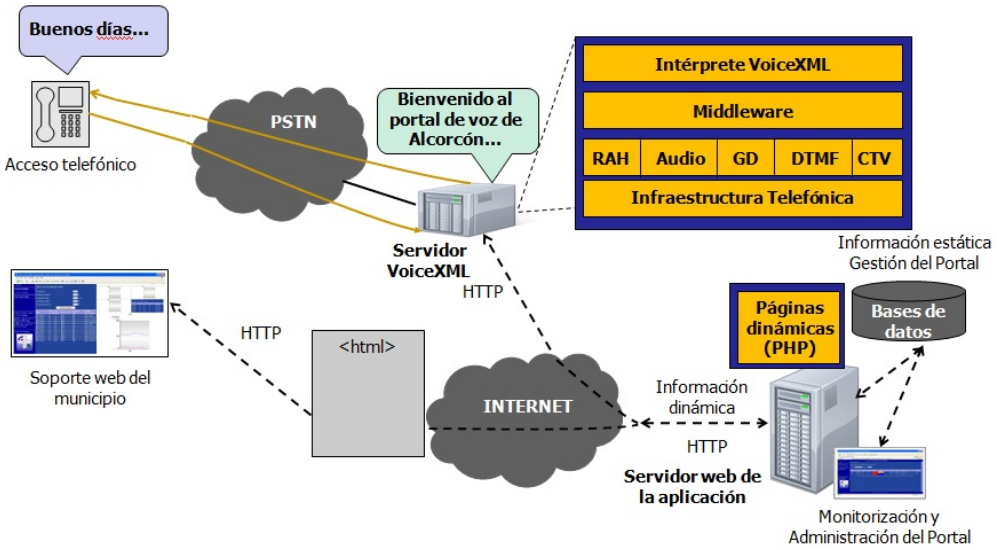


Figura 1: Arquitectura del portal de voz desarrollado

La información dinámica utilizada por la aplicación, en lugar de recargarse periódicamente en las bases de datos de la aplicación, se consulta directamente de portales web externos y se procesa sintácticamente a partir de páginas en PHP. Ejemplos de este tipo de información son las noticias y los eventos municipales, la información meteorológica, la encuesta municipal y la cartelera de los cines de Alcorcón. Las gramáticas dinámicas utilizan información que varía con el tiempo y suelen tratar gran cantidad de datos. Si se quisieran incluir en los ficheros VoiceXML de forma manual habría que modificarlas cada poco tiempo, tarea que sería muy laboriosa debido al gran número de ítems de los que constan. Ejemplos de estas gramáticas son las diseñadas para consultar información sobre los comercios municipales, acceder a las páginas amarillas o rellenar las encuestas disponibles.

2.1 Funcionalidades de la aplicación

El módulo Inicio inicia la interacción con el usuario. Las opciones que éste puede elegir se reparten en 5 módulos bien diferenciados según el tipo de interacción y de datos que se van a proporcionar: información, gestiones y trámites, encuesta, buzón del ciudadano, y operador. Por tanto, es en este módulo inicio donde se bifurca el diálogo entre el resto de módulos. Las principales acciones que se llevan a cabo en él son reproducir un mensaje de bienvenida al usuario, posibilitarle la selección de idioma y darle a elegir la acción que desea realizar. Una vez que el usuario elija una de estas opciones, se continúa la rutina en el fichero principal correspondiente a esta opción elegida, ya almacenado en el servidor web externo.

En el módulo Información el usuario puede acceder a toda la información del municipio de Alcorcón. La información, según su tipo, se divide en 6 grandes submódulos bien diferenciados y clasificados de tal forma que facilitan el acceso del usuario a la información que esté buscando:

- Ayuntamiento: Proporciona toda información relacionada con el Equipo de Gobierno, los Órganos de Gobierno y las Áreas de Gobierno municipales.
- Ciudad: En este submódulo se accede a la información referente a Alcorcón como ciudad. Se pueden consultar los datos de la ciudad, la historia, los accesos, y unas páginas amarillas compuestas por los bares, cafés, restaurantes, tiendas, hostales, hoteles y cines (incluida la información sobre la cartelera) del municipio.
- Áreas temáticas: Cualquier otro tipo de información que busque el usuario se ha repartido y clasificado en una de las 15 áreas que forman esta sección. Para cada una de estas áreas se proporciona información general, competencias y

datos de contacto. Simplemente con la incorporación de más información estática en la base de datos se podrían añadir más funcionalidades y áreas en este módulo de la aplicación.

- Noticias: Proporciona las noticias del municipio. Se facilita la fecha, título y subtítulo de cada noticia.
- Eventos: Reproduce el listado de eventos del municipio. Se proporciona el área temática, título, fecha, lugar y descripción de cada evento.
- Información meteorológica: El usuario puede obtener la información meteorológica actual del municipio y la previsión para los siguientes dos días.

Mediante el módulo Gestiones y Trámites el usuario puede verificar si está incluido en un listado, comprobar el estado de sus expedientes, reservar una instalación municipal o pedir cita para ser atendido en un servicio municipal.

Otra de las opciones del portal de voz es la realización de encuestas. De esta forma, se puede recoger de forma fácil y rápida la opinión de los ciudadanos sobre algún tema en particular que se plantee sobre el municipio. La encuesta se realiza de forma totalmente anónima, no solicitándose ningún tipo de identificación. Una vez que el usuario haya terminado de contestar, se le da la opción de escuchar los resultados parciales almacenados hasta el momento.

En el módulo Buzón del Ciudadano se implementa la funcionalidad de grabar un mensaje de voz por parte del usuario y que este mensaje sea clasificado y almacenado para su posterior tramitación. De esta forma, el ciudadano a cualquier hora y desde cualquier punto, puede hacer llegar al Ayuntamiento sus solicitudes, quejas, reclamaciones o comentarios. Además, en el caso en el que el ciudadano deje sus datos, ya sea teléfono fijo, móvil o correo electrónico, se puede contactar con él para dar una respuesta personalizada a su solicitud.

Finalmente, en el módulo Tele-Operador se transfiere la llamada del usuario al teléfono de la centralita del Ayuntamiento de Alcorcón.

Uno de los aspectos fundamentales en el desarrollo del portal de voz ha consistido en la incorporación de diferentes funcionalidades para adaptar el sistema al estado actual del diálogo y a características específicas de cada usuario. El primero de los puntos es relativo al tratamiento de los diferentes eventos proporcionados por VoiceXML. El portal almacena además en la base de datos de gestión de la aplicación los números de teléfono desde los que se accede al mismo, así como las diferentes funcionalidades de la aplicación que se han consultado en cada una de ellas. Esta información es utilizada por el sistema para asignar cuáles son las preferencias de los usuarios, en cuanto a consultas previas más frecuentes.

3 Conclusiones

En este artículo se ha descrito un portal de voz desarrollado mediante el estándar VoiceXML para proporcionar información municipal. Los servicios ofrecidos por el portal de voz municipal, distribuidos en diferentes módulos por los cuales se conduce al usuario en función de las decisiones que vaya tomando durante el diálogo, permiten que la aplicación del portal para proporcionar estos servicios en la ciudad de Alcorcón pueda considerarse como un avance considerable con respecto al desarrollo actual de estos portales en España.

Las líneas de trabajo que se desarrollan actualmente están relacionadas con la incorporación de funcionalidades adicionales en los diferentes módulos descritos, así como la mejora e incorporación de nuevas técnicas que amplíen los mecanismos descritos para la adaptación del portal. Para ello, se desea incorporar perfiles de usuario que incluyan información más detallada que facilite el uso de la aplicación, sirva además para adaptar la información proporcionada en los mensajes del sistema y disminuya los errores que se pudieran producir durante la interacción.

Bibliografía

Pieraccini, R. 2012. *The Voice in the Machine: Building Computers That Understand Speech.* MIT Press.

Will, T. 2012. *Creating a Dynamic Speech Dialogue: How to implement dialogue initiatives and question selection strategies with VoiceXML agents.* AV Akademikerverlag.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 189-192 recibido 14-04-14 revisado 02-07-14 aceptado 02-07-14

imaxin|software: PLN aplicada a la mejora de la comunicación multilingüe de empresas e instituciones

imaxin|software: NLP applied to enhance multilingual communications for public organisms and companies

José Ramom Pichel, Diego Vázquez, Luz Castro, Antonio Fernández
imaxin|software
Rua Salgueirinhos de abaixo N11 L6, Santiago de Compostela
e-mail: {jramompichel,diegovazquez,luzcastro,afernandez}@imaxin.com

Resumen: imaxin|software es una empresa creada en 1997 por cuatro titulados en ingeniería informática cuyo objetivo ha sido el de desarrollar videojuegos multimedia educativos y procesamiento del lenguaje natural multilingüe. 17 años más tarde, hemos desarrollado recursos, herramientas y aplicaciones multilingües de referencia para diferentes lenguas: Portugués (Galicia, Portugal, Brasil, etc.), Español (España, Argentina, México, etc.), Inglés, Catalán y Francés. En este artículo haremos una descripción de aquellos principales hitos en relación a la incorporación de estas tecnologías PLN al sector industrial e institucional.
Palabras clave: Big Data, Recursos lingüísticos, Análisis de Sentimientos, Minería de Opiniones, Traducción automática, Servicios online con herramientas PLN de código abierto, Aprendizaje de idiomas asistidos por ordenador.

Abstract: imaxin|software is a company created in 1997 by four computer engineers with the aim of developing educational multimedia games and natural language processing tools. After 17 years **imaxin**|software has developed resources, tools and applications for different languages, specially for Portuguese (Galiza, Portugal, Brazil, etc.), Spanish (Spain, Argentina, México, etc.), English, Catalan, French. In this article we will describe the main highlights of this technological and human challenge.
Keywords: Big Data, Language Resources, Sentiment Analysis, Opinion Mining, Machine Translation, Online services using Open-source NLP tools, Computer Aided Language Learning.

1 Introducción

imaxin|software es una empresa dedicada al desarrollo de servicios y soluciones avanzadas de software y multimedia desde el año 1997, especializada en ingeniería lingüística y videojuegos multimedia educativos y formativos (Serious Games, Gamification) (Pichel et al., 2013).

imaxin|software que inicialmente estaba constituída por cuatro socios-trabajadores, tuvo en plantilla hasta veintiseis personas en el año 2010. Las ventas de productos y servicios se han repartido entre público y privado en porcentajes aproximadas de 60 %-40 % variando de año en año entre un sector y otro. Nos centraremos en la primera línea de desarrollos, **imaxin**|software es desde el año 2000 proveedor de tecnología lingüística para Microsoft. Además, podemos destacar entre los principales desarrollos en PLN los sistemas de corrección ortográfica, gramatical, estilística; sumarizadores de textos, sistemas de opinion mining, pesquisa semántica, sistemas de codificación médica de historias clínicas, detección automática de entidades (NER), así como la plataforma líder europea en traducción automática de código abierto: Opentrad (con sus de los motores de traducción Apertium y Matxin) – www.opentrad.com.

ISSN 1135-5948

2 Principales proyectos PLN aplicados a las necesidades empresariales

2.1 Construcción y uso de recursos linguísticos

- **Corrección ortográfica en red Galgo.NET (2001)**:

  El corrector imaxin Galgo.NET es uno de los primeros correctores desarrollados específicamente para la correción ortográfica multilingüe simultánea (gallego y español) para redacciones de periódicos. **imaxin**|software ha desarrollado toda esta tecnología incluyendo un tries propio para compresión de diccionarios (Malvar y Pichel, 2010).

- **Corrección de lenguaje sexista (Exeria)**:

  Hemos desarrollado un corrector para OpenOffice.org que mejora los textos en gallego ofreciendo textos con lenguaje no sexista. **imaxin**|software ha desarrollado toda la tecnología de corrección estilística e integración en Openoffice.org.

- **Coruxa Biomedical Text Mining**:

  Extractor y codificador automático de información médica relevante mediante el uso del PLN. Financiado por la Dirección Xeral de I+D+i (Xunta de Galicia). Investigador principal: **imaxin**|software, USC-GE, IXA Taldea, Doctor QSolutions. Transferencia al sector industrial: servicio de codificación SNOMED-CT para historias clínicas. **imaxin**|software ha desarrollado toda la tecnología de codificación. IXA Taldea ha desarrollado un anonimizador de historias clínicas y la USC-GE el procesamiento de ontologías.

2.2 Optimizadores semánticos de búsquedas

- **Optimizador de búsquedas en bibliotecas mediante ontología (2008)**:

  El objetivo del módulo Optimizador es expandir las búsquedas efectuadas por los/as usuarios/as en los sistemas de consulta bibliográfica del CSBG (Centro Superior Bibliográfico de Galicia) mediante el uso de ontologías construída adhoc a partir de un corpus construído de bibliografía. Está integrado con el software de gestión de bibliotecas de código abierto Koha. **imaxin**|software ha desarrollado toda la tecnología.

2.3 Servicios online mediante el uso de herramientas de PLN de código abierto

- **Traductor de documentos online (www.opentrad.com)**:

  Existe un servicio en línea de e-commerce de Opentrad para traducir documentos entre diferentes lenguas y manteniendo en todo momento el formato original. **imaxin**|software ha desarrollado toda la tecnología web.

- **Traductor de documentos en la Aplicateca de Telefónica**:

  Desde el año 2012 está instalado el traductor de documentos Opentrad especial para PYMES en la Tienda Cloud Aplicateca de Telefónica. **imaxin**|software ha desarrollado toda la tecnología de integración en Aplicateca.

2.4 Traducción automática de código abierto

- **Opentrad: plataforma de servicios de traducción de código abierto (2004-2014)**

  Opentrad (Alegría et al., 2006) es la plataforma de traducción automática en código abierto pionera en el mercado español (www.opentrad.com). Este proyecto se inició en el año 2004, siendo el resultado de diferentes proyectos de I+D+i (PROFIT y Avanza del Ministerio de Industria) desarrollados por un consorcio formado por Universidades y Empresas (Transducens-UA, Eleka, Elhuyar, IXA Taldea, TALP (UPC), **imaxin**|software y SLI-Universidade de Vigo). Como resultado de este proyecto se constituyó una spin-off especialista en uno de los motores del proyecto (Apertium), Prompsit Language Technologies.

  Opentrad está formada por dos ingenios de traducción de código abierto (Apertium y Matxin). Opentrad mejora la comunicación multilingüe, permite publicar información en diferentes idiomas, reduce costes y tiempos de revisión humana, permitiendo incluso la mejora de

los tiempos en la localización de versiones multilingües de aplicaciones empresariales.

Opentrad está o estuvo implantado en administraciones, empresas y portales de Internet traduciendo millones de palabras diariamente (Ministerio de Administraciones Públicas, Xunta de Galicia, Universidades Públicas Gallegas, La Voz de Galicia, Faro de Vigo, Instituto Cervantes, Kutxa, Eroski, etc.)

La mejoría continua de los de los ingenios de traducción (Apertium y Matxin) permite ofrecer sobre todo, una mejor calidad entre lenguas próximas (Español-Francés Español-Portugués, Español-Portugués do Brasil, Español-Catalán, Español-Gallego, etc.) que otros traductores automáticos.

imaxin|software ha desarrollado todas las tecnologías para integrar los prototipos de Opentrad en cliente final y la mejora lingüísticas de los recursos de traducción automática especificamente para los pares español-galego, español-portugués y español-inglés (Pichel et al., 2009).

2.5 Análisis de sentimientos y minería de opinión para un seguimiento de marca inteligente y análisis Big Data

En este campo hemos desarrollado en el año 2009 un prototipo inicial (Coati) de análisis de sentimientos. En la actualidad, hemos trasladado esta experiencia a un proyecto más ambicioso relacionado con el Análisis de sentimientos y minería de opinión relacionado con el Big Data. Explicaremos cada uno de ellos en detalle:

- **Coati Opinion mining (2009)**:

  En este proyecto hemos investigado como extraer automáticamente de blogs opiniones y tendencias interesantes para el ámbito empresarial y la administración pública (2009) mediante el uso de técnicas de Opinion Mining. **imaxin**|software ha desarrollado toda la tecnología del crawler y el corpus de entrenamiento del Opinion Mining basado en support vector machine (Malvar y Pichel, 2011). Pendiente de evaluación.

- **CELTIC: Conocimiento Estratégico Liderado por Tecnologías para la Inteligencia Competitiva (FEDER-INNTERCONECTA)**:

  El proyecto, actualmente en desarrollo, está orientado al campo de la vigilancia tecnológica y el Social Media Marketing mediante el uso de PLN y el procesamiento en Big Data.

  Este proyecto ha sido financiado mediante los fondos tecnológicos europeos para regiones objetivo 1 de la Unión Europea. Estos fondos conocidos como FEDER INNTERCONECTA, son proyectos Integrados de desarrollo experimental altamente competitivos, con carácter estratégico, de gran dimensión y que tienen como objetivo el desarrollo de tecnologías nuevas en áreas tecnológicas de futuro con proyección económica y comercial a nivel internacional, suponiendo a la vez un avance tecnológico e industrial relevante para las autonomías destinatarias de las ayudas, como es el caso de Galicia.

  imaxin|software consiguió en el año 2012 este proyecto con un consorcio formado polas siguientes empresas y Universidades: Indra, Elogia, SaecData, Gradiant, USC-PRONATL (USC), Computational Architecture Group (USC).

  El objetivo del proyecto es el desarrollo de tecnologías capacitadoras que faciliten al tejido empresarial la toma de decisiones estratégicas en tiempo casi-real, a partir del conocimiento tanto del medio científico-tecnológico como de los impactos económicos presentes y futuros. O lo que es el mismo, el desarrollo de tecnologías capacitadoras para la Inteligencia Competitiva en las organizaciones.

  Las tecnologías a desarrollar durante el proyecto cubren el proceso completo de la Inteligencia Competitiva, en sus respectivas fases: agregación de información, análisis de la información extrayendo de ella el conocimiento necesario, y la distribución mediante mecanismos de visualización e iteracción avanzados para facilitar la toma de decisiones estratégicas.

  El ámbito de aplicación es el Social Media Marketing y la Vigilancia tecnológi-

ca. En el primero, la competitividad actual genera la necesidad de disponer de sistemas de monitorización inteligente y en tiempo real de redes sociales y análisis del impacto de los productos de una marca determinada en el consumidor (Gamallo, Garcia, y Pichel, 2013). Esto puede ser posible mediante la integración de tecnologías avanzadas de procesamiento del lenguaje natural y tecnologías semánticas.

En el campo de la Vigilancia tecnológica los los desarrollos a realizar en este proyecto permitirán el acceso y gestión en tiempo real de los conocimientos científicos y técnicos a las empresas, así como la información más relevante sobre su contexto, junto a la comprensión a tiempo del significado e implicaciones de los cambios y novedades.

imaxin|software ha desarrollado en colaboración con Indra, la USC-GE y USC-CA todos los desarrollos de PLN integrados en Big Data. Todas los desarrollos están pendientes de evaluación al final del proyecto.

2.6 Aprendizaje de Lenguas asistido por Ordenador (Juegos y Lexicografía)

Por último hemos desarrollado el "Portal das palabras" en el año 2013, una web educativa que pone en valor el diccionario de la Real Academia Galega mediante juegos relacionados con las palabras para un mejor aprendizaje del gallego por el público en general y sectores más distantes de la lengua como el mundo empresarial.

Con el Portal de las Palabras no solo podemos mejorar nuestra competencia en idioma gallego sino que también aprenderemos jugando. Incluye también el diccionario de la RAG con búsquedas de lemas y sinónimos, videos explicativos y guías didácticas para la lengua.

imaxin|software ha desarrollado toda la tecnología PLN y web en este proyecto.

3 Conclusiones

Este artículo pretende mostrar por un lado un mosaico de tecnologías PLN (productos, servicios y proyectos de I+D) de más de 17 años de una pequeña empresa, y por otro la importancia que para este fin ha tenido la colaboración y transferencia con los organismos públicos de investigación (Universidades y Centros Tecnológicos).

Bibliografía

Alegría, I., I. Arantzabal, M Forcada, X. Gómez-Guinovart, L. Padró, J. R. Pichel, y J. Waliño. 2006. OpenTrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, 37:357–358.

Gamallo, P., M. Garcia, y J. R. Pichel. 2013. A method to lexical normalisation of tweets. En *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural. Workshop on Sentiment Analysis at SEPLN*, páginas 81–85.

Malvar, P. y J. R. Pichel. 2010. Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. the galician and portuguese case. En *Internacional de Lingüística de Corpus (CILC10)*, páginas 529–536.

Malvar, P. y J. R. Pichel. 2011. Métodos semiautomáticos de generación de recursos de opinion mining para el gallego a partir del portugués y el español. *Novática: Revista de la Asociación de Técnicos de Informática*, 214:61–64.

Pichel, J. R., P. Malvar, O. Senra, P. Gamallo, y A. García. 2009. Carvalho: English-galician SMT system from english-portuguese parallel corpus. *Procesamiento del Lenguaje Natural*, 43:379–381.

Pichel, J. R., D. Vázquez, L. Castro, y A. Fernández. 2013. 16 anos desenvolvendo aplicações no campo do processamento da linguagem natural multilingue. *Linguamática*, 5(1):13–20.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 193-196 recibido 15-04-14 revisado 02-07-14 aceptado 02-07-14

Integration of a Machine Translation System into the Editorial Process Flow of a Daily Newspaper

Integración de un sistema de traducción automática en el entorno de redacción de un periódico

Juan Alberto Alonso Martín, Anna Civil Serra
Lucy Software Ibérica SL
c/ Copèrnic 44, 1r 08021-Barcelona
juan.alonso, anna.civil@lucysoftware.com

Resumen: El artículo describe el proceso de integración del traductor automático de Lucy Software en el entorno de redacción de *La Vanguardia*, donde se utiliza diariamente como herramienta auxiliar para publicar una edición bilingüe del diario en catalán y en castellano. Este proceso de integración incluye adaptaciones técnicas y lingüísticas, y un proceso final de post-edición.
Palabras clave: traducción automática, post-edición, integración de la traducción automática en procesos productivos

Abstract: This paper describes the integration process of Lucy Software's machine translation system into the editorial process flow of *La Vanguardia* newspaper, where it is used on a daily basis as a help-tool in order to produce bilingual editions of the daily newspaper in Catalan and Spanish. The integration process includes both technical and linguistic adaptations, as well as a final post-edition process.
Keywords: machine translation, post-edition, integration of machine translation into productive processes.

1 Introduction

Established in 1881, *La Vanguardia* is the leading daily newspaper in Catalonia and the fourth best-selling in Spain, with a daily circulation of over 200.000 copies. It is widely recognized as a quality newspaper

In 2010 *La Vanguardia* decided to prepare a parallel edition in Catalan, which was officially launched on May 3rd 2011. In order to be able to do this parallel edition in Catalan, they decided to use post-edited Machine Translation (MT), and after surveying possible candidates, they finally chose Lucy Software's MT system.

This paper describes different aspects on the integration of the Lucy LT machine translation system into the editorial process flow of *La Vanguardia*. This integration involved both linguistic and IT aspects. You can find more details on this integration in Vidal and Camps (2012).

2 The Lucy LT MT System

Lucy LT is a rule-based machine translation system which is the ultimate successor of the old METAL MT system. Lucy LT is a transfer-based MT system with an island chart parser and three translation phases: analysis, transfer and generation. In each of these phases, and for each language-direction, computational grammars – one analysis grammar, one transfer grammar and one generation grammar -, and computational lexicons – source and target language monolingual lexicons and one source-to-target transfer lexicon – are used. Lucy LT runs on Windows workstations and has a number of APIs (e.g. Web Services) that allow integrating the system within external applications and

ISSN 1135-5948

workflows. For more details, please refer to Schwall and Thurmair (1997).

3 The Challenge

Whatever the final solution was, the following general requirements had to be met:

- One daily copy of *La Vanguardia* includes over 60.000 words, all of them to be translated, revised and post-edited.
- Both editions should be ready for printing every day at 23:30 the latest.
- The Catalan edition should comply with the linguistic requirements stated in the Style Guide of La Vanguardia.
- Even though most journalists at La Vanguardia write in Spanish, which was the base edition at the time, out of which the Catalan edition was to be created, at short/mid-term every journalist should be free to write in the language of his/her choice (Catalan or Spanish), so that, actually, after some time, there should be no base edition.
- Both the MT-system and the post-edition environment should be completely integrated into their editorial flow (both IT-integration and human team integration).

4 Possible Solutions

Given the task of making bilingual daily editions of a newspaper, three possible options could be considered:

4.1 The *MT-less* Option

This option would imply using no MT at all. This would imply:

- Duplicating the whole editorial human team or/and hire a team of N human translators to translate the entire newspaper content on time in order to keep both editions synchronized for publishing.
- Duplicating most of the IT infrastructure (Content Management System, etc.)

Given these factors, the question arises of whether it would be feasible to produce bilingual editions of a newspaper this way because of dramatic increase of costs and very tight time constraints.

This approach was therefore rejected.

4.2 The *full-MT* Option

This option would imply using only MT, without any post-editing phase. This means running all the contents of the (Spanish) base edition through an MT translation system and publishing the raw MT-translation of the original contents in the Catalan language edition.

It was immediately clear that this was not an option because, even for language-pairs for which the quality of MT is very high (as it is the case for Spanish-Catalan, where a quality higher than 95% can be achieved), the output mistakes would be unacceptable for publishing: proper nouns being translated, homographs, etc. Moreover, the Catalan style coming out from the MT system would not always sound "natural" to Catalan speakers.

This approach was also rejected.

4.3 The Sensible-MT Option

This option implied using a customized MT-system and a team of human post-editors. This option implied:

- Customizing the MT-system grammars and lexicons to the specific linguistic needs of La Vanguardia (style guide, corporate terminology, proper nouns, etc.).
- Integrating the MT-flow within the newspaper editorial flow (document and character formats, connection to a post-edition environment, feedback processing, etc.)
- Incorporating a post-edition environment to be used by a team of human post-editors into the editorial flow.

Here we have a compromise between the MT-use (time and effort saving) and the translation quality, so this was the approach that was finally chosen.

5 The Solution

The solution that was finally adopted by *La Vanguardia* implied the following general aspects:

5.1 Pre-launch Phase

There was a pre-launch ramp-up phase during which computational linguists from Lucy Software, post-edition experts, and part of the editorial team from *La Vanguardia* worked together for six months in order to

- Customize the MT-system to the linguistic requirements posed by *La Vanguardia* (as far as possible). This linguistic customization implied that over 20.000 lexical entries had to be added/changed in the MT-system lexicons and many grammar rules had to be adapted in the MT-system grammars, mainly for the Spanish→Catalan direction in a first phase.
- Integrate the MT-system into their IT editorial environment. This integration included:
    - The integration of our MT-system with *La Vanguardia*'s HERMES CMS.
    - Enabling Lucy Software's MT system to be able to handle *La Vanguardia*'s specific character format and XML tags.
    - Inclusion of markups in the MT-output specifically designed for post-editors
    - Configuring the MT-system installation so that translation performance could meet the expected translation load & peak requirements.
- Last, but not least, a team of around 15 persons were trained on post-editing the MT-output before publishing, and the corresponding shifts and work-flow for these post-editors was organized.

5.2 Post-launch Phase

In the post-launch phase, the lexicons and grammars of the MT-system continued to be adapted to the news that were translated every day in the Spanish→Catalan direction. Adaptation works also started for the Catalan→Spanish direction, also both in the lexicons and in the grammars of the system, in order to enable *La Vanguardia* journalists to write in the language of their choice (i.e., Spanish or Catalan).

This post-launch phase lasted for some six months right after the launch of the Catalan edition of the newspaper.

5.3 Maintenance Phase

The maintenance phase started right after the final of the post-launch phase and involves ongoing maintenance works, mainly in the computational grammars of both directions, Spanish→Catalan and Catalan→Spanish.

Previous to this, a training session was done with personnel of the newspaper's editorial team in order for them to get familiar with the lexicon coding tool of the Lucy MT system. Therefore, during this maintenance phase they are taking care of the system lexicons and Lucy Software is responsible of providing at least two annual updates of the computational grammars, where a number of reported errors have been fixed.

Beside the computational lexicons and grammars, the system has a so-called pre- and post-editing filters which allow the users to define strings that should not be translated (typically proper nouns). These filters are maintained by the staff of *La Vanguardia*, with the technical support of Lucy Software.

5.4 Examples of Linguistic Adaptations

Most of the MT lexicons adaptations that have been carried out for *La Vanguardia* correspond to

- Specialized lexicon entries on very specific domains:
    - Bullfight: albero/arena, morlaco/toro (bull)
    - Castellers (human towers): **cinc de vuit amb folre i manilles** (human tower of eight levels of five persons each), **pila de set**, etc.
- Proper noun lists, including lists of place names (villages, rivers, mountains, etc.), well-known person names (**Leo Messi**, **Rodríguez Zapatero**, etc.), etc.
- Latin words and expressions (**in dubio pro reo**, **tabula rasa,** etc.).
- New words (neologisms) or fashion words (Spanish/Catalan): **dron**/**dron** (drone), **bitcoin**/**bitcoin**, **autofoto/autofoto** (selfie), **crimeano**/**crimeà** (Crimean), **watsap**/**watsap** (a Whatsapp message))
- Words that appear often at *La Vanguardia*: **perroflauta/rastaflauta** (anti-system young person), **cantera azulgrana/planter blaugrana** (Barcelona F.C. team), **iniestazo/iniestada** (a score from Andrés Iniesta), **arena política** (political arena), **stajanovista/estakhanovista**.
- Idioms or colloquial language: **tartazo/cop de pastís** (pie hit), **hacer un corte de mangas/fer botifarra** (a rude gesture somehow similar to a *two-finger salute*), **cocinillas/cuinetes** (*kitchen wizard*, sometimes said in a derogatory sense).

6 Post-Edition

As already mentioned, the post-edition phase is a key factor in the final output quality of the Catalan edition. The post-editors typically work from 17:00 to 23:00. Their main goal is to revise and eventually correct mistakes that may appear in the MT output, and – also very important – give the *human flavor* to the MT Catalan output, whenever it is possible because of time constraints. The reason for this last point is that often, the MT output can be perfectly correct from a grammatical point of view but, still, sound a little *awkward or artificial* to a native speaker, in the sense that s/he would never use these words or construction to express this idea. The post-editor task is then to paraphrase the output sentence with a more *natural* wording. Again, because of time constraints, this is typically done in news headlines, where, in addition, more often than not puns can be used in the source language that are impossible to be correctly translated (or actually, localized) into the target language by an MT system.

7 Conclusions

The conclusions of this project can be summarized as follows: producing two parallel bilingual editions of a daily newspaper only seems to be feasible if the following three conditions are met:

- MT is used in the process,
- The MT-system is properly customized, adapted and integrated to the newspaper linguistic and IT requirements,
- There is a team of trained specialized human post-editors who correct MT mistakes and "give the human flavor" to the output.

References

Bernardi, U., Bocsak, A & Porsiel, J, 2005.: *Are we making ourselves clear? Terminology management and machine translation at Volkswagen.* Proceedings of the 10th EAMT conference "Practical applications of machine translation", 30-31 May 2005, Budapest*;* pages 41-49.

Schwall, U. & Thurmair G., 1997. *From METAL to T1: systems and components for machine translation applications*. Proceedings of the MT Summit VI. Machine Translation Past, Present, Future. Proceedings, 29 October – 1 November 1997, San Diego, California, USA; pages 180-190.

Vidal, B. & Camps, M., 2012: *Catalan Daily Goes Catalan* (www.localizationworld.com/lwparis2012/presentations/files/A4.pdf) presented at Localization World 2012, Paris.

Wolf, P., & Bernardi, U, 2013: *Hybrid domain adaptation for a rule based MT system.* Proceedings of the XIV Machine Translation Summit, Nice, September 2-6, 2013; ed. K.Sima'an, M.L.Forcada, D.Grasmick, H.Depraetere, A.Way; pages.321-328.

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 197-200 recibido 15-04-14 revisado 30-06-14 aceptado 02-07-14

track-It! Sistema de Análisis de Reputación en Tiempo Real

track-It! Real-Time Reputation Analysis System

Julio Villena-Román
Janine García-Morera
Daedalus, S.A.
Av. de la Albufera 321
28031 Madrid, España
{jvillena, jgarcia}@daedalus.es

José Carlos González-Cristóbal
Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, España
jgonzalez@dit.upm.es

Resumen: Este artículo presenta un sistema automático para recoger, almacenar, analizar y visualizar de manera agregada información publicada en medios de comunicación sobre ciertas organizaciones junto con las opiniones expresadas sobre ellas por usuarios en redes sociales. Este sistema permite automatizar la elaboración de un análisis de reputación completo y detallado, según diferentes dimensiones y en tiempo real, permitiendo que una organización pueda conocer su posición en el mercado, medir su evolución, compararse con sus competidores, y detectar lo más rápidamente posible situaciones problemáticas para ser capaces de tomar medidas correctoras.
Palabras clave: Reputación, extracción de información, análisis semántico, análisis de sentimiento, clasificación, opinión, redes sociales, RSS.

Abstract: This paper presents an automatic system to collect, store, analyze and display aggregated information published in mass media related to certain organizations together with user opinions about them expressed in social networks. This system automates the production of a complete, detailed reputation analysis, in real time and according to different dimensions, allowing organizations to know their position in the market, measure their evolution, benchmark against their competitors, and detect trouble situations to be able to take early corrective actions.
Keywords: Reputation, information extraction, semantic analytics, sentiment analysis, classification, opinion, topics, social networks, RSS.

1 Introducción

La reputación corporativa es el conjunto de percepciones que tienen sobre una organización todos los grupos de interés implicados: clientes, empleados, accionistas, proveedores, etc. Se ve afectada por todas las noticias sobre la organización en medios de comunicación y las opiniones, recomendaciones, etc. (beneficiosas o perjudiciales) de usuarios en redes sociales.

La gestión de esta información se convierte en algo cada vez más valioso, ofreciendo la oportunidad de responder a las expectativas sobre la organización y estar más protegidas frente a crisis eventuales. Pero el volumen de contenido es tan grande que las tecnologías de análisis automático se hacen indispensables para poder procesar toda esta información.

Este artículo describe track-It!, un sistema automático para recoger, almacenar, analizar y visualizar de manera agregada opiniones recogidas en Internet. Existen sistemas similares en el mercado, enfocados a vigilancia de marca (SproutSocial, comScore, Engagor, etc.) o al análisis de la voz del cliente (Vocus, Customerville, etc.).

El objetivo final es conocer "qué, cómo y cuánto" se dice de la organización en tiempo real, y medir cómo afecta esta información a su reputación y a la de sus competidores, actuando en consecuencia.

[1] Este trabajo ha sido financiado por los proyectos Ciudad2020: Hacia un nuevo modelo de ciudad inteligente sostenible (INNPRONTA IPT-20111006) y MA2VICMR: Mejorando el Acceso, el Análisis y la Visibilidad de la Información y los Contenidos Multilingüe y Multimedia en Red para la Comunidad de Madrid (S2009/TIC-1542).

2 *Arquitectura del sistema*

El sistema se compone de los cinco componentes mostrados en la Figura 1. El **recolector** de información es el responsable de capturar de manera continua y en tiempo real la información proveniente de diferentes fuentes de Internet. El **analizador** utiliza técnicas de procesamiento del lenguaje natural para procesar semánticamente la información. La información obtenida se almacena en el **datawarehouse**. El módulo de **agregación** selecciona, filtra, ordena y consolida toda la información referida a una misma organización y la analiza de forma global, obteniendo los valores de reputación agregados. Finalmente el componente de **visualización** presenta la información de forma gráfica e interactiva.

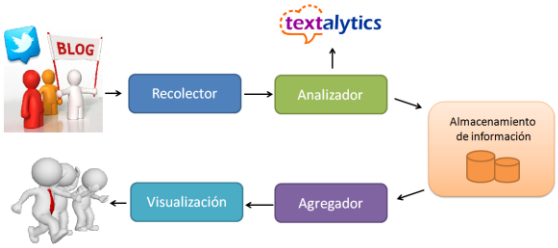


Figura 1: Arquitectura del sistema

3 *Recolector*

El recolector comprueba con una cierta periodicidad si en una serie de fuentes configuradas se ha recibido nuevo contenido. Se utilizan peticiones a la API de Twitter[2] para recolectar tweets con el nombre de la organización o alguno de sus alias. Además se utiliza un lector RSS para recoger las noticias publicadas en medios de comunicación relevantes.

4 *Analizador*

El analizador procesa y extrae información relevante de todo el contenido recogido de la red, empleando las funcionalidades lingüísticas de procesamiento y análisis de texto que ofrece Textalytics[3], nuestro portal de servicios lingüísticos en la nube:

1) Detección de idioma para filtrar el idioma de interés, si la fuente no lo incluye.
2) Clasificación automática, para determinar de qué aspectos del modelo de clasificación reputacional se habla (ética empresarial, organización de la empresa, trato a clientes).
3) Análisis de sentimientos para determinar si la polaridad del mensaje es positiva o negativa para la organización.
4) Detección de entidades para establecer la entidad a la que se refiere el texto y evitar problemas de desambiguación.

El objetivo es almacenar información clasificada sobre todos los aspectos que pueden afectar a la reputación de la empresa.

4.1 Detección de idioma

Se utiliza la detección automática de idioma cuando no se conoce el idioma a priori o no se puede extraer de la fuente. El detector se basa en técnicas estadísticas basada en la representación de N-gramas y es capaz de detectar hasta 60 idiomas distintos.

4.2 Modelo de clasificación

El módulo de clasificación automática de textos utiliza un modelo previamente entrenado para determinar la temática de la información. El algoritmo de clasificación utilizado consiste en un modelo híbrido que combina un procedimiento de clasificación estadístico con un filtrado basado en reglas (Villena-Román et al., 2011). El resultado es una lista de las categorías más representativas, ordenadas de mayor a menor relevancia. El clasificador ha sido evaluado con diferentes modelos de ámbito general, por ejemplo, Reuters-21578, en el que se obtienen precisiones superiores al 80%.

Categoría	Subcategorías
10 Oferta	11 Satisfacción necesidades
	12 Gestión de reclamaciones
	13 Trato a clientes
	14 Relación calidad/precio
	15 Calidad de productos y servicios
	16 Garantía de productos y servicios
20 Trabajo	21 Capacidad de los empleados
	22 Bienestar de los empleados
	23 Igualdad de oportunidades
	24 Remuneración a empleados
30 Integridad	31 Ética
	32 Asuntos corporativos
	32 Transparencia
40 Estrategia y liderazgo	41 Organización y estructura
	42 Liderazgo
	43 Visión y estrategia
50 Innovación y flexibilidad	51 Innovación
	52 Adaptabilidad al cambio
60 Responsabilidad social	61 Contribución a la sociedad
	62 Apoyo de causas sociales
	63 Protección del medio ambiente
70 Situación financiera	71 Potencial de crecimiento futuro
	72 Gestión financiera
	73 Resultados financieros

Figura 2: Ontología del modelo de reputación

El sistema emplea el modelo de reputación mostrado en la Figura 2. Para mejorar la

[2] https://dev.twitter.com/docs/api/1.1
[3] http://textalytics.com

precisión de la clasificación, se han desarrollado dos variantes de este modelo, una para textos largos (para noticias), entrenado estadísticamente con textos de entrenamiento e incluyendo en las reglas los términos empresariales más utilizados en cada categoría, y otro para fragmentos de texto cortos, en el que la clasificación se basa fundamentalmente en el filtrado mediante reglas muy específicas.

4.3 Análisis de sentimientos

El módulo de análisis de sentimientos permite determinar si la opinión o el hecho expresado en el texto es positivo, negativo o neutro, o bien si no expresa sentimiento. El análisis se basa en la información incluida en un modelo semántico que incluye reglas y recursos etiquetados (unidades con polaridad, modificadores).

Además, se utiliza un análisis morfosintáctico del texto para dividir el texto en segmentos, controlar mejor el alcance de las unidades semánticas del modelo, detectar la negación e identificar entidades. Se utiliza un algoritmo de agregación (basado en el uso de medias y desviaciones típicas que incluye la detección de valores *outliers*) para calcular el valor de la polaridad global del texto a partir de la polaridad de los diferentes segmentos y el de la polaridad final de las entidades y conceptos, a partir del valor de cada una de sus menciones.

El sistema ha sido evaluado en foros competitivos obteniendo valores de medida-F superiores al 40% y siendo el mejor sistema de los presentados (Villena-Román et al., 2012).

4.4 Extracción de entidades

La extracción de entidades se lleva a cabo mediante procedimientos de extracción de información basados en análisis morfosintáctico y semántico del texto, apoyado en recursos lingüísticos y reglas heurísticas, permitiendo identificar los elementos significativos. La salida será el listado de entidades encontradas junto a su información asociada (como el tipo de entidad). Evaluaciones internas sitúan al sistema en niveles de precisión y cobertura similares a otros sistemas existentes.

5 Almacenamiento de información

La información se almacena en un repositorio centralizado. Actualmente se utiliza una base de datos relacional (MySQL) pero el *roadmap* de producto plantea la migración del sistema a Elasticsearch (Elasticsearch, 2014), por sus características de tratamiento de grandes volúmenes de datos, eficiencia de búsquedas, escalabilidad y soporte a fallos.

6 Agregación de la información

La reputación de una organización en cada una de las categorías del modelo se calcula aplicando un algoritmo de agregación similar al del módulo de análisis de sentimientos, sobre todo el conjunto de textos comprendidos en un rango de fechas y provenientes de una o varias fuentes. Además, a partir de la reputación por categorías se obtiene la reputación general de la organización, teniendo en cuenta que las diferentes categorías no están interrelacionadas.

7 Visualización

Finalmente se muestra toda la información mediante tablas y gráficos.

Se ha desarrollado un escenario piloto de seguimiento de empresas del IBEX-35, recogiendo información en Twitter y en los medios económicos Expansión, Invertia, Intereconomía, El Economista y Cinco Días. En este caso sólo son de interés las entidades de tipo COMPANY o COMPANY GROUP, para resolver ambigüedades (p.ej., descartar "Santander" si se refiere a la ciudad).

Inicialmente la interfaz web muestra la información de la reputación de cada una de las empresas durante el día actual. Sobre esta tabla se pueden aplicar filtros: por rango temporal (día, semana o mes), tipo de empresas (tecnológica, financiera) o por el origen de la fuente de datos (Twitter o noticias RSS).

Figura 3: Información general

Por ejemplo, la Figura 3 muestra la información de las entidades financieras: Bankia, BBVA, Santander y Bankinter. La tabla presenta la reputación general y por categorías

y el número de menciones. Los colores indican la polaridad: rojo oscuro (N+), rojo (N), amarillo (NEU), verde (P) y verde oscuro (P+).

A partir de ella, se puede obtener información de detalle. La Figura 4 muestra la información específica de Bankinter: el número de menciones y la reputación asociada por cada categoría detectada. El gráfico circular indica la distribución de los textos por categoría y el gráfico de barras muestra la distribución de la polaridad de los textos en cada categoría. Además se puede obtener información de detalle de cada texto concreto.

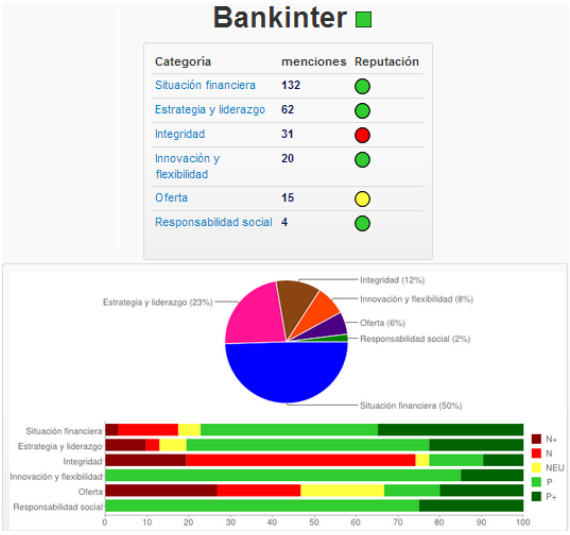


Figura 4: Información detallada

El sistema también ofrece información comparativa de las entidades seleccionadas, general y para cada categoría (Figuras 5 y 6).

Figura 5: Vista comparativa general

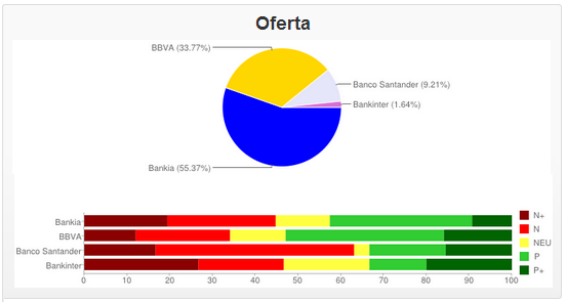


Figura 6: Vista comparativa de una categoría

8 *Conclusiones y Trabajos Futuros*

El sistema automatiza la recogida de información en la red y permite la elaboración de análisis de reputación. Se encuentra en fase beta y en proyecto de implantación en diferentes escenarios. En el piloto del IBEX35 se han recogido más de 10,2 millones de piezas de información desde agosto de 2013, con 500 mil de entidades y 17 millones de etiquetas.

De cara a un futuro sería interesante ampliar su funcionalidad a un mayor número de idiomas, recoger diferentes tipos de fuentes, por ejemplo, blogs o sitios web de opiniones, y calcular la reputación considerando no sólo las menciones de la empresa, sino también sus productos y servicios.

También se podría incluir la relevancia que puede tener un mensaje, estableciendo prioridades. Esto permitiría detectar alertas y dar avisos a la empresa para que puedan gestionar problemas rápidamente.

Bibliografía

Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press.*

Villena-Román, J., S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J.C. González-Cristóbal. 2012. DAEDALUS at *RepLab 2012: Polarity Classification and Filtering on Twitter Data. CLEF 2012 Labs and Workshop Notebook Paper.*

Elasticsearch.org. Open Source Distributed Real Time Search & Analytics. 2014. [En línea] *http://www.elasticsearch.org*

Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 201-204 recibido 15-04-14 revisado 02-07-14 aceptado 02-07-14

Aplicación de tecnologías de Procesamiento de lenguaje natural y tecnología semántica en Brand Rain y Anpro21

Natural language processing and semantic technologies. The application on Brand Rain and Anpro21

Oscar Trabazos, Silvia Suárez, Remei Bori y Oriol Flo
Anpro21 y Brand Rain
Carrer de Barcelona, 2
17002, Girona
{otc,ssb,rem}@anpro21.com
oflo@bluemapconsulting.com

Resumen: Este artículo presenta la aplicación y resultados obtenidos de la investigación en técnicas de procesamiento de lenguaje natural y tecnología semántica en Brand Rain y Anpro21. Se exponen todos los proyectos relacionados con las temáticas antes mencionadas y se presenta la aplicación y ventajas de la transferencia de la investigación y nuevas tecnologías desarrolladas a la herramienta de monitorización y calculo de reputación Brand Rain.
Palabras clave: Procesamiento de lenguaje natural, web semántica, ontologías, minería de datos, análisis de reputación, análisis de sentimiento, aprendizaje automático.

Abstract: This paper presents the application and results on research about natural language processing and semantic technologies in Brand Rain and Anpro21. The related projects are explained and the obtained benefits from the research on this new technologies developed are presented. All this research have been applied on the monitoring and reputation system of Brand Rain.
Keywords: Natural language processing, semantic web, ontologies, data mining, text mining, reputation analysis, sentiment analysis, machine learning.

1 Introducción

Somos una empresa de tecnología que investiga técnicas de Inteligencia artificial (IA) y de procesamiento del lenguaje natural (PLN) para desarrollar herramientas y servicios para el campo del marketing, la comunicación y el branding.

Con una trayectoria de casi 10 años a las espaldas, en 2010 lanzamos al mercado uno de los software de análisis de la reputación de marca más completos de Europa, se trata de Brand Rain.

Una meta-herramienta de monitorización y análisis de la reputación en la cual se aplica la labor de investigación que desarrolla Anpro21 y que ya incorporan en su día a día centenares de gabinetes de comunicación y empresas de sectores bien diversos.

Brand Rain permite capturar, monitorizar y analizar toda la información que se publica respecto a una marca, entidad o persona, medir la reputación y el sentimiento de estas informaciones y determinar la reputación de que goza la marca en los medios off-line (prensa, radio y televisión) on-line y en las redes sociales. Precisamente, esta es una de las particularidades de Brand Rain, tres herramientas en una, para analizar todos los impactos de una marca mediante un software de uso sencillo e intuitivo.

En lo que se refiere al campo de I+D+i centramos nuestras líneas de investigación en el área de la Inteligencia Artificial, PLN, técnicas de indexación, recuperación de la información y Big data. Este artículo pretende exponer los proyectos que llevamos a cabo, las

ISSN 1135-5948

tecnologías que usamos y el cómo aplicamos las técnicas a nuestro software Brand Rain.

2 Estado del arte

En los últimos años, las aportaciones que se han hecho desde el PLN han mejorado sustancialmente, permitiendo el procesamiento de ingentes cantidades de información en formato texto con un grado de eficacia aceptable. Muestra de ello es la aplicación de estas técnicas como una componente esencial en los motores de búsqueda web, en las herramientas de traducción automática, o en la generación automática de resúmenes.

Para el desarrollo de PLN se han usado diferentes herramientas y analizadores sintácticos, semánticos y de dependencias. Un analizador sintáctico (o parser) es una de las partes de un compilador que transforma su entrada en un árbol de derivación. Algunos ejemplos de analizadores sintácticos y de dependencias son Freeling [Lluis Padro, 2013], Gate [Sonal Gutap, 2014], Stanford Parser [H. Cuniham, 2012].

3 Proyectos de I+D+i de Anpro21

Actualmente trabajamos en 3 grandes proyectos, en colaboración con la Universitat de Girona, la Universidad de Salamanca y consolidadas empresas del sector tecnológico como Ibermática y Verbio Technologies.

3.1 Vídeo-reputación

El proyecto de video-reputación persigue el objetivo final de desarrollar un sistema de análisis de los contenidos multimedia que sea capaz de interpretar imágenes, vídeos y voz para determinar su temática, su significado y su tono para finalmente poder cuantificar y determinar la reputación corporativa de las empresas en internet.

Esta plataforma permite monitorizar y analizar el contenido audiovisual mediante técnicas de interpretación de la voz, el audio, los gestos y el tono que se desprenden de los mensajes.

Para este proyecto se aplican técnicas de IA relacionadas con reconocimiento de patrones en imágenes y vídeos, reconocimiento automático de voz y tono o modulación de la voz y procesamiento de lenguaje natural.

El desarrollo de técnicas de análisis audiovisual nos permitirá también acercarnos a la detección de la ironía y el sarcasmo, uno de los grandes retos de las herramientas de monitorización y análisis de la reputación.

3.2 Redes complejas

El proyecto de redes complejas tiene que ver con la detección de crisis de reputación o amenazas en la red. Mediante tecnologías de redes complejas llegamos a establecer, localizar y definir una red para la detección de personas que influencian la reputación de una marca hasta dar con el núcleo de la red, el influenciador clave.

Esto será de gran ayuda para la gestión de crisis de reputación y también tendrá otras aplicaciones como pueden ser la detección de amenazas en la red, como el acoso escolar o la violencia de género.

Para este proyecto se aplican tecnologías de minería de texto, machine learning, boosting de autor y procesamiento de lenguaje natural.

3.3 Find Your Fund (FUF)

Find your Fund está diseñado para enlazar emprendedores y empresas en búsqueda de financiación con inversores mediante técnicas de procesamiento del lenguaje natural. Para llevarlo a cabo se usan analizadores sintácticos (Freeling, Gate), ontologías y minería de textos.

4 Aplicaciones de las tecnologías en Brand Rain

Los distintos proyectos de investigación desarrollados por Anpro21 nutren continuamente el sistema de Brand Rain, y es el mismo software el que muestra las necesidades de seguir investigando para perfeccionar y afinar en el análisis de la reputación y en la detección de la ironía y el sarcasmo, el gran caballo de batalla del sector.

Por lo tanto, las tecnologías desarrolladas en Anpro21 se incorporan al software Brand Rain como nuevas funcionalidades que permiten actualizar y perfeccionar el servicio que da esta herramienta día a día a los profesionales de la comunicación.

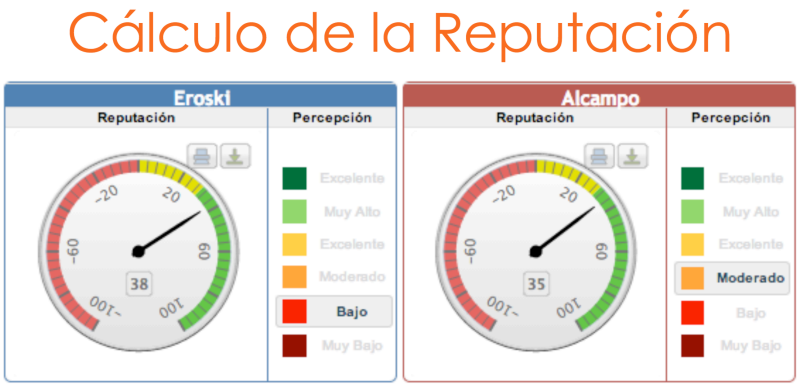


Figura 1: Cálculo de la reputación tal y como se muestra en Brand Rain.

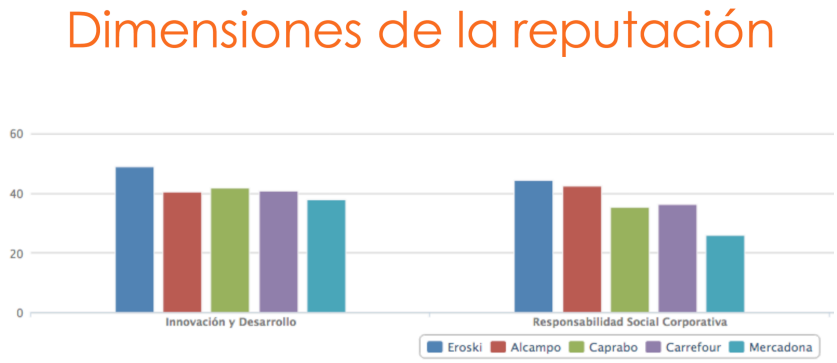


Figura 2: Dos dimensiones de la reputación. Innovación y desarrollo y responsabilidad social.

En este capítulo vemos las aplicaciones de las tecnologías mencionadas en el capítulo anterior en dos de las funciones más destacadas de Brand Rain, el cálculo de la reputación y la detección de influenciadores.

4.1 Cálculo de la reputación y análisis de sentimiento

En Brand Rain el cálculo de la reputación (Figura 1) de una marca se realiza a partir del análisis de las menciones, noticias y conversaciones capturadas sobre la misma. Para llevar a cabo este análisis contamos con un sistema de diseño propio que se encarga del estudio semántico y del contexto que envuelve cada una de las noticias. Para el análisis de las menciones utilizamos algoritmos de análisis de texto, de PLN (Analizadores sintácticos y morfológicos) y patrones lingüísticos.

Estos algoritmos permiten examinar semánticamente las frases, mensajes o noticias capturadas y detectar dónde se habla de la marca que queremos estudiar y sobretodo el cómo se habla.

Mediante el análisis de sentimiento sabemos si una mención es positiva (se habla bien de la marca), negativa (se habla mal de la marca) o neutra (el mensaje no se puede considerar ni positivo ni negativo para la marca). Para llevarlo a cabo contamos con una serie de ontologías de sentimiento que se basan en un estudio científico realizado a más de 20000 personas y adaptado en el cual se hace un ranking de las palabras con una puntuación de -10 a 10, siendo -10 muy negativo y 10 muy positivo. Esta ontología cuenta con mas de 15.000 palabras rankeadas y etiquetadas con una puntuación para cada sentimiento. Aun así, está ontología se adapta y personaliza a las necesidades de cada cliente y sector con métodos de machine learning.

Para ir más allá del análisis de sentimiento, Brand Rain contempla en sus cálculos otros

análisis semánticos que tienen en cuenta el universo corporativo de cada marca. Este universo abarca conceptos que entendemos como dimensiones de la reputación de marca como el desempeño financiero, el respeto al medio ambiente, la calidad de los productos, la inversión en I+D+i o la responsabilidad social corporativa. (Figura 2). Las dimensiones tienen alto valor para objetivar la reputación de marca y intervienen en el cálculo final.

El análisis se realiza utilizando técnicas de análisis de sentimiento, PLN, machine learning y ontologías de sentimiento.

4.2 Detección de influenciadores y redes complejas

La detección de la red de influenciadores de una marca se realiza mediante tecnología de redes complejas. En primer lugar, se reconocen los autores que publican contenido sobre las marcas y todos sus alias posibles. Esto se hace mediante técnicas de boosting y reconocimiento de patrones en texto.

Una vez identificados los autores y medios más destacados para la marca, el software aporta datos sobre el autor y su capacidad de impactar al público.

En un futuro, está previsto que ofrezca también datos sobre el grado de peligrosidad del autor dependiendo de cómo habla de la marca y que llegue a dibujar toda la red de personas que influencian la marca hasta dar con el influenciador clave.

De esta forma, analizando la actividad del influenciador clave y el sentimiento de los mensajes que emite se podrá llegar a detectar el foco de una crisis de reputación y actuar en consecuencia para evitarla.

5 Conclusiones y valoraciones

La investigación desarrollada en PLN, representan una ventaja competitiva para Brand Rain. Nuestros clientes se han visto beneficiados al poder utilizar una herramienta cada vez más inteligente, y adecuada a sus necesidades, permitiendo un análisis de reputación y de marca con características muy avanzadas, que reúne los beneficios de la IA y de la aplicación de PLN.

Sin embargo, este es un tema retador y existen tópicos como la ironía y aspectos más complejos relacionados con la lingüística, que nos impulsa a querer investigar cada vez mas para la mejora continua de nuestros sistemas.

En definitiva, son muchos los beneficios obtenidos tras la aplicación del PLN en nuestros sistemas. Sin duda los más destacados son aquellos que nos han permitido abrir nuevas vías de negocio. Entre éstos está el análisis de sentimiento, con el que hemos podido analizar aspectos cualitativos de una marca, como el tono contextual del que está rodeada, o la reputación de la misma tal y como la entenderíamos las personas.

Nuestra herramienta analiza diariamente más de 600.000 menciones y hace un seguimiento exhaustivo de más de 1.000 marcas, la aplicación del PLN ha supuesto poder realizar análisis semántico, que junto con los algoritmos de big-data representan una gran mejora para empresas como la nuestra que gestionan un enorme volumen de datos.

Bibliografía

H. Cunningham, V. Tablan, A. Roberts, K.Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Comput Biol,* 9(2):e1002854.doi: 10.1371/ journal.pcbi.100284.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. ELRA. (Istanbul, Turkey).

Sonal Gupta and Christopher D. Manning 2014. SPIED*:* Stanford Pattern-based Information Extraction and Diagnostics. En *Proceedings of the ACL 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces ACL-ILLVI.*

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/revistaSEPLN/Instrevista.php

Hoja de Inscripción para Instituciones

Datos Entidad/Empresa

Nombre : .....
NIF : ..... Teléfono : .....
E-mail : ..... Fax : .....
Domicilio : .....
Municipio : ..... Código Postal : ..... Provincia : .....
Áreas de investigación o interés: .....
.....

Datos de envío

Dirección : ..... Código Postal : .....
Municipio : ..... Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

Sr. Director de:

Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....
Núm Cuenta : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

.....

Cuotas de los socios institucionales*: 300 €.*

Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Hoja de Inscripción para Socios

Datos Personales

Apellidos : .....
Nombre : .....
DNI : ..... Fecha de Nacimiento : .....
Teléfono : ..... E-mail : .....
Domicilio : .....
Municipio : ..... Código Postal : .....
Provincia : .....

Datos Profesionales

Centro de trabajo : .....
Domicilio : .....
Código Postal : ..... Municipio : .....
Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....
Áreas de investigación o interés: .....
.....

Preferencia para envío de correo:

[ ] Dirección personal [ ] Dirección Profesional

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

En.....a.....de.....de.....
(firma)

Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN

Sr. Director de:
Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

Cuotas de los socios*: 18 €* (residentes en España) o *24 €* (socios residentes en el extranjero).
Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

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
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:

José Gabriel Amores	Universidad de Sevilla
Toni Badía	Universitat Pompeu Fabra
Manuel de Buenaga	Universidad Europea de Madrid
Irene Castellón	Universitat de Barcelona
Arantza Díaz de Ilarraza	Euskal Herriko Unibertsitatea
Antonio Ferrández	Universitat d'Alacant
Mikel Forcada	Universitat d'Alacant
Ana García-Serrano	UNED
Koldo Gojenola	Euskal Herriko Unibertsitatea
Xavier Gómez Guinovart	Universidade de Vigo
Julio Gonzalo	UNED
José Miguel Goñi	Universidad Politécnica de Madrid

Cartas al director

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

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://www.sepln.org/revistaSEPLN/revistas.php
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/revistaSEPLN/edirevista.php
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/revistaSEPLN/lectrevista.php

Demostraciones y Artículos de la Industria

Información General