sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Artículos

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA 2015

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

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

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2016

Editores:

M. Teresa Martín Valdivia	Universidad de Jaén	maite@ujaen.es
Eugenio Martínez Cámara	Technische Universität Darmstadt	camara@ukp.informatik.tu-darmstadt.de
M. Carlos Díaz Galiano	Universidad de Jaén	mcdiaz@ujaen.es
Salud María Jiménez Zafra	Universidad de Jaén	sjzafra@ujaen.es
L. Alfonso Ureña López	Universidad de Jaén	laurena@ujaen.es
Patricio Martínez Barco	Universidad de Alicante	patricio@dlsi.ua.es

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

Consejo asesor

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

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

Patricio Martínez-Barco	Universidad de Alicante (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Raquel Martínez	Universidad Nacional de Educación a Distancia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lidia Moreno	Universidad Politécnica de Valencia (España)
Lluís Padró	Universidad Politécnica de Cataluña (España)
Manuel Palomar	Universidad de Alicante (España)
Ferrán Pla	Universidad Politécnica de Valencia (España)
German Rigau	Universidad del País Vasco (España)
Horacio Rodríguez	Universidad Politécnica de Cataluña (España)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé	Universidad de Barcelona (España)
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignoon / Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maillo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Iñaki Alegría	Universidad del País Vasco (España)
Miguel Anxo Solla Portela	Universidad de Vigo (España)
Maxux Aranzabe	Universidad del País Vasco (España)
Arantza Casillas	Universidad del País Vasco (España)
Francis De la Caridad Fernández Reyes	Universidad Autónoma del Estado de Morelos (México)
Manuel Carlos Díaz Galiano	Universidad de Jaén (España)
Miguel Ángel García Cumbreras	Universidad de Jaén (España)
Salud María Jiménez Zafra	Universidad de Jaén (España)
Marina Lloberes	Universidad de Barcelona (España)
Adrián-Pastor López Monroy	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Eugenio Martínez Cámara	Technische Universität Darmstadt (Alemania)
Esaú Villatoro	Universidad Autónoma Metropolitana (México)

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preámbulo

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

Las áreas temáticas tratadas son las siguientes:

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

El ejemplar número 57 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a tres apartados diferenciados: comunicaciones científicas, resúmenes de

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

proyectos de investigación y descripciones de aplicaciones informáticas (demostraciones). Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista. Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 40 trabajos para este número de los cuales 26 eran artículos científicos y 14 se correspondían a resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas. De entre los 26 artículos recibidos 14 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 53,8%.

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. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Septiembre de 2016
Los editores

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA 2015

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 57th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, research project summaries and description of Natural Language Processing software tools. All

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

of these were accepted by a peer reviewed process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Fourteen papers were submitted for this issue of which twenty-six were scientific papers and fourteen were either projects or tool description summaries. From these twenty-six papers, we selected fourteen (53.8%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation 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 were supervised by the editorial board. The cut-off criteria adopted was the average of the three scores given.

September 2016
Editorial board

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Artículos

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

Proyectos

Demostraciones

Información General

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Lexicografía y Terminología Computacionales

Procesamiento de Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 15-22 recibido 20-03-2016 revisado 18-04-2016 aceptado 03-05-2016

Gestión de resúmenes para dispositivos móviles

Managing summaries for mobile devices

Fernando Llopis, José M. Gómez, Elena Lloret, Patricio Martínez, Yoan Gutiérrez
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{llopis,jmgomez,elloret,patricio,ygutierrez}@dlsi.ua.es

Resumen: Los dispositivos móviles han cambiado notablemente la forma en la que los usuarios acceden a la información disponible en Internet. Estos dispositivos permiten un acceso instantáneo desde cualquier lugar, pero tienen una serie de limitaciones importantes sobre los ordenadores personales. Su limitada pantalla, así como en ocasiones la limitada capacidad de recepción de la información, dado el coste, hacen que la selección de información a acceder sea todavía más importante. La generación automática de resúmenes multi-documentos es una alternativa de interés que es el objeto de este artículo. Así en este artículo se presentan y evalúan un modelo de generación automática de resúmenes, junto con un sistema de recuperación de información basado en pasajes.
Palabras clave: Gestión de resúmenes, Recuperación de Información, Recuperación de Pasajes, Turismo, COMPENDIUM, IR-n

Abstract: Mobile devices have significantly changed the way users access the information available on Internet. These devices allow instant access anytime and anywhere, but they have a number of important limitations with respect to personal computers. The limited screen space and, sometimes, the limited capacity to receive the information, make the selection of information even more important. Automatic summary generation from multi-document summarization is an interesting alternative which is the subject of this paper. Therefore, in this article is presented and evaluated a model of automatic summarization with an information retrieval system based on passages.
Keywords: Summarization, Information Retrieval, Passage Retrieval, Tourism, COMPENDIUM, IR-n

1 Introducción

La evolución de los hábitos y necesidades de los usuarios de Internet ha sido una constante. Una vez que los usuarios podían acceder a millones de documentos, el principal problema a resolver fue el de cómo facilitar la búsqueda de los documentos relevantes para el usuario entre todos ellos. La aparición de los buscadores como Yahoo, Google, Bing y muchos más resolvió, de una forma razonablemente satisfactoria, este problema. Una persona sentada cómodamente frente a la pantalla del ordenador podía realizar una búsqueda de documentos. El esquema de prácticamente todos los buscadores es el de proporcionar una lista de referencias a 10/20 páginas de internet acompañados con una selección de fragmentos de textos incluidos en ellas y que suelen contener palabras de la consulta del usuario.

La siguiente tarea del usuario era la de ir accediendo a aquellas páginas que podía considerar relevantes, en base a esos fragmentos de texto que le mostraba el buscador. El proceso se convertía en una repetición de estas tareas hasta que el usuario daba por satisfechas sus necesidades de información, o si decidía refinar su búsqueda inicial dado el poco éxito de la misma.

Los principales objetos de la investigación en la recuperación de información se basaban en ser lo más eficientes y eficaces a la hora de seleccionar los documentos relevantes ante las consultas del usuario. Pero algo ha cambiado notablemente en la forma en los dispositivos

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

con los que los usuarios acceden a Internet[1]. Los dispositivos móviles son uno de los medio más utilizado para el acceso a la información. La principal ventaja de los dispositivos móviles es que prácticamente están disponibles en cualquier momento. Los principales inconvenientes a la hora de facilitar el procesado de la información son dos: el primero el tamaño de la pantalla, cuatro o cinco veces más pequeña que los monitores que suelen acompañar a los ordenadores de sobremesa o portátiles; el segundo son las limitaciones con las que la información puede llegar al dispositivo móvil. La progresiva aparición de las redes 4G ha solucionado en parte el problema de la escasa velocidad en accesos a la información desde un dispositivo móvil, pero siguen sin solucionar el coste todavía alto que cobran las compañías por la transmisión de información a los móviles. Hay otro aspecto que introdujo el llamado Internet 2.0, y es la importancia del usuario no solo como receptor pasivo de la información sino como generador activo de la misma a través de opiniones. Así, suelen ser más valoradas las aportaciones que los usuarios realizan sobre un tema, que lo que podría ser considerada la versión oficial. Esto ha supuesto la aparición de innumerables páginas que complementan la información con opiniones de todo tipo acerca de algún elemento.

Así una búsqueda del tipo "Iglesias Alicante" en la conocida página de viajes Trip Advisor devolvería un resultado como en la Figura 1.

Figura 1: Búsqueda de Iglesias Alicante en Trip Advisor

Ahora le tocaría al usuario navegar a través de los dos enlaces que se proponen. Al seleccionar en una de las primeras, se pueden acceder a las opiniones que los usuarios han ido escribiendo sobre la misma como se puede ver en la Figura 2.

Figura 2: Opiniones de usuarios en aplicación móvil

Para facilitar la búsqueda a los usuarios se trataría de conseguir en primer lugar obtener información relevante ante una petición del usuario y en vez de mostrarle ordenados una serie de documentos o enlaces en función de su relevancia, se propone realizar un procesamiento adicional a los documentos iniciales con el objeto de identificar la parte más relevante de los mismos y mostrar en la pantalla del dispositivo móvil del usuario un pequeño fragmento de texto que resuma o identifique partes clave de los documentos relevantes. Con un esquema parecido al que se muestra en la Figura 3.

Figura 3: Resumen de opiniones

La idea es resumir las opiniones de los usuarios en base a los términos más relevan-

[1] http://www.idc.com/getdoc.jsp?containerId=prUS40664915

tes en el total de las opiniones. Hay términos que se repiten como *bonita*, *Barroco*, *siglo XIV* y *horarios*. Los tres primeros son positivos y el último algo negativo por los problemas de acceso que indican algunos usuarios. Con esto se facilita al usuario decidir si un lugar merece la pena visitarlo leyendo tres o cuatro frases. Siempre tiene la opción, si así lo desea, de leer todas las opiniones, pero la opción de leer un resumen es cómodo y rápido, sobre todo si debe seleccionar por ejemplo que monumentos debe visitar con restricciones de tiempo.

El objetivo es ser capaz de obtener resúmenes desde diversos documentos. Para ello vamos a basarnos en los modelos propuestos en la generación automática de resúmenes cuya misión fundamental es la es reducir la dimensión de un texto manteniendo la información relevante del mismo, incluida dentro del ámbito del procesamiento del lenguaje natural (PLN).

Los modelos de resúmenes que presentaremos en este artículo se centran en los de tipo extractivo, los cuales son los que se generan a partir de la selección y extracción de frases del texto original. La frase es la unidad mínima de información con el significado suficiente para ser presentada al usuario y que éste pueda entender razonablemente su significado.

En nuestro trabajo se han realizado experimentos para analizar la forma óptima de utilizar sistemas de resúmenes pero que utilicen fuentes de texto diversas. Para llevar a cabo esto, consideramos todos los documentos de estas distintas fuentes como un único texto que, posteriormente, fue el que se resumió. Esta primera experimentación la hemos preferido realizar sobre una colección de textos y preguntas que va a ser evaluada dentro de las conferencias NTCIR.

La experimentación realizada en este artículo se basa en el uso combinado de un sistema de recuperación de Información basado en pasajes, el IR-n (Llopis and Vicedo, 2001) y el sistema de generación automática de resúmenes COMPENDIUM (Lloret and Palomar, 2012). Para las pruebas iniciales hemos utilizado la colección propuesta en la tarea Mobile Click del NTCIR-12.

El artículo se estructura del siguiente modo: en primer lugar se hace un análisis de la situación actual de los modelos de resúmenes de opiniones en la sección 2 para luego presentar el sistema generador de resúmenes automáticos COMPENDIUM y el sistema de recuperación de Pasajes IR-n en las secciones 3 y 4. Las secciones 5 y 6 describen la experimentación y resultados obtenidos por los sistemas. Finalmente, se exponen las conclusiones obtenidas en la sección 7.

2 Análisis del contexto actual

En la mayoría de los enfoques de sistemas de resúmenes de comentarios se integra un motor de minería de opiniones que tiene como objetivo fundamental identificar y clasificar las opiniones presentes en los comentarios diferenciando positivas, negativas o neutras.

Por ejemplo, el enfoque propuesto en (Kokkoras et al., 2008) genera múltiples resúmenes de opiniones, considerando la información de metadatos que suelen acompañar. El proceso de resumen incluye como primera etapa la creación de un diccionario que contiene el vocabulario específico del dominio. En una segunda etapa, las frases se puntúan con respecto a su relevancia utilizando métodos estadísticos basados en recuentos de frecuencias y la presencia de las palabras en el diccionario obtenido en la fase anterior.

En la tercera etapa, la puntuación asignada inicialmente a una sentencia se ajusta teniendo en cuenta la contribución de los metadatos. Los metadatos incluyen información que no está directamente presente en el texto de la revisión, pero pueden proporcionar información sobre la utilidad de la revisión, o el nivel técnico de la crítica, siendo por tanto útil para determinar la importancia de una oración.

En la etapa final, la calidad del resumen generado se mejora mediante la eliminación de frases redundantes. Esto se realiza mediante la comparación de las palabras en la frase resumen con las palabras de las frases restantes, y el recálculo de la puntuación de las frases disminuyendo la de aquellas que contienen palabras que ya aparecen en el resumen.

En el enfoque mencionado anteriormente, las opiniones son consideradas como un texto único. El problema es que una valoración puede contener diferentes opiniones con respecto a varios aspectos de ese tema. Por ejemplo, acerca de un hotel, un usuario puede escribir un comentario expresando que “el hotel estaba muy bien situado, pero el servicio fue un poco decepcionante”; o para un teléfono

móvil, los usuarios pueden expresar sus opiniones sobre su precio, su batería, su diseño, su pantalla, etc.

Ante esto, las técnicas de minería de opiniones comenzaron a desarrollar enfoques que teniendo en cuenta las diferentes características del producto, servicio, lugar, etc. Esto se conoce como multi-aspecto o la minería opinión multi-faceta, apareciendo también la tarea de resúmenes de opinión multi-aspecto, cuando el objetivo es resumir la revisión, en relación con las diferentes características. En (Abulaish et al., 2009) se genera el resumen teniendo en cuenta los resultados de un sistema de minería de opinión, y que representa de una manera visual, a través de un gráfico. El enfoque de minería de opiniones propuesto extrae diferentes características de la revisión, junto con los modificadores asociados a la opinión de que disponga para cada uno de los aspectos que considera (por ejemplo, el precio - excelente, alto, caro, barato). Tras haber evaluado la lista de las opiniones y modificadores asociados para cada característica extraída, se establece su polaridad usando (Baccianella and Sebastiani, 2010), asociando a cada elemento a tres puntuaciones numéricas, que lo definen como positivo o negativo. En (Ly et al., 2011), las diferentes características expresadas en los comentarios se llaman facetas. Estas facetas son identificados a través de una etapa de identificación de las facetas de productos, que comprende etiquetado gramatical, análisis de dependencias y el recuento de la frecuencia. Una vez que se han determinado las facetas de productos, el enfoque de integración se limita únicamente a las frases con opiniones (positivas o negativas), obtenidos a través de un enfoque de minería de opiniones. Estas frases de opinión resultantes se agrupan en función de su similitud contenido, distinguiendo entre frases positivas y negativas.

Finalmente se selecciona la frase más representativa de cada grupo para formar el resumen.

Pero en ocasiones las opiniones no marcan en sí aspectos positivos o negativos. El hecho de que un iglesia del siglo XIV sea bonita, es positivo, pero a su vez incorpora información acerca de que es del siglo XIV lo que puede ser positivo para un interesado por las iglesias de esa época y negativo para una persona que desea visitar edificios modernos.

El modelo que planteamos en este artículo se centra en el resumen de textos provenientes de varias fuentes, como puedan ser los comentarios pero que traten de responder a una petición concreta del usuario.

3 Sistema de resúmenes COMPENDIUM

COMPENDIUM es un sistema automático que es capaz de realizar resúmenes de uno o varios documentos. El sistema utiliza diferentes técnicas para la identificación, selección y extracción de la información más relevante. Para ello se utilizan cinco etapas en la que la salida de cada una de ellas es la entrada de la siguiente hasta producir el resultado final.

- **Análisis lingüístico**. Esta etapa consiste en el preproceso del texto, realizando un análisis lingüístico básico, utilizando herramientas y recursos externos. Este pre-procesamiento incluye la segmentación, tokenización, etiquetado y el análisis sintáctico, así como la identificación y borrado de stopwords.
- **Detección de redundancias**. El objetivo de esta etapa es identificar la información redundante en los documentos de origen, con el objeto de no incluirla en el resumen. Para este propósito, las técnicas de Textual Entailment (TE) han demostrado ser las adecuada para esta etapa, ya que pueden determinar si el significado de un fragmento de texto puede deducirse de otro (Glickman, 2006).
- **Identificación de temas principales**. El objetivo de esta etapa es determinar los temas principales del documento/s que se pretende resumir.

  En COMPENDIUM, los temas de un documento están representados por la frecuencia de los términos que contiene. A raíz de esta declaración, se supone que los términos más frecuentes de un documento son indicativos de los temas incluidos en él. Por lo tanto, las oraciones que contienen términos frecuentes se puntúan más alto tal como se comentará en la etapa de *detección de relevancia.*

  Sin embargo, es importante señalar que las "stopwords", que fueron identificadas previamente en fases anteriores no se tienen en cuenta para el cálculo de la frecuencia de un documento, por tanto, no

forma parte de los temas de un documento.

- **Detección de relevancia**. En esta etapa, se calcula y asigna un peso a cada frase, en función de su relevancia.

  Este peso toma la frecuencia de los términos calculada en la etapa anterior, y se combina con otra característica basada en *El principio de la cantidad de código* (Givón, 1990).

  El fundamento de esta teoría es que una determinada información en un texto, independientemente de que sea nueva o dada, debe recibir una codificación tal que resalte más o menos según el grado de relevancia que tiene esta información en el texto. Si la información es más relevante recibirá una "cantidad mayor de codificación", esta información se codificará con mayor peso léxico; y, al revés, si es menos relevante se utilizará una "cantidad menor" para su codificación de menor peso léxico.

  La puntuación total de la frase se calcula considerando la frecuencia de las palabras que forman parte de los sintagmas nominales que contiene una frase y la longitud de los mismos, de manera que frases que tengan sintagmas nominales con palabras más frecuentes obtendrán mayor relevancia respecto al resto.

- **Generación del resumen**. El objetivo de esta etapa es la de generar un resumen con una longitud específica. Esta longitud se expresa en forma de una tasa de compresión (es decir, el porcentaje de información que el resumen contiene con respecto al documento de origen). Dada una tasa de compresión, las frases más relevantes (es decir las que tienen una puntuación más alta) serán las seleccionadas para formar el resumen final hasta dicha longitud deseada. Con el fin de minimizar los posibles problemas con la coherencia, las frases seleccionadas se presentan en el mismo orden en que están en el documento de origen.

4 Sistema IR-N

Los sistemas de Recuperación de Pasajes (PR) son sistemas de Recuperación de Información que determinan la relevancia de un documento con respecto de una pregunta en base a la similitud de diferentes fragmentos (pasajes) de dicho documento con respecto a la misma pregunta. Estos modelos no solo permiten mejorar la localización de documentos relevantes sino que además permiten localizar con mayor exactitud la parte realmente relevante del documento (Kaszkiel and Zobel, 1997).

Los sistemas de PR se clasifican en función de cómo determinan los pasajes de cada documento. El sistema IR-n (Llopis and Vicedo, 2001) es un sistema de PR que define los pasajes en base a un número determinado de frases. Esto permite dotar a los pasajes de cierto contenido sintáctico.

Así el modelo de trabajo del sistema IR-n es el siguiente

1. Un documento se divide en frases. Cada pasaje puede estar formado por hasta un número fijo de frases. Para la experimentación realizada en este trabajo se ha trabajado en pasajes de una única frase.

2. La relevancia de un pasaje o frase p con respecto a una pregunta o query q se obtiene de la siguiente forma:

$$Relevancia = \sum_{t \in p \wedge q} W_{p,t} * W_{q,t} \qquad (1)$$

   Siendo:

   $W_{p,t} = log_e(f_{p,t} + 1)$,

   $f_{p,t}$ es el número de apariciones del término t en el pasaje. p,

   $W_{q,t} = log_e(f_{q,t} + 1) * idf$,

   $f_{q,t}$ es el número de apariciones del término t in pregunta q,

   $idf = log_e(n/f_t + 1)$,

   n es el número total de documentos de la colección

   f_t es el número de documentos donde aparece el término t

La fórmula es muy parecida a la definida en (Salton, 1989) y conocida como "la del Coseno". La principal diferencia es que en el sistema IR-n se omite los cálculos de normalización por el tamaño del documento, dado que estamos trabajando con frases, y se entiende que cada una de ellas es una unidad comparable a cualquier otra.

El planteamiento de utilizar el sistema IR-n como un sistema de generación de resúmenes se basa en el concepto de favorecer las frases que contienen los términos que se repiten más frecuentemente en todos los documentos. Para ello se plantea la propuesta de una vez seleccionados todos los documentos generar una pregunta con la concatenación de todos ellos y luego evaluar la relevancia para cada documento de forma individual con respecto a esa pregunta generada.

Por ejemplo la pregunta 1 era "*Iglesias en Alicante*" se lanzaría una búsqueda con un buscador web tradicional o en el caso de ejemplo se seleccionarían las opiniones de cada usuario de la web TripAdvisor. A cada uno de estos elementos les llamaremos iUnits.

- iUnit 1 Es la construcción religiosa más antigua de la ciudad, puesta esta data del siglo XIV y se encuentra asentada sobre el mismo lugar que antaño ocupaba una mezquita... para los alicantinos es la CONCATEDRAL. La visita es gratuita... pero mejor consultar horario.
- iUnit 2 Externamente la iglesia está bien, pero es una pena de no ponerla más al alcance de los turistas. No indican en el exterior los horarios, lo cual dice muy poco a favor de los cuidadores de dicho edificio, Debe de estar abierta de 11 a 12:30 y de 19 a 20:30, en los horarios de misa.
- iUnit 3 Alicante reúne varias iglesias realmente preciosas y la de Santa María es la que más me gusto, su fachada, sus imágenes procesionales, sus pinturas, altar mayor, capillas, techo, me pareció bellísima y que debe visitarse.

Posteriormente el sistema concatena todas las iUnits obtenidas generando una nueva consulta, cuya relevancia se evalúa con cada una de las frases de cada iUnit por separado obteniendo el valor de relevancia Rel_{IRn} tal como se puede ver en la figura 4:

El resumen se formaría por las iUnits con valor de relavancia mayor. Aunque dado el modelo que propone el sistema IR-n se podría aplicar de forma sencilla a recuperar las frases más relevantes en general de todas las iUnits.

5 *Evaluación*

Actualmente estamos trabajando en la recopilación de un corpus de prueba que pueda servir para las evaluar la bondad del modelo dentro del sector turístico, pero aprovechando la tarea Mobile Click de las conferencias NTCIR hemos podido realizar una serie de experimentaciones sobre generación y ordenación de frases sobre documentos de información general.

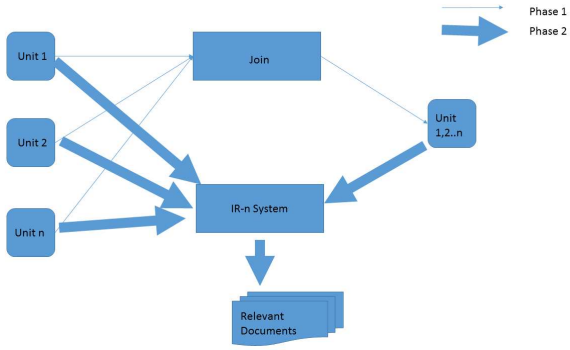


Figura 4: Sistema de Referencia: IR-n

Una de las tareas propuestas presentaba el siguiente planteamiento: se disponía de una consulta Q que se utilizaba para obtener documentos relevantes utilizando buscadores web tradicionales obteniendo una serie de IUnits (documentos de una frase). El objetivo era ordenar las iUnits en función de la relevancia con respecto a la pregunta original.

Por ejemplo la pregunta 1 era "*Pregunta 1(Q1) michael jackson death*" y las iUnits (iU) que ponía la organización a disposición de los participantes eran entre otras:

- Q1 iU1 family concerned about murray role.
- Q1 iU2 giving singer nightly doses of propofol.
- Q1 iU3 murray first met jackson in las vegas.

Los sistemas se evaluaban en base a la conocida medida Q-measure (Sakai, 2007), la que valora el orden en el que las frases relevantes son devueltas por los participantes teniendo en cuenta que la tarea exigía devolver todas las iUnits originales.

5.1 Resultados

Para las pruebas se disponía de la colección de preguntas y unidades utilizadas el año anterior. Así hemos decidido utilizar el modelo IR-n como caso base. Posteriormente hemos probado con el sistema COMPENDIUM

con varios porcentajes de comprensión (10 %, 20 % y 40 %), y utilizando o no los módulos de Textual Entailment que permiten eliminar información redundante.

5.1.1 Sistema Referencia: IR-n

Nuestro sistema base fue el de utilizar únicamente el sistema IR-n sin ningún preprocesamiento de las IUnits ni eliminación de las stopwords (ya que el modelo de calculo de relevancia les da peso ínfimo al prácticamente hallarse en todas las iUnits). La pregunta entrada del sistema es la concatenación de todas las iUnits referidas a cada pregunta.

Así el sistema concatenaba todas las iUnits generando una consulta que se evaluaba contra cada iUnit por separado obteniendo el valor de relevancia. Para el ejemplo anterior la consulta generada sería la siguiente.

Gen1 family concerned about murray role. giving singer nightly doses of propofol. murray first met jackson in las vegas.

5.1.2 Resultados con COMPENDIUM

Los sistemas de generación automática de resúmenes tienen como entrada un documento con la concatenación de todos los iUnits. A ese documentos se le aplica el sistema COMPENDIUM con 3 diferentes grados de compresión (10 %, 20 % y el 40 %).

Para cumplir los requisitos de la tarea, se le daba un valor de relevancia Rel_{Com} 1 a aquellas frases que el sistema de resúmenes seleccionaba y un 0 a aquellas frases que no se encontraban en el resumen.

También probamos el módulo de detección de redundancia basado en textual entailment. Los resultados de todos los experimentos se pueden ver en la Tabla 1.

	Q-measure		
	10 %	20 %	40 %
Baseline (IR-n)		**0,8621**	
COMPENDIUM	0,8196	0,8291	**0,8403**
COMPENDIUM+TE	0,8201	0,8218	0,8246

Tabla 1: Comparativa de resultados con sistemas aislados para el 10 %, 20 % y 40 % de ratio de resumen. En el baseline este ratio no se aplica.

El análisis de los resultados nos permitió obtener una serie de conclusiones:

- Textual Entailment empeoraba los resultados, al discriminar frases relevantes por estar incluidas en otras.
- En estos experimentos se obtuvieron mejores resultados con el sistema base (IR-n), ya que los sistemas de resúmenes no reodenaban las frases seleccionadas y simplemente las colocan en el orden de aparición en el texto original, siendo penalizadas en la medida Q-measure.

5.1.3 Combinando IR-n + COMPENDIUM

Dada la última conclusión obtenida, se optó por reordenar las frases obtenidas por el sistema de resúmenes teniendo en cuenta la relevancia que el sistema IR-n les había dado a cada una de ellas. Así, las primeras iban a ser las frase seleccionadas por el sistema de resúmenes, pero manteniendo el orden que el sistema IR-n había dado a cada una de ellas.

Para cada iUnit, se obtiene la siguiente relevancia:

$Rel_{def} = Rel_{IRn} + (Rel_{Com} * 1000)$

Los resultados se muestran en la Tabla 2. Como se puede observar, la combinación de ambos sistemas produce mejores resultados que utilizando sólo los sistemas de una manera independiente. El motivo es que IR-n reordena las frases que obtiene Compendium en lugar de utilizar el orden inicial . La medida Q-measure premia esta acción

	Q-measure
Baseline (IR-n)	**0.8621**
COMPENDIUM	**0.8403**
COMPENDIUM + IRn	**0.8648**

Tabla 2: Results con la combinación IR-n+COMPENDIUM

6 Evaluación

6.1 iUnit Ranking Subtask

Los resultados utilizando el modelo COMPENDIUM+IR-n sobre la colección de test se pueden ver en la Tabla 3. Se obtuvo el mejor resultado con IR-n+COMPENDIUM.

7 Conclusiones

Teniendo en cuenta las necesidades actuales de los usuarios actuales que utilizan sus dispositivos móviles para la obtención de información a través de Internet, este tipo de tareas tienen un enorme interés. El objetivo es ser capaz de reducir y condensar la información de tal forma que el usuario puede obtener de un simple vistazo la respuesta a sus dudas.

	Q-measure
IRn + COMPENDIUM (*)	**0.9027**
TITEC	0.9003
UHYG	0.8994
ORG	0.8975
RISAR	0.8972
RISAR	0.8962
IRn	0.8959
COMPENDIUM	0.8934

Tabla 3: Resultados para tarea iUnit Ranking Subtask

Los resultados obtenidos en las tareas del NTCIR 12 nos animan a seguir en la línea de explotación combinada de los sistemas COMPENDIUM e IR-n. Es necesario realizar una serie de trabajos adicionales para mejorar la experiencia de usuario. Por un lado evitar la redundancia de alguna de la información generada, que si que es valorada tal como se aplicaba la tarea pero no es interesante para el usuario.

Esto requiere la inclusión de técnicas adicionales de procesamiento del lenguaje natural pero con la problemática que deben estructurarse de forma que no consuman un tiempo importante que pudiera desesperar al usuario de este tipo de tareas.

Agradecimientos

Investigación realizada gracias a la financiación de los proyectos: DIIM2.0 (PROMETEOII/2014/001) de la Generalitat Valenciana; TIN2015-65100-R, DIGITY (TIN2015-65136-C2-2-R) del Ministerio de Economía y Competitividad y SAM (FP7-611312) de la Unión Europea

Bibliografía

Abulaish, M., Jahiruddin, Doja, M., and Ahmad, T. (2009). Feature and opinion mining for customer review summarization. In Chaudhury, S., Mitra, S., Murthy, C., Sastry, P., and Pal, S., editors, *Pattern Recognition and Machine Intelligence*, volume 5909 of *Lecture Notes in Computer Science*, pages 219–224. Springer Berlin Heidelberg.

Baccianella, A. E. S. and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Givón, T. (1990). *Syntax: A functional-typological introduction, II*. John Benjamins.

Glickman, O. (2006). *Applied Textual Entailment*. PhD thesis.

Kaszkiel, M. and Zobel, J. (1997). Passage Retrieval Revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Text Structures, pages 178–185, Philadelphia, PA, USA.

Kokkoras, F., Lampridou, E., Ntonas, K., and Vlahavas, I. (2008). Mopis: A multiple opinion summarizer. In Darzentas, J., Vouros, G., Vosinakis, S., and Arnellos, A., editors, *Artificial Intelligence: Theories, Models and Applications*, volume 5138 of *Lecture Notes in Computer Science*, pages 110–122. Springer Berlin Heidelberg.

Llopis, F. and Vicedo, J. L. (2001). IR-n system, a passage retrieval system at CLEF 2001. In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, pages 244–252, Darmstadt, Germany. Springer-Verlag.

Lloret, E. and Palomar, M. (2012). COMPENDIUM: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering*, 19(02):147–186.

Ly, D. K., Sugiyama, K., Lin, Z., and Kan, M.-Y. (2011). Product review summarization from a deeper perspective. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 311–314, New York, NY, USA. ACM.

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *IP&M*, 43(2):531–548.

Salton, G. A. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley, New York.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 23-30 recibido 29-03-2016 revisado 07-04-2016 aceptado 28-04-2016

Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain

Savana: Un Entorno Integral de Extracción de Información y Expansión de Terminologías en el Dominio de la Medicina

Luis Espinosa-Anke*†, Jorge Tello†, Alberto Pardo, Ignacio Medrano†**
Alberto Ureña†, Ignacio Salcedo†, Horacio Saggion*
*TALN-DTIC, Universitat Pompeu Fabra, Carrer Tànger, 122 08018 -Barcelona
†Medsavana S.L., C/ Jiloca 4 - 5 Derecha, 28016 - Madrid
**Fundación Jiménez Díaz
{luis.espinosa,horacio.saggion}@upf.edu
{jtello,ihmedrano,aurena,isalcedo}@savanamed.com
alberto.pardo@quironsalud.es

Abstract: Terminological databases constitute a fundamental source of information in the medical domain. They are used daily both by practitioners in the area, as well as in academia. Several resources of this kind are available, e.g. CIE, SnomedCT or UMLS (Unified Medical Language System). These terminological databases are of high quality due to them being the result of collaborative expert knowledge. However, they may show certain drawbacks in terms of faithfully representing the ever-changing medical domain. Therefore, systems aimed at capturing novel terminological knowledge in heterogeneous text sources, and able to include them in standard terminologies have the potential to add great value to such repositories. This paper presents, first, SAVANA, a Biomedical Information Extraction system which, combined with a validation phase carried out by medical practitioners, is used to populate the Spanish branch of SnomedCT with novel knowledge. Second, we describe and evaluate a system which, given a novel medical term, finds its most likely hypernym, thus becoming an enabler in the task of terminological database enrichment and expansion.
Keywords: Medical terminologies, knowledge bases, snomed, word2vec, semantics, savana

Resumen: Las bases terminológicas médicas constituyen una fuente de información fundamental en el dominio médico, ya que son utilizadas a diario tanto por profesionales en el sector como en el ámbito académico. Existen numerosos recursos de este tipo, tales como la Clasificación Internacional de Enfermedades (CIE), SnomedCT, o UMLS (Unified Medical Language System). La calidad de estas bases terminológicas es en general alta, dado que están construidas manualmente por expertos. Sin embargo, su capacidad para representar fielmente un dominio como el médico, que se encuentra en constante evolución, es limitada. Por tanto, el desarrollo de sistemas capaces de capturar nuevo conocimiento en fuentes textuales heterogéneas e incluirlas en terminologías estándar tienen el potencial de añadir un gran valor añadido a dichas terminologías. Este artículo presenta, en primer lugar, SAVANA, un sistema de extracción de información biomédica que, combinado con validación por parte de profesionales médicos, es utilizado para popular la rama española de SnomedCT con nuevo conocimiento. En segundo lugar, describimos y evaluamos un sistema que, dado un término médico nuevo, le asigna su hiperónimo más probable, constituyendo así un facilitador en tareas de enriquecimiento y expansión de bases terminológicas médicas.
Palabras clave: Terminologías médicas, bases de conocimiento, snomed, word2vec, semántica, savana

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Today, in the Information Age, we are witnessing an unprecedented growth in the production and availability of knowledge stored in the web, a massive source of publications, source code, data, research websites, wikis and blogs (O'Donoghue et al., 2014). The Health domain is not oblivious to this (r)evolution, and both academics and practitioners are starting to leverage medical evidence obtained from large-scale knowledge repositories in order to enhance the so-called Evidence-Based Medicine (Kumar, 2011). One of the great challenges for enabling data-driven support to clinical decisions is making sense of unstructured information appearing in research papers, text books, social networks or, what is more relevant for this paper, clinical records.

Clinical Health Records (CHRs) are documents where doctors take notes on a patient's medical condition, his or her progress, and suggest possible medication and treatment. They are a rich source of information because they provide personalized empirical data on treatment and evolution of medical conditions, and hence this type of document is receiving interest from the NLP community as enabler not only for medical support systems but also for its potential as training and evaluation data for Machine Learning algorithms in the field of Bioinformatics.

Examples of the interaction between NLP and CHRs include MedEx (Xu et al., 2010), a system for extracting medication information from clinical narratives, or a system for drug reaction event extraction (Santiso et al., 2016). However, and despite their potential, CHRs pose great challenges for automatic processing, as they are often unstructured, ill-defined and arduous to analyse at scale (Iqbal et al., 2015).

In this context, Medical Terminological Databases (MTDs) play a crucial role, as they provide a structured ground where medical concepts and their relations are encoded by medical experts and can be used as a benchmark for developing algorithms that leverage medical concept extraction to some extent. One of the best known MTDs is SnomedCT (Spackman, Campbell, and Côté, 1997), which is part of the UMLS (Bodenreider, 2004). One of the main drawbacks of MTDs is that creating and maintaining them manually is arduous. More importantly, keeping them up to date is not possible considering the amount of novel information that is generated daily. Furthermore, even if they are manually created, there is certain discussion even on their quality, since it is difficult to control the fitness of every single addition to the database (Morrey et al., 2009).

In this paper we propose to bridge the gap between unstructured medical knowledge stored arbitrarily in CHRs, on one hand, and the automatic maintaining of MTDs, on the other. In the remainder of this paper, we first describe Savana, a Biomedical Information Extraction system, which we run on a large collection of CHRs. In a second phase, Savana's predictions are presented to medical practitioners, who validate novel associations between SnomedCT concepts and their lexicalizations (i.e. the way they are expressed in free text). We exploit the combination of Savana and the validation stage to obtain a validation dataset of nearly 500 novel medical terms in Spanish, on which we evaluate several unsupervised systems aimed at finding, for each candidate novel term, its best point of attachment in the Spanish SnomedCT Database. These systems are based on both syntactic and semantic properties. Our results suggest that this is a promising direction for performing large-scale medical terminology extraction for Spanish, along with its *semantification*.

2 Related Work

The availability of MTDs is in constant growth. Paramount examples range from well-established collaborative efforts like UMLS (Bodenreider, 2004), which serves as an umbrella for multilingual resources such as SnomedCT (Spackman, Campbell, and Côté, 1997), or even the CIE database (*Clasificación Internacional de Enfermedades*), published by the Organización Panamericana de la Salud (1995). In addition, general purpose resources are increasingly playing more important roles in biomedical NLP tasks, as is the case of Wikipedia, which has been exploited for identifying medical disorders in CHRs (Bodnari et al., 2013).

Structured knowledge resources may present drawbacks such as staticity, and hence they may become obsolete in highly active and ever-changing domains such as the biomedical. Research in NLP has

proposed several approaches to alleviate this problem. For instance, in the area of learning lexical taxonomies, novel terms are discovered and included in domain-specific is-a hierarchies (Velardi, Faralli, and Navigli, 2013; Luu Anh, Kim, and Ng, 2014; Kozareva, 2014; Espinosa-Anke et al., 2016). Another approach is to combine Open Information Extraction paradigms with predefined semantic criteria. These criteria may be based on Wikipedia (Nakashole, Weikum, and Suchanek, 2012), BabelNet (Delli Bovi, Espinosa-Anke, and Navigli, 2015) or may be *ad-hoc* semantic hierarchies (Carlson et al., 2010). Finally, there are (fewer) approaches on discovering novel terminology from domain-specific dictionaries, glossaries or web pages and providing semantics via intersection with WordNet (Miller, 1995), e.g. by processing associated glosses or definitions (Jurgens and Pilehvar, 2015). Most of these approaches, however, while having shown notable success in the English language, did not address other languages, for which availability of resources and tools is scarcer.

The medical domain has also received attention in terms of automatically expanding existing resources. Prominent examples include (1) The development of novel MTDs from Wikipedia (Pedro, Niculescu, and Lita, 2008); (2) Enriching SnomedCT terminology with associated definitions (Ma and Distel, 2013); and in multilingual settings, (3) Expansion of SnomedCT in Swedish by processing CHRs (Henriksson et al., 2013).

3 Savana

We use SAVANA, a Biomedical Information Extraction System[1], integrated in several public and private healthcare institutions in Spain, for obtaining and validating ground truth data. The SAVANA algorithm is designed to retrieve prominent biomedical information from CHRs in the Spanish language. It does so by combining in its pipeline modules for, among others, sentence segmentation, tokenization, spell checking, acronym detection and expansion, negation identification, and a multi-dimensional ranking scheme which combines linguistic knowledge, statistical evidence, and state-of-the-art continuous vector representations of words and documents in the biomedical domain learned via shallow neural networks. We run SAVANA over several thousand CHRs, and ask medical practitioners to validate matches of SAVANA's association between a mention of a medical concept in text, and an existing SnomedCT entry, by means of a web interface (Figure 1). The subset of the Spanish SnomedCT branch on which we run our experiments contains over 401,126 concepts, which are linked by means of 2,722,877 hypernymic (is-a) relations. The validation procedure may yield *novel terminology* in terms of either novel lexicalizations for an existing term (synonyms), or novel terms which can be attached to a more general SnomedCT concept (hyponyms). In this paper we are interested in the latter case: Finding the best point of attachment for novel concepts, rather than finding additional ways of expressing the same idea. At validation stage, if human experts consider that a concept identified by SAVANA has a meaning which is missing in SnomedCT, this concept makes it to our ground truth novel terminology, and hence will constitute the testbed for the experiments we describe in Section 4. We collect gold standard data of up to 492 novel terms, with an average of 3.2 hypernymic relations encoded by human experts. There was no restriction in the type of concept to be included. Therefore, this dataset includes diverse terms which are related to infrastructure, e.g. SERVICIO DE ODONTOLOGÍA $\rightarrow$ {servicio hospitalario}[2], or actual medical conditions, e.g. GONARTROSIS$\rightarrow$ {trastorno de la rodilla, enfermedad de la rodilla}. In the following section, we describe the experiments carried out to discover the most appropriate hypernym for each of the 492 novel terms we incorporated to SnomedCT thanks to combining the SAVANA algorithm with an expert validation stage.

4 Enriching SnomedCT

In this section we describe the SnomedCT enrichment experiments. Given a novel term, we aim at finding its best point of attachment, expressed as its closest hypernym. Our approach is unsupervised and hence requires no prior annotation or training. Moreover, we do not exploit any web or Wikipedia-based textual evidence (which we may inves-

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

[2] We denote is-a relations between terms and sets of hypernyms as *term*$\rightarrow${*hypernym1*, *hypernym2*}.

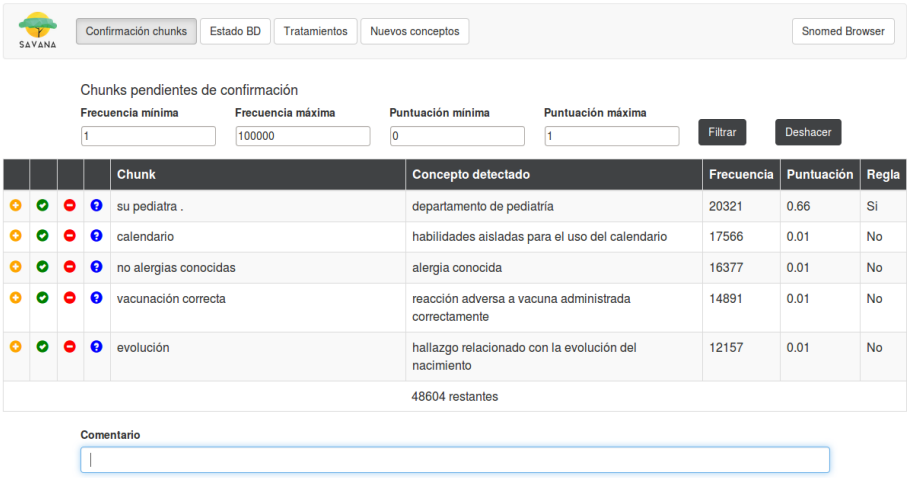


Figure 1: A snapshot of the validation web interface. Let us highlight how the validation procedure allows the medical expert to assign to the SnomedCT concept `departamento de pediatría`, a novel lexicalization in the context of CHRs, namely the string `su pediatra`.

tigate in future work). However, we do leverage two main resources in our experiments, which are described briefly.

- For syntactic parsing, we use a transition-based parser based on the parsing technology included in the Mate framework (Bohnet, 2010).
- For computing similarities between concepts, we exploit word embeddings derived from training a shallow neural net model (Mikolov, Yih, and Zweig, 2013) with the *word2vec*[3] tool, implemented in *gensim*[4]. The model used for our experiments comes from a 2015 dump of the Spanish Wikipedia preprocessed and lemmatized with Freeling (Atserias et al., 2006). Our model is 300-dimensional, and is trained using the skip-gram with negative sampling algorithm, using a minimum count of 10 for each word.

Having described the two main technological pivots of our approach, let us describe each of the systems evaluated:

- **Substring***[5] This is a substring inclusion baseline which, for each novel term, assigns as term hypernyms all Snomed concepts that are subsumed in the novel term. For example, given the unseen concept GONARTROSIS, candidate hypernyms are ARTROSIS and ARTROSIS (TRASTORNO). Note that this approach fails short when dealing with longer and more complex terminology, as in the case of the concept NO OTROS HÁBITOS TÓXICOS, where incorrect hypernyms are captured, such as TOS or OTRO.
- **Head Fuzzy Match*** Multiword terms (mwt) may be generalized via their syntactic dependencies. For example, given the novel concept INSUFICIENCIA CARDÍACA CONGESTIVA LEVE, after dependency parsing we are able to isolate INSUFICIENCIA as the mwt's head. This configuration of our approach collects all Snomed concepts of which this head is substring. In this example, we would correctly match INSUFICIENCIA CARDÍACA, but also generate false positives such as INSUFICIENCIA HEPÁTICA or INSUFICIENCIA RESPIRATORIA TIPO 2.
- **Head Exact Match** This is a restricted version of *Head Fuzzy Match*, in which in most cases we only obtain one candidate, i.e. the Snomed concept which matches exactly the out-of-vocabulary (OOV) term's head. For instance, for the concept NO OTROS HÁBITOS TÓXICOS, the retrieved candidate would be the Snomed concept

[3] code.google.com/archive/p/word2vec/
[4] radimrehurek.com/gensim/models/word2vec.html
[5] We distinguish baseline systems with *.

HÁBITO.

- **Distributional** The first of our distributional approaches, exploiting word embeddings, stems from the intuition that similar concepts may occur in similar contexts. This property has been confirmed to hold in many semantic relations (Mikolov, Yih, and Zweig, 2013; Mikolov et al., 2013). In this configuration of our system, given a term t consisting of a set of words $\{w_i, ...w_n\}$ (after stopword removal), we compute the centroid vector μ of the set of associated word vectors $\overrightarrow{w} \in t$. We obtain $\mu(t)$ as follows:

$$\mu(t) = \frac{1}{|t|} \sum_{\overrightarrow{w} \in t} \frac{\overrightarrow{w}}{\|\overrightarrow{w}\|} \quad (1)$$

  We perform the same operation on all candidate Snomed concepts. Specifically, we obtain, given a Snomed terminology $\mathcal{S}$, for each Snomed term $t_s \in \mathcal{S}$, its corresponding centroid vector $\mu(t_s)$. Then, our algorithm returns as best match the Snomed concept maximizing the semantic similarity between t and t_s, denoted as $\text{SIM}(t, t_s)$, and computed via cosine score as follows:

$$\text{SIM}(t, t_s) = \frac{\mu(t) \cdot \mu(t_s)}{\|\mu(t)\| \, \|\mu(t_s)\|} \quad (2)$$

  This operation yields a ranked list of candidates by score, where score is the cosine score above, and thus the predicted candidate is the term t_s with the highest similarity with the input term t.

- **DistDep** Our last experiment is performed with the **DistDep** system, which combines head word lookup with similarities derived from word embeddings. It simply consists in comparing the vector associated to the head node (as extracted in any of the **Head**-based approaches) of the novel term with the centroid of all available concepts in $\mathcal{S}$, and keeping as best match the highest scoring candidate.

5 Evaluation

5.1 Distance-based Evaluation

The evaluation of a system's performance in terms of its ability to attach novel terminology to an existing knowledge repository is traditionally performed by considering distance between reference nodes and predicted nodes, i.e. the best point of attachment, and the decision made by the system. While evaluation metrics exist for computing semantic similarity over lexical databases like WordNet, these are not suitable in our case because our branch of Snomed is designed in a slightly different fashion, as it can be considered a multiroot directed acyclic graph (DAG), and hence in many cases, given two concepts, there is no *least common subsumer* other than the root node. For instance, the path between TRASTORNO CON TALLA BAJA and PRUEBA DE VARIANTE DE HEMOGLOBINA includes one of the root nodes in the SnomedCT taxonomy, namely SNOMED CLINICAL TERMS (ENERO 2014).

This makes metrics like Wu&Palmer Similarity (Wu and Palmer, 1994), which considers lowest common subsumers in their similarity computation, unsuitable. For this reason, we propose a distance metric for evaluation purposes sensitive to terminological databases shaped as DAGs rather than trees (like WordNet). We account for the fact that there may be several valid points of attachment and hence compute an average of node-based shortest path $sp(\cdot)$ over all predicted candidates and all gold standard nodes.

Let $\mathcal{G}$ be the set of gold standard points of attachment to a novel term t, and $\mathcal{P}$ the set of predictions generated by a system. We define an Error-Score function E that, given a novel term t, computes the average shortest path of all predicted points of attachment $p \in \mathcal{P}$:

$$E(t) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{\sum_{g \in \mathcal{G}} sp(p, g)}{|\mathcal{G}|} \quad (3)$$

In addition, we report a second evaluation based on whether a system is able to capture all the gold standard points of attachment, regardless of additional incorrect predictions. This is performed only on those systems which return a *set of candidates*, which is not the case of the distributional systems **Distributional** and **DistDep**. We included evaluation under this Recall score, which we denote as $R(t)$, as we foresee a real world scenario where human post-edition of false positives may be less time-consuming than finding in SnomedCT the best points of attachment for each novel term. We simply

set $R(t) = 1$ if a given approach is able to cover, with its n predictions, all the possible gold standard attachments, and $R(t) = 0$ otherwise, and average results over the total prediction sets. We provide the evaluation results for both criteria in Table 1. The two main conclusions that can be drawn from our experimental results are that, first, leveraging similarities derived from word embeddings improve the performance of MTD enrichment systems, and second, that exploiting a greedy approach of fuzzy syntactic head matching is a reasonable strategy for increasing recall.

	Error-Score	Recall
Substring*	8.51	26%
Head Fuzzy	7.07	**84**%
Head Exact	4.72	13%
Distributional	**3.34**	N/A
DistDep	3.36	N/A

Table 1: Evaluation results for our proposed systems in terms of average performance of all its predictions (Error-Score), and Recall.

Finally, we plotted the performance in Error-Score of our three proposed systems (not baselines) over all the novel terms present in the evaluation data. We can observe that the two distributional systems based on word embeddings show a similar behaviour, much better in general than the third best system, **Head Exact** (Figure 2).

5.2 Human Evaluation

We assume in our automatic evaluation that human experts in the biomedical domain will provide a solid ground truth against which system predictions can be evaluated. However, given the size of SnomedCT, our system may provide correct points of attachment for novel terminology which were not included in the first place, and this is penalized in the automatic evaluation. For this reason, we presented human experts with the set difference between the sets of gold and predicted points of attachment, and asked them to label them as correct or incorrect. We find an average of 27% correctness over all systems, which suggests that certain cases of *false positives* were actually correct predictions and hence were valid inclusions of novel terms along

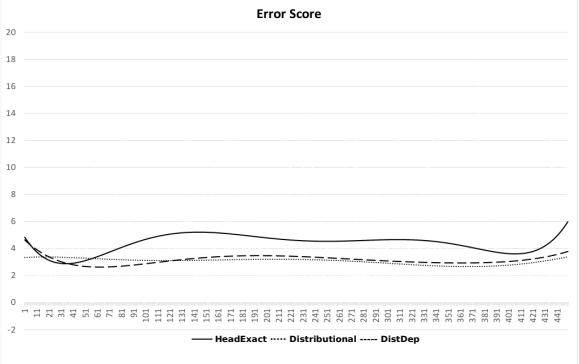


Figure 2: Error-Score results (y axis) for the three systems we propose, namely **Distributional** (dotted), **DistDep** (dashed), and **Head Exact Match** (line), over the whole test terminology (x axis).

with their associated hypernymic relations. We illustrate a few cases of *false positives* in Table 2 together with their correctness according to a domain expert.

6 Conclusions and Discussion

The rapidly growing interplay between Artificial Intelligence and healthcare is producing innovative assistive technologies (e.g. adaptive and rehabilitative devices), as well as medical support systems which leverage large quantities of heterogeneous data. Among the latter, let us highlight SAVANAMED, which thanks to the SAVANA algorithm, provides a real-time medical support system by making sense of textual information present in CHRs. In this paper, we described how the SAVANA algorithm, backed up by a validation stage carried out by medical practitioners, was used to produce a ground truth for evaluating a system in the task of MTD enrichment. We evaluated several systems against this data and found that combining linguistic information derived from syntactic dependencies, as well as similarities computed over word produces the best results. To the best of our knowledge, both SAVANA and the MTD enrichment system are the first systems of their kind developed for the Spanish language.

Our encouraging results open up a promising line of research in NLP tasks like semantics and Information Extraction in the medical domain, both industrial and academic

Novel Term	$\mathcal{G}$	**Novel PoA** (fp)	**Correctness**
dermatitis seborreica leve	eccema seborreico	dermatitis	Yes
servicio de cardiología pediátrica	servicio hospitalario	servicio de cardiología	Yes
artrosis cervical	artrosis	linfadenopatía cervical	No
talla baja idiopática	trastorno con baja estatura	al examen: estatura baja	No

Table 2: Illustrative cases where some of the novel concepts discovered by our approach, and evaluated as false positive (fp) by the automatic criteria, were considered correct in a second pass by human domain experts.

purposes, where we expect that close collaboration may result in larger and better datasets both for AI and NLP, and also for medical practitioners.

7 Future Work

We have presented an unsupervised approach for enriching medical terminological databases in Spanish. Our original idea was to explore to what extent we could *do well* in a setting without external knowledge being included in the pipeline (web, Wikipedia, and so forth), mainly having scalability in mind. However, we are interested in expanding our initial approach with this external knowledge, in the hope that it will contribute to discover additional candidates a medical terms, and as evidence for finding their best point of attachment.

We would also like to expand our experiments by including word embeddings coming from neural nets training, but trained on medical corpora. We are currently gathering data both from known medical websites, as well as a pool of medical records. We would also like to expand the system described in this paper with knowledge mined from scientific papers in order to also take advantage of information expressed in more canonical and less noisy fashion.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful comments. This work is partially funded by the Spanish Ministry of Economy and Competitiveness under the following sponsorships: Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), and TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE). The authors also acknowledge support from Savana and Hospital Universitario Fundación Jiménez Díaz.

References

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In *Proceedings of LREC*, volume 6, pages 48–55.

Bodenreider, Olivier. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Bodnari, Andreea, Louise Deleger, Thomas Lavergne, Aurelie Neveol, and Pierre Zweigenbaum. 2013. A Supervised Named-Entity Extraction System for Medical Text. In *CLEF (Working Notes)*.

Bohnet, Bernd. 2010. Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In *COLING*, pages 89–97.

Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of AAAI*, pages 1306–1313.

Delli Bovi, Claudio, Luis Espinosa-Anke, and Roberto Navigli. 2015. Knowledge Base Unification via Sense Embeddings and Disambiguation. In *Proceedings of EMNLP*, pages 726–736.

Espinosa-Anke, Luis, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*.

Henriksson, Aron, Maria Skeppstedt, Maria Kvist, Martin Duneld, and Mike Conway. 2013. Corpus-driven Terminology Development: Populating Swedish

SNOMED CT with Synonyms Extracted from Electronic Health Records. *ACL 2013*, page 36.

Iqbal, Ehtesham, Robbie Mallah, Richard George Jackson, Michael Ball, Zina M Ibrahim, Matthew Broadbent, Olubanke Dzahini, Robert Stewart, Caroline Johnston, and Richard JB Dobson. 2015. Identification of Adverse Drug Events from Free Text Electronic Patient Records and Information in a Large Mental Health Case Register. *PloS one*, 10(8):e0134208.

Jurgens, David and Mohammad Taher Pilehvar. 2015. Reserating the Awesometastic: An Automatic Extension of the WordNet Taxonomy for Novel Terms. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies, Denver, CO*, pages 1459–1465.

Kozareva, Zornitsa. 2014. Simple, Fast and Accurate Taxonomy Learning. In *Text Mining*. Springer, pages 41–62.

Kumar, Dhavendra. 2011. The Personalised Medicine: A Paradigm of Evidence-based Medicine. *Annali dell'Istituto superiore di sanitÃ*, 47(1):31–40.

Luu Anh, Tuan, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy Construction Using Syntactic Contextual Evidence. In *EMNLP*, pages 810–819, October.

Ma, Yue and Felix Distel. 2013. Learning Formal Definitions for SNOMED CT From Text. In *Artificial Intelligence in Medicine*. Springer, pages 73–77.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Morrey, C Paul, James Geller, Michael Halper, and Yehoshua Perl. 2009. The Neighborhood Auditing Tool: A Hybrid Interface for Auditing the UMLS. *Journal of biomedical informatics*, 42(3):468–489.

Nakashole, Ndapandula, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of EMNLP-CoNLL*, pages 1135–1145.

O'Donoghue, DP., H Saggion, F Dong, D Hurley, Y Abgaz, X Zheng, O Corcho, JJ Zhang, J-M Careil, B Mahdian, et al. 2014. Towards Dr Inventor: A Tool for Promoting Scientific Creativity. In *ICCC*.

Organización Panamericana de la Salud. 1995. Clasificación estadística internacional de enfermedades y problemas relacionados con la salud: décima revisión: CIE-10.

Pedro, V, S Niculescu, and L Lita. 2008. Okinet: Automatic extraction of a Medical Ontology from Wikipedia. In *WiKiAI08: a workshop of AAAI2008*.

Santiso, Sara, Arantza Casillas, Alicia Pérez, Maite Oronoz, and Koldo Gojenola. 2016. Document-level Adverse Drug Reaction Event Extraction on Electronic Health Records in Spanish. *Procesamiento del Lenguaje Natural*, 56:49–56.

Spackman, Kent A, Keith E Campbell, and Roger A Côté. 1997. SNOMED RT: A Reference Terminology for Health Care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Velardi, Paola, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.

Wu, Zhibiao and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Xu, Hua, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. MedEx: A Medication Information Extraction System for Clinical Narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Desarrollo de Recursos y Herramientas Lingüísticas

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 33-40 recibido 14-03-2016 revisado 19-04-2016 aceptado 03- 05-2016

Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego

Building a UD treebank using existing resources from related languages: the case of Galician

Marcos Garcia
Grupo LyS, Dep. de Galego-Portugués, Francés e Lingüística
Universidade da Coruña
marcos.garcia.gonzalez@udc.gal

Carlos Gómez-Rodríguez
Grupo LyS
Dep. de Computación
Universidade da Coruña
carlos.gomez@udc.es

Miguel A. Alonso
Grupo LyS
Dep. de Computación
Universidade da Coruña
miguel.alonso@udc.es

Resumen: En este trabajo presentamos una nueva estrategia para crear *treebanks* de lenguas con pocos recursos para el análisis sintáctico. El método consiste en la adaptación y combinación de diferentes *treebanks* anotados con *dependencias universales* de variedades lingüísticas próximas, con el objetivo de entrenar un analizador sintáctico para la lengua elegida, en nuestro caso el gallego. Durante el proceso de selección y adaptación de los *treebanks* de origen, analizamos el impacto de propiedades de tres niveles diferentes: (i) la distancia entre las lenguas de origen y destino, (ii) la adaptación de características léxico-ortográficas, y (iii) las directrices de anotación entre los *treebanks*. Usando la estrategia propuesta, entrenamos un analizador sintáctico estadístico para etiquetar, con resultados prometedores y sin datos previos de gallego, un pequeño corpus de esta lengua. La corrección manual de este corpus, usado como *gold-standard*, nos permitió probar la eficacia del método propuesto.
Palabras clave: análisis sintáctico, *treebank*, *dependencias universales*, gallego

Abstract: This paper presents a novel strategy for creating a Universal Dependencies (UD) treebank of a low-resource language. The method consists of adapting and combining different UD treebanks from related varieties in order to train a parser for the target language. More precisely, the paper explores the influence of three different levels for the selection and adaptation of the source treebanks: (i) the relatedness of the linguistic varieties, (ii) the adaptation of features based on lexical and spelling data, and (iii) the agreement in annotation criteria between different treebanks. The proposed strategy allowed us to train a parser for analyzing, with promising results, a small Galician corpus without previous availability of labeled data for this language. After a few bootstrapping iterations, we obtained a UD gold-standard corpus, used for proving the effectiveness of the proposed method.
Keywords: parsing, treebank, universal dependencies, Galician

1 Introducción

El uso de corpus anotados sintácticamente (*treebanks*) se ha demostrado útil en diferentes áreas, como los estudios en lingüística de corpus o trabajos de análisis sintáctico automático (*parsing*), que es a su vez beneficioso para tareas como la minería de opiniones, o la traducción automática, entre otras (Socher et al., 2013; Gimpel y Smith, 2014). Con todo, la creación de este tipo de recursos es una tarea costosa, ya que implica la etiquetación manual de una gran cantidad de información lingüística de diferentes niveles. El proceso de anotación sintáctica se puede aliviar mediante la aplicación previa de un analizador automático, corrigiendo así únicamente los errores producidos por este sistema. En lenguas para las que no existen este tipo de herramientas, se han propuesto diferentes estrategias que aprovechan recursos de otros idiomas para entrenar *parsers* estadísticos. Entre estas técnicas encontramos el uso de corpus paralelos de las lenguas de origen y destino (Zeman y Resnik, 2008), a veces enriqueciendo el *parser* con reglas específicas del

* Este trabajo ha sido parcialmente financiado por el MINECO (proyectos FFI2014-51978-C2-1-R y FFI2014-51978-C2-2-R, y un contrato *Juan de la Cierva formación*: FJCI-2014-22853), y por la Xunta de Galicia (programa *Oportunius*).

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

idioma de destino (Hwa et al., 2005). Sin embargo, tanto diferencias lingüísticas (u otras divergencias de anotación entre los corpus) como la escasez de este tipo de recursos pueden dificultar este proceso.

En un intento de homogeneizar —en la medida de lo posible— las directrices de anotación sintáctica, el proyecto *Universal Dependencies* (UD) promueve una anotación consistente de los diferentes *treebanks* de las lenguas naturales (McDonald et al., 2013). Así, utilizando un conjunto universal de dependencias sintácticas (aunque permitiendo etiquetas diferentes para anotar fenómenos específicos de algunas lenguas), UD facilita, por ejemplo, el aprovechamiento de recursos entre varias lenguas o el análisis interlingüístico de fenómenos sintácticos.

Con el objetivo de crear un corpus con anotación sintáctica UD para gallego, en el presente trabajo proponemos una estrategia de combinación y adaptación de *treebanks* de variedades lingüísticas próximas, que permiten una anotación inicial de alta calidad. En los procesos de selección y adaptación de los *treebanks* de origen, se tienen en cuenta características de tres niveles (en relación al idioma de destino): (i) proximidad lingüística, (ii) distancia léxico-ortográfica y (iii) particularidades de anotación interlingüística.

La estrategia aquí propuesta, evaluada en $\approx$ 12.000 *tokens* corregidos manualmente, obtiene resultados prometedores en lo que respecta al aprovechamiento de recursos de lenguas próximas para la creación de un nuevo *treebank* UD, y muestra que tanto la proximidad lingüística (sintáctica y léxica) como las variaciones de anotación son relevantes en el proceso de transferencia.

Además de esta sección introductoria, el artículo se organiza de la siguiente manera. La sección 2 incluye una revisión del trabajo relacionado, mientras que la sección 3 presenta las principales características del proyecto UD y de la adaptación del corpus gallego a este proyecto. En las secciones 4 y 5 presentamos y evaluamos, respectivamente, el método de transferencia propuesto. Finalmente, la sección 6 contiene las conclusiones del estudio, así como ideas para el trabajo futuro.

2 Trabajo Relacionado

Diversos trabajos han analizado el uso de recursos sintácticos de una o más lenguas para crear un *treebank* de un idioma diferente, con resultados dispares. Así, antes de la existencia de las UD, varios trabajos utilizaron corpus paralelos para proyectar la anotación sintáctica de una lengua origen (con recursos) a la lengua de destino (Hwa et al., 2005; Ganchev, Gillenwater, y Taskar, 2009).

En Zeman y Resnik (2008) se entrena un *parser* únicamente con información sintáctica y morfosintáctica de la lengua de origen (*parser* deslexicalizado), para analizar posteriormente textos en la lengua de destino. La deslexicalización obtiene mejores resultados que el uso de información léxica en el par de lenguas evaluado (sueco–danés). Trabajos posteriores mejoraron esta técnica al combinarla con el uso de corpus paralelos y comparables, añadiendo también más de un idioma al conjunto de *treebanks* de origen (Søgaard, 2011; McDonald, Petrov, y Hall, 2011).

Utilizando las UD, McDonald et al. (2013) también evalúan el rendimiento de *parsers* entrenados para un idioma diferente del que posteriormente analizan. La estrategia de deslexicalización —con corpus paralelos— proporciona mejoras en el análisis, y la transferencia entre lenguas próximas obtiene mejores resultados que la realizada entre variedades lingüísticamente más distantes.

Con todo, las evaluaciones de Lynn et al. (2014) (para el irlandés), o de Vilares, Alonso, y Gómez-Rodríguez (2016) (donde se entrenan y evalúan varios *parsers* bilingües) sugieren que el resultado de la transferencia de recursos sintácticos entre idiomas no tiene por qué estar relacionado con la proximidad lingüística entre ellos (entendiendo la proximidad en términos de pertenencia —o no— a la misma familia lingüística).

En lo que respecta a *treebanks* de gallego, no conocemos hasta este momento ningún corpus disponible con anotación sintáctica, si bien durante el desarrollo de este trabajo la página web del proyecto UD informó sobre un *treebank* en desarrollo, que estará disponible a partir de la versión 1.3.[1]

Los trabajos sobre *parsing* para gallego tampoco son muy abundantes, aunque existen varios artículos que implementan reglas sintácticas en analizadores automáticos. Así, Gamallo Otero y González López (2011) presentan una *suite* multilingüe de análisis de dependencias que incluye un *parser* de gallego.

[1] `http://universaldependencies.org`

Por su parte, las versiones más recientes de FreeLing también disponen de un *parser* para gallego, que realiza análisis tanto de constituyentes como de dependencias sintácticas (Padró y Stanilovsky, 2012). Las dependencias utilizadas por ambos sistemas (DepPattern y FreeLing) no son UD, por lo que su utilización en el presente trabajo supondría un proceso de adaptación mayor. Además, la inexistencia de *treebanks* tampoco facilita la realización de evaluaciones empíricas de los diferentes analizadores.

Finalmente, existen algunos trabajos que —como el actual— han aprovechado la proximidad lingüística entre portugués y gallego para generar recursos de este último a partir del primero: entre otros, Malvar et al. (2010) obtienen corpus bilingües para entrenar modelos de traducción automática, mientras que Garcia y González (2012) generan, para un sistema de transcripción fonética automática, léxicos de gallego utilizando léxicos de portugués europeo.

En este trabajo analizamos el uso de recursos sintácticos de UD en español y portugués (entre otras lenguas) para el análisis de un corpus gallego, estudiando también el impacto de las características léxico-ortográficas y de anotación entre los diferentes *treebanks* de origen y destino.

3 Dependencias Universales y Corpus Gallego

McDonald et al. (2013) fueron los primeros en utilizar, en varios corpus, el conjunto de *dependencias sintácticas universales*, publicando *treebanks* de 6 lenguas diferentes. En el origen de este conjunto de dependencias está, por un lado, una versión de las etiquetas sintácticas del *parser* de inglés del NLP Group de la universidad de Stanford (De Marneffe y Manning, 2008) y, por otro lado, el conjunto de etiquetas morfosintácticas universales propuestas por Google (Petrov, Das, y McDonald, 2012). Así, el proyecto UD tiene entre sus objetivos facilitar tanto el desarrollo de analizadores multilingües y el aprovechamiento mutuo de recursos de diferentes lenguas, como el estudio interlingüístico de fenómenos sintácticos.

Como hemos referido, UD promueve una anotación (no sólo sintáctica, sino también morfosintáctica y de *tokenización*) consistente entre *treebanks* de diferentes lenguas, mediante el uso de un conjunto universal de etiquetas y unas directrices de anotación homogéneas. Sin embargo, teniendo en cuenta que existen fenómenos lingüísticos particulares, cada *treebank* puede utilizar variantes propias de las *dependencias universales* para anotar este tipo de fenómenos.

A este respecto, durante la actual etapa preliminar de etiquetación estamos definiendo unas directrices propias que, siguiendo las recomendaciones UD, nos permitan analizar satisfactoriamente los fenómenos lingüísticos específicos del gallego. Estas directrices, en su versión inicial sujeta a posibles revisiones o ampliaciones en el futuro, se basan en tres pilares básicos:

1. Utilización —siempre que sea posible— de los principios de UD
2. Uso del menor número posible de dependencias y directrices de anotación diferentes de las etiquetas universales
3. Coherencia (si es posible) con la anotación sintáctica del *treebank* portugués, en aquellos casos en los que UD permita varias soluciones de anotación

En este sentido, la principal divergencia de anotación con respecto a las directrices UD ha sido la utilización de la etiqueta *iobj* (objeto indirecto) en aquellos casos en los que el objeto directo (*dobj*) no está explícito (en estas situaciones, UD recomienda etiquetar el *iobj* como *dobj*). Esta decisión ha sido tomada porque la discriminación de estas etiquetas favorece tanto el análisis lingüístico como la extracción de información del *treebank*, dado que en el corpus gallego la preposición que introduce el objeto (normalmente *a*) aparece tanto en *dobj* como en *iobj*. La Figura 1 contiene un ejemplo de una oración del corpus gallego con dependencias UD (cuya traducción al español podría ser "La competencia le corresponderá a la RAG"), en donde se puede observar la anotación del único objeto como *iobj*, y del pronombre clítico como *expl*.

El corpus elegido para iniciar el proceso de construcción del *treebank* gallego fue el XIADA 2.6 (Rojo et al., 2015), un recurso con más de 740.000 *tokens* lematizados y con anotación morfosintáctica corregida manualmente. XIADA se compone de textos de dominio periodístico, económico y narrativo en gallego. Durante la adaptación de este corpus hemos mantenido algunas particularida-

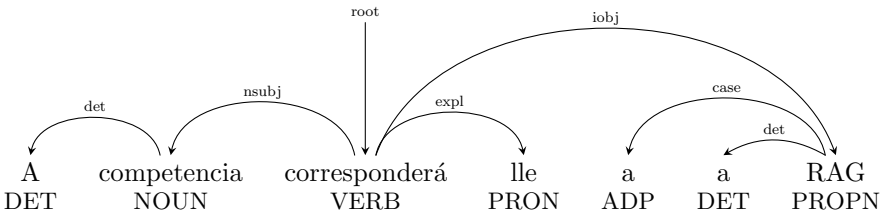


Figura 1: Oración del corpus gallego con anotación (sintáctica y morfosintáctica) UD.

des que, desde el punto de vista del proyecto *Universal Dependencies*, cabe mencionar:

***Tokenización*:** con el objetivo de preservar la *tokenización* de XIADA se ha mantenido la división original del corpus. Así, tanto los nombres propios compuestos (de más de un *token*) como algunas locuciones son etiquetadas como elementos individuales, y no separadas en *tokens* como recomienda UD.

Anotación morfosintáctica: UD usa un *tagset* universal de 17 etiquetas que incluyen las categorías morfosintácticas básicas (adjetivo, adverbio, verbo, etc.), codificando —si existe— otra información de carácter morfosintáctico (género, número, etc.) como características independientes de las categorías.

En nuestro caso, hemos extraído de las etiquetas XIADA tanto la categoría UD como las restantes características morfosintácticas, manteniendo en el corpus la información original y la extraída automáticamente.

En general, la anotación morfosintáctica del corpus no se ha modificado durante la corrección del *treebank*, salvo en aquellos casos en que se han detectado errores inequívocos de anotación.

4 Selección y Adaptación de treebanks *de origen*

La disponibilidad de *treebanks* en gallego es necesaria tanto para diferentes tareas del procesamiento computacional de esta lengua como para realizar estudios interlingüísticos con otros *treebanks* con los que se comparta anotación. Así, ante la inexistencia de recursos ya etiquetados, hemos optado por estudiar diferentes métodos para la transferencia de *parsers* desde otros idiomas.

La estrategia propuesta en este trabajo enfoca la transferencia de analizadores sintácticos de una o más lenguas de origen (que dispongan de *treebanks*) a una lengua de destino, con base en tres parámetros:

1. Proximidad lingüística —especialmente sintáctica— entre las variedades de origen y destino
2. Distancia léxico-ortográfica entre los corpus
3. Variación en las directrices de anotación (dentro del conjunto de etiquetas UD)

Proximidad lingüística: como hemos visto, las evaluaciones de diferentes trabajos no confirman firmemente que la distancia lingüística sea un factor decisivo en la transferencia de *parsers* de un idioma a otro. Con todo, diversas evaluaciones aquí realizadas (con *treebanks* UD de idiomas de diferentes familias lingüísticas) nos sugieren que algunas lenguas se pueden analizar con resultados aceptables utilizando recursos de variedades muy próximas desde el punto de vista sintáctico y léxico (véase la sección 5).

Así, con el objetivo de analizar el corpus XIADA, seleccionamos (después de evaluaciones preliminares) los *treebanks* UD de español y portugués europeo como origen. La elección de estas variedades se debe, por un lado, a que ambas lenguas tienen estructuras sintácticas muy similares a las de gallego. Además, el español coexiste con el gallego en el mismo territorio, y las interferencias sintácticas —y otras— son frecuentes entre las dos lenguas (también en el corpus XIADA). El uso del portugués como lengua origen está basado en el hecho de que tanto el gallego como el portugués provienen del mismo sistema lingüístico (*galego-português*), siendo considerados por algunos lingüistas todavía en la actualidad como variedades del mismo idioma (Cintra y Cunha, 1984).[2]

Distancia léxico-ortográfica: las diferencias léxicas entre varios idiomas propiciaron el uso de estrategias de deslexicalización,

[2]Sea como fuere, entre los dos estándares existen diferencias cuyo impacto en la transferencia tratamos de reducir usando métodos de adaptación ortográfica.

diseñadas con el objetivo de minimizar el impacto negativo al analizar idiomas con mayor distancia léxica. A este respecto, este trabajo propone como una de las estrategias de adaptación de *treebanks*, la transliteración ortográfica del corpus portugués. Para ello, hemos construido automáticamente una versión del *treebank* portugués con ortografía muy próxima a la del estándar gallego, usando la estrategia adoptada en Malvar et al. (2010).

A pesar de que este método solo es aplicable entre variedades lingüísticas muy próximas, técnicas similares (con base en diccionarios bilingües o en similitud léxica) se podrían evaluar en otros pares de lenguas.

Directrices de anotación: el proyecto UD promueve unas directrices estándar de anotación para las diferentes lenguas, pero los *treebanks* individuales pueden tener características de etiquetación propias, no sólo por el uso de dependencias específicas de un idioma, sino por decisiones particulares de los anotadores (recuérdese, por ejemplo, nuestra decisión de priorizar el uso de *iobj* sobre *dobj*, explicada en la sección 3).

Hasta el momento, el principal cambio relativo a las directrices de anotación que hemos realizado durante la adaptación de los corpus español y portugués ha sido el uso de la dependencia *expl* (expletivo). Para fortalecer la coherencia entre los *treebanks* de origen (que no utilizan la dependencia *expl*) y de destino (que sí la utiliza, de acuerdo con las directrices UD), hemos evaluado el impacto de una transformación automática de los pronombres reflexivos en español y portugués (anotados originariamente como *dobj* o *iobj*) a *expl* (véase el ejemplo de la Figura 1). Otras sustituciones automáticas, como la anotación de algunos pronombres clíticos, determinados usos de la dependencia *case* en el inicio de oraciones subordinadas, o la anotación de expresiones multipalabra están siendo estudiadas para futuros procesos de adaptación.

5 Experimentos

En la presente sección explicamos sucintamente el proceso de corrección de la versión actual del *treebank* de gallego (usado como *gold-standard*), y también evaluamos y discutimos diferentes métodos de transferencia.[3]

[3]Todos los recursos utilizados durante las evaluaciones se pueden obtener en la siguiente dirección `http://grupolys.org/~marcos/pub/sepln16.zip`

Los *parsers* utilizados durante los diferentes experimentos fueron creados con base en los conjuntos de entrenamiento de los *treebanks* de la versión más reciente del proyecto *Universal Dependencies* (1.2). Así mismo, todos los analizadores fueron entrenados con MaltParser (1.8), con la configuración por defecto (dejando, por lo tanto, margen para optimización). Todos los resultados incluyen tanto valores LAS (*Labeled attachment score*) como UAS (*Unlabeled attachment score*).

5.1 *Bootstrapping*

Para evaluar los métodos referidos hemos iniciado la anotación sintáctica del corpus XIADA del siguiente modo: los primeros ≈ 1.000 *tokens* (el número exacto varía en función de la frontera de oración) del subcorpus "xeral" (con 198.231 *tokens* de dominio periodístico general) fueron analizados con un modelo de MaltParser (Nivre et al., 2007) entrenado en una combinación de los *treebanks* UD de portugués (201.845 tokens) y español (382.436 tokens), que fue seleccionado por obtener los mejores resultados en una evaluación subjetiva (al no disponer todavía de datos anotados para la evaluación).

La anotación automática de estos ≈ 1.000 *tokens* fue corregida manualmente por uno de los autores de este trabajo utilizando la herramienta DepAnnotator (Ribeyre, 2015). Una vez finalizada la corrección, aplicamos una estrategia de *bootstrapping* para entrenar un nuevo modelo con los *treebanks* español, portugués, y las oraciones gallegas corregidas. Este proceso se repitió cada ≈ 1.000 *tokens*, hasta llegar a los 12.054 (500 oraciones corregidas), utilizando el corpus resultante como *gold-standard* para evaluar la estrategia propuesta.

5.2 Evaluación

En primer lugar, utilizamos el *gold-standard* de gallego para conocer cómo la distancia lingüística puede influir en el análisis sintáctico de una lengua diferente. La Tabla 1a contiene los resultados de aplicar directamente al gallego *parsers* entrenados sobre los *treebanks* UD de idiomas con diferente grado de distancia lingüística (sueco, inglés, francés, italiano, español y portugués). Durante el proceso de aprendizaje, se utilizó también una variante deslexicalizada (entrenada con corpus sin tokens ni lemas, únicamente con información sintáctica y morfosintáctica) de cada uno de

los *treebanks*, con el objetivo de conocer el impacto de las características léxicas en función de la distancia lingüística.

Los resultados indican que, para el análisis sintáctico del gallego, la distancia lingüística del *treebank* de origen es un factor importante (con diferencias de más de 12 % entre sueco y portugués, por ejemplo).

En relación al impacto de la información léxica en los resultados del *parsing*, los valores obtenidos en las diferentes lenguas parecen indicar que el proceso de deslexicalización es más efectivo en idiomas distanciados léxicamente de la lengua de destino (con los que, por lo tanto, comparten un menor número de palabras). Así, los modelos 'delex' de sueco e inglés obtienen mejores resultados que sus variantes lexicalizadas (entre $\approx 1\,\%$ y $\approx 2\,\%$, en función de la lengua y tipo de evaluación), mientras que en francés e italiano la mejora no es tan clara. Por último, en español y portugués (variedades más próximas al gallego), los modelos con información léxica obtienen sistemáticamente mejores resultados.

Una vez observado el impacto de la distancia lingüística (tanto sintáctica como léxica) en el proceso de transferencia, el siguiente conjunto de evaluaciones analizó (i) combinaciones de los mejores modelos individuales, (ii) la adaptación de las características léxicas —a través de la transliteración del *treebank* portugués— y (iii) la unificación de determinadas directrices de anotación entre *treebanks*.

Así, se han evaluado combinaciones lexicalizadas y deslexicalizadas de español y portugués ('es+pt'), modelos transliterados de portugués a gallego ('pt2')[4] y modelos (tanto de español como de portugués transliterado) en cuyos *treebanks* se han anotado automáticamente los pronombres reflexivos como expletivos ('expl'). Los resultados de estos experimentos se pueden ver en los diferentes bloques de la Tabla 1b.

Los valores de las combinaciones de español y portugués (tanto la variante completa como la deslexicalizada) son ligeramente superiores a los que habíamos obtenido únicamente con los modelos 'pt' y 'pt-delex', lo que sugiere que las combinaciones de recursos complementarios pueden mejorar el análisis de una lengua diferente.

En relación a la adaptación léxico-ortográfica, los resultados del modelo 'pt2' superan en casi 2 % los obtenidos por el *parser* 'pt', por lo que esta estrategia se muestra una vez más efectiva en la adaptación de recursos entre portugués y gallego.

Así mismo la adición del *treebank* español al modelo 'pt2' ('es+pt2') mejora el rendimiento de la transferencia en cerca de 2 % con relación al modelo 'pt2', y en más de 4 % (LAS) en relación al *parser* de portugués.

El último de los niveles definidos (las divergencias entre las directrices de anotación de diferentes *treebanks*) se ha evaluado a través de los modelos 'expl'. A pesar de tratarse de una conversión simple (no se han convertido todos los pronombres reflexivos y expletivos sino únicamente los anotados como "Reflex=Yes" en los corpus de origen), los resultados, tanto en modelos individuales de español y portugués como en la combinación 'es+pt2', sugieren que este tipo de adaptaciones pueden ser útiles durante el proceso de aprendizaje. A este respecto, salvo en el valor LAS del *parser* 'pt2' (con resultados $< 0,01\,\%$), las variantes 'expl' obtienen mejores resultados que aquellos que utilizan la anotación original de los *treebanks* español y portugués.

Así, los diferentes experimentos aquí presentados, realizados en función de los tres parámetros definidos en la sección 4, muestran que para el análisis sintáctico del gallego, la selección de variedades lingüísticas próximas es un factor decisivo en el rendimiento de un *parser* transferido.

Además, la adaptación ortográfica (o léxico-ortográfica, ya que la transliteración modifica directamente las palabras del corpus de origen para adaptarlas a la ortografía de la lengua de destino), es útil para aprovechar recursos sintácticos de portugués en el procesamiento del gallego.

Sobre la uniformización de ciertas variantes de anotación (incluso utilizando un mismo *tagset*, como UD), los experimentos realizados también sugieren que criterios de etiquetación más homogéneos entre las lenguas de origen y destino permiten entrenar *parsers* más precisos.

En suma, la combinación de los diferentes métodos presentados nos permite realizar un análisis sintáctico inicial de un corpus gallego con resultados competitivos con relación al *parsing* de otras lenguas con un mayor número de recursos, por lo que estamos ante un

[4] La transliteración fue realizada con *port2gal*: `http://gramatica.usc.es/~gamallo/port2gal.htm`

Modelo	LAS	UAS
sv	56,39	66,48
sv-delex	**58,24**	**67,92**
en	59,77	68,18
en-delex	**60,84**	**69,52**
fr	66,75	**75,18**
fr-delex	**67,28**	74,45
it	**69,13**	76,54
it-delex	68,98	**77,79**
es	**69,96**	**78,71**
es-delex	69,30	77,59
pt	**71,33**	**79,20**
pt-delex	69,70	76,59

(a) Resultados de modelos individuales (sueco: sv; inglés: en; francés: fr; italiano: it; español: es, y portugués: pt).

Modelo	LAS	UAS
es+pt	**74,21**	**81,65**
es+pt-delex	70,13	77,69
pt2	73,09	80,43
es+pt2	**75,45**	**81,98**
es_expl	70,92	78,82
pt2_expl	73,08	80,51
es_expl+pt2_expl	**75,85**	**82,03**

(b) Resultados de los mejores modelos combinados (líneas superiores) y modelos adaptados: portugués transliterado ('pt2') y español y portugués con conversión automática de la dependencia *expletivo* ('expl').

Tabla 1: Resultados de diferentes *parsers* lexicalizados y deslexicalizados (delex) evaluados sobre el corpus de gallego.

buen punto de partida para la ampliación de un *treebank* para esta lengua.

6 Conclusiones y Trabajo Futuro

En este trabajo hemos presentado una estrategia de combinación y adaptación de *treebanks* de lenguas próximas para el análisis sintáctico de un idioma que, hasta el momento, no disponía de *treebanks* publicados.

El método consiste en combinar recursos de idiomas similares, etiquetados con dependencias universales, y reducir las divergencias tanto léxico-ortográficas como de anotación, para incrementar la precisión de análisis en la lengua de destino.

La etiquetación de un *gold-standard* en gallego, disponible libremente, nos ha permitido probar la eficacia del método propuesto, que no necesita procesos de deslexicalización para transferir analizadores sintácticos de las lenguas origen a la lengua de destino.

Actualmente nos encontramos en proceso de ampliación y corrección del *treebank* inicial presentado en este trabajo, al mismo tiempo que revisamos las directrices de anotación. Un *treebank* de mayor tamaño (así como la publicación de otros recursos UD para gallego) nos permitirá evaluar el impacto de añadir datos propios de gallego a mejores modelos de transferencia.

Así mismo creemos necesario estudiar otras estrategias de adaptación, a través de un análisis más detallado, de recursos de otras lenguas con el fin de aumentar la precisión en los procesos de transferencia. Entre estas estrategias podría estar la realización de un mapeado de las dependencias sintácticas específicas de diferentes idiomas, o el tratamiento homogéneo de estructuras como perífrases verbales, entre otras.

Bibliografía

Cintra, L. F. L. y C. Cunha. 1984. *Nova gramática do português contemporâneo*. Sá da Costa, Lisboa.

De Marneffe, M.-C. y C. D. Manning. 2008. The Stanford typed dependencies representation. En *COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, páginas 1–8, Manchester. ACL.

Gamallo Otero, P. y I. González López. 2011. A grammatical formalism based on patterns of Part of Speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

Ganchev, K., J. Gillenwater, y B. Taskar. 2009. Dependency grammar induction via bitext projection constraints. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volumen 1, páginas 369–377, Singapur. ACL.

Garcia, M. y I. J. González. 2012. Automatic Phonetic Transcription by Phonologi-

cal Derivation. En H. Caseli A. Villavicencio A. Teixeira, y F. Perdigão, editores, *Computational Processing of the Portuguese Language (PROPOR 2012)*, volumen 7243 de *Lecture Notes in Artificial Intelligence*. Springer, Coimbra, páginas 350–361.

Gimpel, K. y N. A. Smith. 2014. Phrase Dependency Machine Translation with Quasi-Synchronous Tree-to-Tree Features. *Computational Linguistics*, 40(2):349–401.

Hwa, R., P. Resnik, A. Weinberg, C. Cabezas, y O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Lynn, T., J. Foster, M. Dras, L. Tounsi, y others. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. En *Proceedings of the First Celtic Language Technology Workshop*, páginas 41–49, Dublin. ACL.

Malvar, P., J. R. Pichel, Ó. Senra, P. Gamallo, y A. García. 2010. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática*, 2(2):31–38.

McDonald, R., S. Petrov, y K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, páginas 62–72, Edimburgo. ACL.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, y J. Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. En *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, páginas 92–97, Sofia. Association for Computational Linguistics.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, y E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*, Estambul. ELRA.

Petrov, S., D. Das, y R. McDonald. 2012. A Universal Part-of-Speech Tagset. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Estambul. ELRA.

Ribeyre, C. 2015. *Méthodes d'Analyse Supervisée pour l'Interface Syntaxe-Sémantique*. Ph.D. tesis, Université Paris 7 Diderot.

Rojo, G., M. L. Martínez, E. D. Noya, y F. M. Barcala. 2015. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6. `http://corpus.cirp.es/xiada/corpus_xiada_2_6.tar.gz`.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, y C. Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, páginas 1631–1642, Seattle. ACL.

Søgaard, A. 2011. Data point selection for cross-language adaptation of dependency parsers. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers (ACL HLT 2011)*, volumen 22, páginas 682–686, Portland. ACL.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin. ACL.

Zeman, D. y P. Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. En *Proceedings of the Workshop on NLP for Less Privileged Language at the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, páginas 35–42, Hyderabad. Asian Federation of Natural Language Processing.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 41-48 recibido 01-04-2016 revisado 09-05-2016 aceptado 02-05-2016

La negación en español: análisis y tipología de patrones de negación*

Negation in Spanish: analysis and typology of negation patterns

M. Antònia Martí, Mariona Taulé, Montserrat Nofre, Laia Marsó
CLiC- Universitat de Barcelona
Gran Via 585, 08007, Barcelona
{amarti, mtaule, mnofre}@ub.edu
marso.laia@gmail.com

M. Teresa Martín-Valdivia
Salud María Jiménez-Zafra
Departamento de Informática
Universidad de Jaén
E-23071 – Jaén, España
{maite, sjzafra}@ujaen.es

Resumen: En este artículo se presentan los criterios aplicados para la anotación del corpus *SFU Review$_{SP}$-NEG* con negación y la tipología lingüística correspondiente. Esta tipología presenta la ventaja de ser fácilmente expresable en términos de un *tagset* para la anotación de corpus, de presentar tipos claramente delimitados, evitando así la ambigüedad en el proceso de anotación, y de presentar una amplia cobertura, es decir, que ha servido para resolver todos los casos que han aparecido. El corpus contiene 400 comentarios y 198.551 palabras. Actualmente está anotado en un 75% y, de un total de 6.331 oraciones revisadas, se han identificado 2.953 estructuras de negación.
Palabras clave: Negación, anotación de corpus, tipos de negación, análisis de opiniones, anotación de la polaridad

Abstract: In this paper we present the criteria applied for the annotation of the *SFU Review$_{SP}$-NEG* corpus and the corresponding linguistic typology. This typology has the advantage that it is easy to express in terms of a tagset for corpus annotation: the types are clearly defined, which avoid the ambiguity in the annotation process, and they present a wide coverage (i.e. they covered/solved all the cases occurring in the corpus). The corpus consists of 400 reviews and 198,551 words. Currently, we have annotated 75% and from a total of 6,331 annotated sentences 2,953 contain at least one negation.
Keywords: Negation, scope, corpus annotation, sentiment analysis, polarity annotation

1 Introducción: Motivación

En el marco del Procesamiento del Lenguaje Natural (PLN) el tratamiento de la negación ha cobrado un especial interés en la medida en que afecta directamente a la polaridad de los textos, en concreto los que expresan opiniones sobre artículos, productos, tendencias y servicios ((Pang et al., 2002), (Wiegand et al., 2010), (Polanyi y Zaenen, 2006), (Councill, McDonald y Velikovich, 2010) y (Morante and Sporleder, 2012)). El carácter idiosincrático de la expresión de la negación en cada lengua requiere un análisis lingüístico específico. Todo proceso de anotación requiere una definición previa de los fenómenos que se van a anotar y una tipología de los mismos. Siendo la negación un fenómeno dependiente de la lengua, las tipologías sobre negación existentes para el inglés -con mucho la lengua en la que se han realizado más esfuerzos en el tratamiento de este fenómeno- no se pueden reutilizar para la anotación de corpus en otras lenguas.

En este artículo presentamos nuestra aproximación al tratamiento de la negación en un corpus del español, el *SFU Review$_{SP}$* (Taboada et al., 2006). En concreto, se presenta una clasificación de las distintas maneras de expresar la negación en base a una tipología; se

* Financiado por fondos FEDER, los proyectos: TIN2015-65136-C2-1-R y TIN2015-71147-C2-2 del MINECO y FPU014/00983 del MECD.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

discuten los conceptos de foco, evento y alcance y se presenta el esquema general de anotación que estamos utilizando para la anotación del corpus antes mencionado.

El corpus *SFU Review$_{SP}$* contiene 400 comentarios repartidos en 50 opiniones (la mitad positivas y la otra mitad negativas) de cada uno de los siguientes temas: coches, hoteles, lavadoras, móviles, ordenadores, música, libros y películas, extraídos de la página web Ciao.es.

En la sección 2, se presenta un estado de la cuestión en la anotación de corpus con negación y se sitúa el corpus que estamos anotando en este contexto; en las secciones 3 y 4 se define y delimita el concepto de negación que está en la base de nuestra anotación. En la sección 5 se describe la tipología desarrollada para la anotación del corpus *SFU Review$_{SP}$*. En la sección 6 se presenta brevemente el esquema de anotación general y en la sección 7 se presentan las conclusiones y se apuntan las líneas futuras.

2 Antecedentes: corpus anotados con negación

Los corpus anotados con negación disponibles, todos ellos del inglés, son de tamaño muy diverso (desde 20.000 a 1.000 oraciones anotadas) y difieren en el sistema de anotación empleado. Tienen en común que todos ellos anotan tanto las partículas negativas como el alcance; sólo uno de ellos, el de Blanco y Moldovan (2011), marca el foco y sólo el *ConanDoyle-neg* (Morante y Daelemans, 2012) marca el evento (véase la sección 4).

El primer corpus anotado con negación fue *BioInfer* (Pyysala et al., 2007), que incluye 1.100 oraciones extraídas de *abstracts* de artículos biomédicos. Se etiquetan los predicados con negación, pero no su alcance. Destacan por su tamaño los corpus *BioScope* (Vincze el al. 2008) y *SFU Review$_{EN}$* (Konstantinova et al., 2012), que además de anotar la negación también incluyen la anotación de las expresiones especulativas y su alcance, información clave para identificar los enunciados subjetivos.

BioScope es un corpus formado por textos biomédicos en el que se anotaron por primera vez tanto las partículas negativas (y especulativas) como su alcance. El corpus contiene más de 20.000 oraciones anotadas, de las cuales el 13% incluye algún tipo de negación. *BioScope* está formado por textos clínicos (6.383 oraciones), *abstracts* de artículos de biología (11.871 oraciones) y los 9 artículos completos de biología (2.670 oraciones) del corpus *Genia Event* (Kim et al., 2008).

Konstantinova et al., (2012) han anotado con negación (y especulación) el corpus *SFU Review$_{EN}$*. Este corpus está formado por un total de 400 comentarios (17.263 oraciones de las cuales el 18% contienen negación) escritos en inglés y de distinta temática -comentarios sobre libros, coches, ordenadores, utensilios de cocina, hoteles, películas, música y teléfonos- extraídos de la página web Epinions.com. El corpus *SFU Review$_{EN}$* contiene 50 documentos de cada una de las temáticas seleccionadas y cada uno de ellos tiene asignada una etiqueta que indica si se trata de un comentario positivo o negativo. Para la anotación de la negación siguen fundamentalmente los criterios utilizados en *BioScope* adaptados al dominio de los comentarios (Konstantinova y Sousa, (2011).

Entre los corpus de menor tamaño, en el mismo ámbito de los comentarios, cabe destacar el corpus *Product Review* (Councill, McDonald y Velikovich, 2010) formado por 268 comentarios de productos extraídos de *Google Product Search*. El corpus contiene 2.111 oraciones de las cuales 679 incluyen negación, es decir, el 32%. Los autores utilizan este corpus para desarrollar un sistema cuyo objetivo es identificar el alcance de la negación en el contexto del análisis de los sentimientos.

ConanDoyle-neg es el corpus de entrenamiento y evaluación desarrollado para la tarea 10 de SemEval-2010, *Linking events and their participants in discourse*[1] (Ruppenhofer et al. 2010). El corpus incluye textos literarios de dos obras de Arthur Conan Doyle[2], anotados con las partículas negativas, su alcance y el evento o propiedad explícitamente negada. El corpus se encuentra en formato xml TIGER/SALSA (Erk y Padó, 2004)[3] y, además de la negación, también está anotado con correferencia, roles semánticos y argumentos implícitos. El corpus contiene 4.423 oraciones

[1] http://www.coli.uni-saarland.de/projects/semeval2010_FG/

[2] Las obras son: *The Hound of the Baskervilles* y *The adventure of Wisteria Lodge.*

[3] El corpus está disponible en: http://www.clips.ua.ac.be/BiographTA/corpora.html

de las cuales el 22,49% incluyen al menos una partícula negativa.

Blanco y Moldovan (2013) seleccionaron 3.993 negaciones verbales del corpus *PropBank* (Palmer et al., 2005) para establecer el alcance y el foco de estas negaciones con el objetivo de representar su semántica. Siguiendo a Huddeleston y Pullum, (2002), definen el foco como la parte del alcance que está más destacada y explícitamente negada.

Los corpus *SFU Review*, *Product Review*$_{EN}$ y *ConanDoyle-neg* se basan o inspiran en la guía de anotación (Vincze, 2010) utilizada para anotar *BioScope*. Las diferencias residen principalmente en la manera de anotar el alcance, en concreto, qué elementos quedan dentro o fuera del mismo.

En este artículo utilizaremos el corpus *SFU Review*$_{SP,}$ y lo anotaremos con negación, siguiendo parcialmente el sistema ABSA utilizado en la tarea 12 de SemEval[4]. *SFU Review*$_{SP}$*-NEG* tiene un total de 198.551 palabras. El corpus está constituido por 400 comentarios, de los cuales ya se ha anotado un 75%[5], lo que corresponde a un total de 6.331 oraciones, de las cuales 2.953 contienen al menos una estructura negativa. De éstas, 1.430 contienen una sola estructura negativa y 620 contienen más de una. Está organizado en ocho bloques de 50 ficheros cada uno. De estos 50 ficheros, 25 corresponden a opiniones positivas y 25 a opiniones negativas. Cada fichero contiene la opinión de un usuario acerca de un producto. Además, el corpus está anotado morfológicamente, con su categoría gramatical y lema correspondiente.

3 Definición y delimitación de la negación

La negación es un fenómeno lingüístico mediante el cual se invierte el valor de verdad de la unidad lingüística (proposición, sintagma o palabra) a la que se aplica. En las lenguas la negación se expresa mediante diversos mecanismos, siendo los más comunes el uso de partículas de negación sintácticamente independientes (‘no’, ‘nunca’, ‘nadie’, etc.), prefijos (‘imposible’, ‘ilícito’) y frases hechas (‘en la vida’), entre otros.

En nuestra aproximación al tratamiento de la negación para la anotación del corpus en español nos hemos centrado, de momento, en la negación a nivel sintáctico, es decir, la que afecta a sintagmas y a la oración. Queda excluida de nuestra tipología la negación léxica (‘dudar’, ‘ausencia de’, ‘falta de’, etc.) y la morfológica, es decir, palabras con un afijo de negación (‘descontento’, ‘incoherente’).

Esta aproximación es acorde con la definición propuesta por la RAE (2009: 3631): “En sus múltiples manifestaciones gramaticales, la negación se considera un operador sintáctico en un sentido similar al de los cuantificadores y determinados adverbios, es decir, un elemento que condiciona (…) la referencia de otras unidades que se hallan en su ámbito de influencia”. Las palabras que expresan negación pertenecen a diferentes categorías gramaticales: adverbios (‘no’, ‘jamás’, ‘nunca’, ‘tampoco’, ‘nada’); pronombres (‘nada’, ‘nadie’, ‘ninguno’, ‘nunca’); conjunciones (‘ni’, ‘sino’); preposiciones (‘sin’, ‘en vez de’, etc.); determinantes indefinidos (ningún, ninguna, etc.). Como se puede observar, algunas palabras como ‘nada’ pueden pertenecer a más de una categoría.

4 Foco y alcance de la negación: <scope> y <event>

En los tratados gramaticales ((RAE, 2009) y (Bosque y Demonte, 1999))- se distingue entre el foco y el alcance de la negación. Según la gramática, el alcance de la negación corresponde a la totalidad de palabras afectadas por la misma, mientras que el foco corresponde a la palabra o sintagma dentro del alcance que se niega explícitamente.

(1) No pienso ir al concierto ni contigo ni con nadie. (RAE, 2009: 3638)

En la oración (1), el alcance sería la oración entera y el foco ‘ni contigo ni con nadie’. Lo que se niega no es el hecho de ir al concierto sino el hecho de ir acompañado (foco).

El modo en que estos dos conceptos se han plasmado en los diferentes corpus anotados es muy diverso. En lo que se refiere al alcance, la RAE (2009: 3655) considera que si el sujeto es postverbal, queda incluido en el alcance, mientras que si es preverbal, queda fuera. De los corpus descritos en la sección 2, solo en el corpus de *ConanDoyle-neg* el sujeto se incluye en el alcance.

Respecto de la partícula negativa, la RAE no se pronuncia sobre su inclusión o no inclusión

[4] http://alt.qcri.org/semeval2015/task12/

[5] Faltan por anotar los comentarios de películas y ordenadores.

en el alcance. De los corpus mencionados, solo *Bioscope* la incluye en el alcance.

La mayoría de corpus no anotan el foco, por ser un componente de la negación de carácter semántico-pragmático, que muchas veces resulta difícil de identificar. La resolución del foco requiere las más de las veces disponer de información contextual que no siempre se encuentra disponible. Entre los corpus revisados, sólo Blanco y Moldovan (2011) lo tratan, ya que su objetivo es la representación semántica de la negación. Como contrapartida, en algunos corpus anotados con negación, por ejemplo en *ConanDoyle-neg*, se anota un componente de la misma, el evento, que no aparece en los tratados gramaticales, y con el que se pretende marcar el elemento directamente afectado por la negación, siempre dentro del alcance.

En nuestra propuesta, el alcance siempre corresponde a un constituyente sintáctico, es decir un sintagma o una oración (2) y el sujeto queda incluido cuando la negación afecta al predicado verbal. En el sistema de anotación se marca con la etiqueta <*scope*>[6].

(2) a. [Sin mirar el aceite.]$_{sn}$
b. [Cero fiabilidad.]$_{sn}$
e. [No llegaron a tiempo.]$_{o}$

En lo que respecta al foco, no lo hemos tratado en la versión actual del corpus, pero sí que hemos considerado interesante marcar la palabra directamente negada por el operador negativo, es decir, el evento o núcleo del constituyente que se niega (el nombre, el adjetivo, el verbo y el adverbio). Utilizamos la etiqueta <*event*> para anotar este elemento. En el caso de los sintagmas preposicionales introducidos con la partícula negativa 'sin', el evento es el sintagma nominal o la oración afectados por la preposición. En el caso de los verbos copulativos, el evento de la negación es el verbo más el atributo. En el caso de los verbos con complemento predicativo, este último se incluye también en el evento. En el caso de las perífrasis ('no acaba de salir'), las colocaciones ('no da problemas') y los verbos '*light*' con complemento ('no se dio por vencido', 'no decir mucho [a cerca de/sobre/…]') el evento incluye a toda la forma verbal compleja.

[6] En los ejemplos, utilizamos los corchetes para marcar el alcance y subrayamos el evento.

Son casos especiales de evento y alcance los pronombres indefinidos de negación cuando se usan antepuestos al verbo, es decir, cuando no van acompañados de la partícula 'no' (3). En 'Nadie [='ninguna persona'] vino', el alcance y el evento coinciden en la forma 'nadie' (3a), del mismo modo que en 'Ningún niño vino' el alcance es 'ningún niño' y el evento 'niño' (3b). En estos casos no se niega el verbo, sino que se le asigna un sujeto que tiene como referente el conjunto vacío (RAE: 3646).

(3) a. [Nadie] vino.
b. [Ningún niño] vino.

5 Tipología

Hemos construido la tipología de expresiones de negación teniendo en cuenta, por un lado, los principios básicos contenidos en las gramáticas descriptivas y normativas ((Bosque y Demonte, 1999) y (RAE, 2009)) y, por otro, la coherencia, la sistemática y la máxima sencillez en la metodología y el conjunto de etiquetas (*tagset*) para la anotación del corpus. Suele ocurrir que en los corpus aparecen estructuras, construcciones o expresiones que no están contempladas en las gramáticas, por lo que se plantean problemas a la hora de expresar el contenido de las mismas en términos de un *tagset*. Es por ello que nuestra tipología, si bien está basada en la gramática, garantiza que es consistente desde el punto de vista de la anotación y que los tipos definidos (o categorías) constituyen clases claramente disjuntas, lo que facilita el proceso de anotación. Todas las expresiones de negación que hemos hallado en el corpus *SFU Review*$_{SP}$, pertenecen a una clase de nuestra tipología, por lo que queda probada suficientemente su validez y consistencia teniendo en cuenta que el corpus tiene un tamaño suficiente para garantizar que incluye una amplia gama de estructuras de negación.

Para definir nuestros tipos de expresiones de negación hemos tenido en cuenta tanto la estructura sintáctica como su interpretación semántica, es decir, si la estructura negativa expresa o no una negación. La tipología se estructura en torno a dos grandes bloques, la expresión de la negación simple (5.1) y compleja (5.2), ambas con la etiqueta '*neg*' asociada. En (5.3) se presentan las estructuras negativas que no expresan negación.

5.1 Negación simple

Se considera 'negación simple' la expresión de la negación mediante una única partícula. Esta partícula va antepuesta al evento y puede ser un adverbio ('no', 'jamás', 'apenas', 'nunca') (4a-b), un pronombre antepuesto al verbo ('nadie', 'nada') (4c), o una preposición ('sin') (4d).

(4) a. (…) para conductores que **apenas**$_{\textbf{adv}}$ tocan el coche.
b. **Nunca**$_{\textbf{adv}}$ tienen las piezas de recambio en el taller.
c. **Nadie**$_{\textbf{pr}}$ quedará decepcionado en este aspecto.
d. **Sin**$_{\text{p}}$ conexión**.**

Incluimos también en esta categoría la coordinación de oraciones negativas simples (5).

(5) a. [Ni puedo desear más] [ni puedo contentarme con menos].
b. El aire acondicionado [ni enfría] [ni calienta].

5.2 Negación compleja

Dentro del tipo 'negación compleja' incluimos la expresión de la negación mediante dos o más partículas, continuas (6) o discontinuas (7)[7], la primera de las cuales suele expresar negación, mientras que la segunda puede expresar también negación (7) reforzando así la primera (véase sección 5.2.1), o puede modular el valor de la negación (6) (véase sección 5.2.2).

(6) **Casi no** llega.
(7) **No** vino **nunca**.

En nuestro sistema de anotación, las partículas de la negación compleja tienen asociada la etiqueta *<discid='1n/1c, 2n/2c,…'>* (discontinua).

A continuación, describimos más detalladamente estas dos clases de negación.

5.2.1 Refuerzo de la negación

En español es frecuente que las expresiones de negación se refuercen mediante una segunda partícula (8a-11a). Es lo que en nuestro sistema de anotación denominamos refuerzo de la negación. Estas expresiones siempre se pueden parafrasear anteponiendo al verbo la segunda partícula negativa, dando como resultado una negación simple (8b-11b):

(8) a. Ustedes **no** pueden hacer **nada.**
b. Ustedes **nada** pueden hacer.
(9) a. En los Nokia que he utilizado **no** he tenido **nunca** este problema.
b. **Nunca** he tenido este problema en los Nokia que he utilizado.
(10) a. Allí **no** me esperaba **nadie**.
b. **Nadie** me esperaba allí.
(11) a. Puede que **ni** siquiera los hayan escuchado **jamás**.
b. Puede que **jamás** los hayan escuchado.

Cuando se da la coordinación de dos estructuras negativas en un mismo sintagma también lo consideramos dentro de esta categoría (12), ya que la repetición de partículas negativas ('ni… ni…') también da idea de refuerzo.

(12) a. No comió **ni** pan **ni** vino.
b. No me sentí **ni** libre **ni** poderoso.
c. **Sin** agua **ni** comida.

5.2.2 Negación con modificadores

La negación, al igual que muchos otros fenómenos lingüísticos, no es categorial, sino que puede presentar gradación. Existen diferentes mecanismos para expresar esta gradación, que en nuestro sistema de anotación denominamos modificadores y que pueden ser incrementadores, cuando potencian la negación (13) y decrementadores, cuando la atenúan (14).

(13) a. Mi coche no frena **en absoluto**.
b. No te molesta **nada**[8].
(14) a. No estoy **muy** segura.
b. No tiene **mucho** sentido.
c. No da **demasiadas** opciones de idioma.

En nuestro sistema de anotación anotamos los incrementadores con la etiqueta

[7] En el 75% del corpus que se ha anotado, se han identificado un total de 2.375 expresiones negativas simples y complejas continuas, -de las cuales 229 no expresan negación- y 449 complejas discontinuas.

[8] Nótese que 'nada' se ha interpretado como un adverbio, en el sentido de 'en absoluto', pero podría ser también un pronombre. Solo el contexto ha permitido desambiguarlo.

<*increment*> y los decrementadores con la etiqueta <*reduction*>. En estos casos, la partícula negativa aparece en primer lugar, y solo en casos de dislocación, la partícula negativa va en segundo lugar, precedida por el modificador (15).

(15) **Más** equivocado **no** pude estar.

5.2.3 Comparativas con negación

Un tipo particular de negación es el que se da en una estructura que expresa comparación, anotadas con la etiqueta '*comp*'. Son siempre estructuras discontinuas (16).

(16) a. **No** me gusta **tanto como** lo otro.
b. Mi amor **no** iba a ser **más** pequeño **que** yo.
c. El ambiente de este local es agradable pero **no** (verbo elidido) **tanto como** el del otro.
d. El motor **no** es **todo lo** potente **que** debería.

5.2.4 Frases hechas que expresan negación

Existen construcciones complejas lexicalizadas que expresan negación (17). En nuestro sistema de anotación las consideramos como una sola unidad, de manera que formarían parte del léxico de partículas de negación.

(17) a. En la vida.
b. En toda mi vida.
c. Ni lo sueñes.

Cabe destacar que los casos que presentan variables (17a y 17b) se tratan como unidades diferentes. Estas expresiones complejas pueden incluir (17c) o no (17a y 17b) una partícula negativa.

5.3 Estructuras negativas que no expresan negación

Existen expresiones que aunque contienen partículas de negación, semánticamente o bien no expresan negación o bien expresan un contraste o contraposición entre dos o más opciones o posibilidades. Dentro de este tipo distinguimos las estructuras simples de las complejas.

5.3.1. Estructuras simples que no expresan negación

Dentro de esta clase se incluyen las partículas negativas en oraciones interrogativas (18), las partículas negativas en contextos que no expresan negación (19) y las partículas negativas con valor expletivo (20).

(18) El coche lo compré para viajar, **no**?
(19) **Nada_más** darle al contacto[9].
(20) No pienso irme hasta que **no** vengas.

Las frases hechas o expresiones lexicalizadas con partícula negativa que no expresan negación (21) como se tratan como una única expresión léxica multipalabra, las incluimos en esta clase.

(21) a. Visto y no visto.
b. Sin pena ni gloria.
c. No hace más que.
d. No hay más que.

Todas estas estructuras se anotan con la etiqueta '*noneg*'.

5.3.2. Estructuras complejas que no expresan negación

Dentro de esta categoría se incluyen las estructuras de contraste entre dos o más elementos que se contraponen bien para introducir una corrección (22a) o para añadir información nueva (22b). En otros casos se expresa una contraposición respecto de un límite o cota que se explicita (23).

(22) a. **No** vinieron 2 soldados, **sino** 6.
b. **No_sólo** lleva rueda de recambio **sino_también** caja de herramientas.
(23) a. BMW **no** suele poner **más_que** lo que considera necesario.
b. **No** veo otra salida **que** pedirle otra lavadora.

La oración de (23a) se parafrasea como 'BMV suele poner sólo lo que considera necesario', de manera que se especifica el límite en las inversiones de BMW. La oración de (23b) se parafrasea como 'La única salida es pedirle otra lavadora', por lo tanto lo que se

[9] Nótese que 'nada_más' lo tratamos como un único elemento léxico.

expresa es la única opción posible, el límite. En ningún caso se expresa una negación.

Todas estas estructuras se anotan con la etiqueta '*contrast*'.

6 Esquema de anotación

En esta sección se describen brevemente los atributos utilizados en la anotación de la negación del corpus *SFU Review$_{SP}$-NEG* recogidos en el esquema general de anotación de la Figura 1.

La etiqueta <*review polarity*> indica la polaridad de todo el comentario, que puede ser positiva o negativa. En *SFU Review$_{SP}$-NEG* solo se anotan las oraciones (<*sentence*>) que contengan al menos una negación. Cuando la oración contiene más de una estructura negativa (<*neg_structure*>) se asigna el valor 'yes' al atributo <*sentence complex*> y cuando solo incluye una única estructura negativa el valor 'no'.

```
<review polarity= 'positive/negative'
 <sentence complex='yes/no'>
  <neg_structure
    polarity='positive/negative/neutral'
    change='yes/no'

polarity_modifier='increment/reduction'
    value='neg/contrast/comp/noneg'
    <scope>
      <negexp discid='1n/1c'>
      </negexp>
      <event>
      </event>
    </scope>
  </neg_structure>
</sentence>
```

Figura 1: Esquema general de anotación.

La etiqueta <*neg_structure*> tiene asociados cuatro atributos:

- <*polarity*>: indica la orientación positiva, negativa o neutra de la estructura negativa (p.e.: 'no es un chico malo', 'no es un chico bueno', 'no es un chico alto').
- <change>: indica si, debido a la negación, la estructura negativa ha visto modificada o no totalmente su polaridad (p.e.: 'chico bueno' vs. 'chico no bueno') o su significado ('chico alto' vs. 'chico no alto').
- <polarity_modifier>: indica si en la estructura negativa hay algún elemento que modifica o matiza su polaridad (p.e.: 'chico bueno' vs. 'chico no muy bueno'). Este atributo tiene dos valores posibles: 'increment' para indicar que se incrementa la polaridad (p.e.: 'no me arrepiento para nada') y 'reduction' para cuando se reduce (p.e.: 'no lo he utilizado mucho').
- <value>: indica el significado expresado por la estructura negativa. Tiene cuatro valores posibles: 'neg' cuando indica negación; 'contrast' cuando expresa contraste u oposición ente términos; 'comp' cuando expresa comparación o desigualdad entre términos; y 'noneg' para indicar las estructuras que contienen una partícula negativa pero que no niegan.

La etiqueta <*scope*> se usa para anotar el alcance de la negación, incluyendo la propia partícula negativa y <*negexp*> para delimitar la palabra o palabras que expresan negación. <*negexp*> puede llevar asociado el atributo <*discid*>, que se aplica en aquellas estructuras negativas donde hay más de un elemento y los casos de estructuras negativas discontinuas. La etiqueta <*event*> sirve para marcar la palabra o palabras directamente negadas por el operador negativo.

7 Conclusiones y líneas futuras

En este artículo hemos presentado los diferentes tipos de negación en español y el sistema de etiquetas utilizado para la anotación del corpus *SFU Review$_{SP}$-NEG*, el primer corpus del español anotado con esta información. Aunque se ha anotado sólo un 75% del corpus, el número de casos observados y anotados (2.050) permite suponer que nuestra tipología es completa y abarca el fenómeno en su totalidad. El corpus es de libre disposición[10].

Tenemos previsto como líneas futuras, por un lado, el tratamiento del foco y de la negación léxica y morfológica y, por otro, terminar la anotación del corpus.

Bibliografía

Blanco E. y D. Moldovan. 2013. Retrieving implicit positive meaning from negated statements. *Natural Language Engineering*, 20 (4): 501-535. Cambridge University Press.

[10] http://sinai.ujaen.es/sfu-review-sp-neg/

Bosque I. y V. Demonte. 1999. Gramática Descriptiva de la Lengua Española, Vol. 2. Espasa Calpe, España.

Councill, I. G., R. McDonald, y L. Velikovich, L. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 51-59, Uppsala, ACL.

Erk K., y S. Padó. 2004. A powerful and versatileXML format forrepresenting role-semantic annotation. *Proceedings of 4th International Conference on Language Resources and Evaluation* (LREC'04), Lisboa, Portugal.

Huddleston, R.D. y G. K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Kim J.D., T. Ohta y J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics, 9:10.

Konstantinova, N., S. C de Sousa, N. P. Díaz, N. P. Cruz, M. J. Maña, M. Taboada y R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), páginas 3190-3195, Turkey.

Konstantinova, N. y S. C de Sousa. 2011. Annotating Negation and Speculation: the Case of the Review Domain. *Proceedings of the Student Research Workshop associated with RANLP 2011*, páginas 139-144, Bulgaria.

Morante, R. y W. Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), páginas 1563-1568, Turkey.

Morante, R. y C. Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2), 223-260.

Palmer, M., P. Kingsbury y D. Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 21 (1).

Pang, B., L. Lee, y S. Vaithyanathan, S. 2002.Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*: 79-86. ACL.

Polanyi L., Zaenen, A. 2006. Contextual Valence Shifters. *Computing affect and attitude in text: Theory and applications, 20:* 1-10. *The Information Retrieval Series.*

Pyysala S., F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen y T. Salakosk. 2007. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics, 8: 50.

RAE. 2009. Nueva Gramática de la Lengua Española. Vol. 2. Espasa Libros, España.

Ruppenhofer J., C. Sporleder, R. Morante, C. Baker y M. Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. *Proceedings of the 5th Workshop on Semantic Evaluations (ACL 2010)*, páginas 45-50, Suecia.

Taboada, M., C. Anthony y K. Voll. 2006. Methods for creating semantic orientation dictionaries. *Proceedings of the 5th Conference on Language Resources and Evaluation* (LREC'06), páginas 427-432.

Vincze, V., Szarvas G., Farkas R., Móra G. y Csirik J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9:1-9.

Vincze, V. 2010. Speculation and negation annotation in natural language texts: what the case of bioscope might (not) reveal. *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 51-59, Uppsala, ACL.

Wiegand, M., A. Balahur, B. Roth, D. Klakow, y A. Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 60-68, ACL.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 49-56 recibido 28-03-2016 revisado 21-04-2016 aceptado 29-04-2016

Ampliación de lexicones de opinión específicos de dominio usando representaciones continuas de palabras

Expansion of domain-specific opinion lexicons using word embeddings

Tomás López Solaz
Universidad de Sevilla
tlopez2@us.es

Fermín L. Cruz
Universidad de Sevilla
fcruz@us.es

Fernando Enríquez
Universidad de Sevilla
fenros@us.es

Resumen: En este trabajo abordamos la ampliación de lexicones de opinión específicos de dominio a partir de textos del dominio elegido. El método se basa en la construcción de clasificadores que catalogan las palabras de entrada como positivas, negativas o neutras, y en un criterio estricto de selección de las palabras que pretende garantizar la precisión de las nuevas incorporaciones al lexicón. Se utilizan representaciones continuas de palabras (*word embeddings*) como espacio de características de los clasificadores. Los resultados confirman que dichas representaciones contienen información relativa a la polaridad de las palabras, obteniéndose una precisión en la selección de los candidatos y en la estimación de su polaridad de alrededor del 94 % para los tres dominios analizados, con una cobertura en torno al 50 % de las palabras de opinión contenidas en los textos de partida.
Palabras clave: Análisis del sentimiento, lexicones de opinión, representaciones continuas de palabras

Abstract: In this work we present a domain-specific opinion lexicon expansion method. The method is based on classifiers which categorize words as positive, negative or neutral, and a strict selection criteria of words intended to ensure the precision of the new additions to the lexicon. We use word embeddings as the feature space of the classifiers. The results confirm that these representations contain information on the polarity of the words, obtaining a precision in the selection of candidates and the estimation of its polarities of about 94 % for the three domains analyzed, covering around 50 % of the opinion words contained in the initial texts.
Keywords: Sentiment analysis, opinion lexicons, word embeddings

1 Introducción

En el contexto del análisis de opiniones textuales, un recurso de amplia utilización son los lexicones de opinión, diccionarios de términos con connotaciones subjetivas, acompañados de estimaciones de las implicaciones positivas o negativas de los mismos. Existen muchas aproximaciones distintas: lexicones de propósito general o específicos de un dominio, lexicones con estimaciones binarias de la polaridad o con un mayor grado de granularidad (incluyendo información no sólo de la polaridad sino también de la intensidad de las connotaciones afectivas), lexicones basados en palabras o en recursos lexico-semánticos como WordNet, etc. En la sección 2 comentamos algunos trabajos sobre generación de lexicones de opinión.

El sistema de extracción automática de opiniones orientado al dominio TOES (Cruz et al., 2013) se basa en recursos generados a partir de documentos de opinión anotados. A pesar de ser supervisado, se intenta usar el menor número posible de documentos anotados en este proceso, ampliando posteriormente los recursos a partir de otros documentos sin anotar. Entre los recursos generados se encuentra un lexicón de opiniones a nivel de palabras orientado al dominio. En este trabajo exponemos un método de ampliación de estos lexicones de opinión, que a partir de un conjunto de documentos sin anotar extrae nuevos términos a incluir en el lexicón, junto a una estimación de sus polaridades. Nuestra intención es asegurar en todo momento una altísima precisión, de manera que podamos estar seguros de que la inmensa mayoría de los términos añadidos y sus polaridades son correctos.

Como objetivo secundario de este trabajo, pretendemos corroborar las bondades de las representaciones continuas de palabras o

ISSN 1135-5948
© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

word embeddings al ser aplicadas a tareas de clasificación de la polaridad, en concreto mediante el uso de la herramienta *word2vec*. Son manifiestas las relaciones entre la geometría de las representaciones vectoriales de palabras obtenidas mediante dicha herramienta y el contenido semántico de las mismas, de manera que por ejemplo palabras semánticamente próximas son representadas mediante vectores similares (a diferencia de lo que ocurre con representaciones vectoriales discretas como las basadas en *tf-idf* u otras). Nuestra hipótesis de partida es que también existe una relación similar entre las representaciones continuas de palabras de la misma polaridad, como ya parecen confirmar algunos trabajos previos que se expondrán en la siguiente sección.

La estructura del resto del artículo es la que se expone a continuación. En la sección 2 se repasan algunos trabajos previos relacionados con la inducción de lexicones de opinión y su ampliación, y se introducen las principales técnicas y algunas aplicaciones de las representaciones continuas de palabras. En la sección 3 se define la tarea de ampliación de lexicones de opinión, se detalla el método utilizado para llevarla a cabo. En la sección 4 se muestran los resultados experimentales y, finalmente, en la sección 5 se comentan las conclusiones del trabajo y se proponen líneas de continuación del mismo.

2 Trabajos relacionados

2.1 Inducción de lexicones de opinión

En los últimos años se han propuesto diversos métodos para la creación de lexicones de opinión. Existen trabajos que abordan la tarea de forma totalmente manual (Stone, 1966; Cerini et al., 2007), y otros que se basan en un corpus anotado a nivel de opiniones, a partir del cual extraen las palabras de opinión y las polaridades de las mismas (Cruz et al., 2013). Algunos de ellos se basan en recursos léxico-semánticos para obtener los lexicones de opinión de manera automática o semiautomática. Por ejemplo, en (Kamps et al., 2004) se utiliza una función de distancia semántica sobre WordNet que es utilizada para decidir la polaridad de las palabras del lexicón en función de la distancia a una semilla positiva ("good") y una negativa ("bad"). Otros trabajos se basan en una idea similar pero utilizando un mayor número de semillas (Hu y Liu, 2004; Kim y Hovy, 2004). En los trabajos de Esuli y Sebastiani (Esuli y Sebastiani, 2006; Baccianella, Esuli, y Sebastiani, 2010) la idea de partida es que si una palabra tiene una polaridad determinada, es probable que las palabras de su glosa (pequeñas definiciones de los términos que conforman WordNet) compartan dicha polaridad. Partiendo de unas semillas positivas y negativas, y aplicando un proceso iterativo de expansión basado en las relaciones de sinonimia y antonimia de WordNet, se obtienen conjuntos de entrenamiento para las clases positiva y negativa, usando las glosas de los términos de cada conjunto. Se construye un clasificador de textos mediante representaciones discretas de las palabras (tf-idf) que es aplicado a todas las palabras que componen el recurso WordNet, para obtener así estimaciones de las polaridades de todas ellas. Posteriormente se aplica un algoritmo de paseo aleatorio (*PageRank*) sobre un grafo construido a partir de las palabras y sus glosas, que refina los valores de orientación semántica asignados a las palabras en el recurso final. Las mismas ideas son aplicadas en (Cruz et al., 2014), variando algunas partes del proceso y aplicando un algoritmo más potente de paseo aleatorio (Cruz et al., 2012) adaptado a la tarea de cálculo de la polaridad. La principal debilidad de los acercamientos basados en recursos léxico-semánticos es su disponibilidad para nuevas lenguas, y que las estimaciones de polaridad obtenidas son independientes del dominio (cuando existen diferencias claras en el vocabulario utilizado en distintos ámbitos).

Otros trabajos se basan en corpus de textos sin anotaciones, como hacemos en el método propuesto en este trabajo. Algunos de ellos se basan en definir conjuntos semilla de términos positivos y negativos y calculan las polaridades de las palabras del lexicón a partir de construcciones conjuntivas observadas en el corpus (Hatzivassiloglou y McKeown, 1997) o de coocurrencias en un contexto dado (Turney y Littman, 2003; Yu y Hatzivassiloglou, 2003) entre las semillas y las palabras objetivo.

2.2 Ampliación de lexicones de opinión

En nuestro trabajo la tarea no consiste en la generación de lexicones de opinión desde cero, sino que pretendemos ampliar lexicones previamente obtenidos a partir de un corpus

anotado (Cruz et al., 2013). Existen trabajos previos que comparten este objetivo; en (Kanayama y Nasukawa, 2006) se propone un algoritmo de expansión automática de lexicones basado en la coherencia contextual, esto es, la tendencia a que términos con la misma polaridad aparezcan en un mismo contexto. En (Qiu et al., 2011) se utilizan los términos del lexicón de partida para identificar patrones sintácticos entre las palabras de opinión y las palabras objetivo de su contexto a las que afectan. Estos patrones son aplicados sobre otras apariciones de estas palabras objetivo conformando un proceso de *bootstrapping*. La polaridad de las nuevas palabras de opinión encontradas es decidida de nuevo a partir de reglas contextuales. En (Cruz et al., 2011) se construye un grafo entre las palabras que participan en expresiones conjuntivas en el corpus de partida. Los nodos que representan a las palabras del lexicón inicial son anotados con los valores de polaridad, y posteriormente se aplica un algoritmo de paseo aleatorio (Cruz et al., 2012) sobre el grafo para obtener las polaridades del resto de nodos. En (Molina-González et al., 2015) se expande un lexicón de opiniones en español independiente del dominio (Molina-González et al., 2013), adaptándolo a un dominio concreto a partir de documentos de opinión del dominio de interés. La selección de las palabras y su polaridad se realiza automáticamente a partir de una fórmula basada en la frecuencia de aparición de las mismas en opiniones positivas y negativas.

2.3 Representaciones continuas de palabras

Cuando se emplean técnicas de aprendizaje automático sobre textos es frecuente utilizar representaciones numéricas de las palabras o los documentos. Tradicionalmente se han empleado representaciones discretas, por ejemplo vectores de unos y ceros que indican la aparición o no de las palabras del vocabulario en un documento, u otras métricas que acumulen más información (basadas generalmente en la frecuencia de aparición de las palabras). Estas representaciones adolecen de un problema de falta de sentido geométrico; por ejemplo dos palabras cualesquiera siempre estarían representadas por vectores con una única componente no nula, de manera que la distancia entre los vectores no está relacionada con la cercanía semántica de las palabras. En los últimos años se han producido esfuerzos encaminados a obtener representaciones continuas de las palabras y los documentos, que sobrepasen la limitación anterior. En estas representaciones, los vectores sí pueden ser interpretados geométricamente, de manera que dos palabras o documentos cercanos desde un punto de vista semántico estarían representados por vectores similares (y viceversa). Las representaciones continuas (también conocidas como *word embeddings*) más utilizadas en los últimos años han sido Latent Semantic Indexing (Dumais, 1995), Latent Semantic Analysis (Dumais, 2004) y, más recientemente, técnicas basadas en redes neuronales (Turian, Ratinov, y Bengio, 2010; Huang et al., 2012; Mikolov et al., 2013b). De estas últimas, *word2vec* (Mikolov et al., 2013a) es la herramienta que más popularidad ha alcanzado. Se basa en la construcción previa de un modelo a partir de un corpus de textos, a partir del cual se entrena una red neuronal de dos niveles en la que las entradas son los contextos de una palabra dada y la salida la palabra en cuestión (o viceversa, intercambiando las entradas por las salidas). El sentido semántico de la geometría de los vectores obtenidos mediante *word2vec*, que puede apreciarse en los ejemplos de operaciones entre vectores mostradas en (Mikolov et al., 2013a), los hace apropiados como entrada a algoritmos de aprendizaje automático para resolver múltiples tareas de NLP (Socher et al., 2013; Tang et al., 2014; Pavlopoulos y Androutsopoulos, 2014; Kim y de Marneffe, 2013), incluida la inducción de lexicones de opinión (Pablos, Cuadros, y Rigau, 2015).

3 Propuesta

Partimos de lexicones de opinión obtenidos a partir de corpus de textos de dominios concretos, anotados a nivel de opiniones. Estos lexicones incluyen, además de la polaridad de las palabras, una probabilidad de que el término sea o no de opinión; una probabilidad igual a 1 indica que siempre que ha aparecido en el corpus ha sido anotado como participante en una opinión. Valores inferiores indican cierta ambigüedad en cuanto a si el término expresa opinión o no. Pueden consultarse los detalles sobre el método concreto utilizado para obtener los lexicones en (Cruz et al., 2013).[1] La tarea consiste en ampliar

[1] Los lexicones concretos utilizados en este trabajo no se corresponden con los presentados en dicho

dicho lexicón a partir de términos de textos del dominio sin anotar. A partir de las palabras que aparecen en dichos textos, debemos seleccionar cuáles pasarán a formar parte del lexicón y decidir sus polaridades.

3.1 Método

Nuestro método consiste en el entrenamiento de clasificadores ternarios que deciden para cada palabra candidata si se trata de una palabra neutra (no debe ser incluida en el lexicón), positiva o negativa. Estos clasificadores toman como entrada representaciones continuas de las palabras. A continuación detallamos los pasos del método.

En la fase de entrenamiento, seleccionamos aquellas palabras positivas y negativas del lexicón cuya probabilidad sea igual a 1. Como representantes de la categoría neutra, se seleccionan aleatoriamente un número similar de palabras de entre las contenidas en los textos sin anotar. Estas palabras deben ser revisadas para asegurarse de que no se incluyen palabras que pudieran expresar opinión. Para obtener la representación continua de las palabras seleccionadas hacemos uso de la herramienta *word2vec* (Mikolov et al., 2013a). La elección del modelo que utilice dicha herramienta, que puede haber sido construido a partir de textos del dominio o usando un corpus genérico, puede tener consecuencias en los resultados obtenidos. Ambas opciones serán estudiadas en la sección 4. Una vez obtenidos los datos de entrenamiento, entrenamos un clasificador multiclase de tipo *Support Vector Machines (SVM).*

Tras la fase de entrenamiento, y para cada una de las palabras contenidas en los textos no anotados, obtenemos representaciones continuas de las mismas usando *word2vec* con el mismo modelo utilizado en la fase de entrenamiento. Los vectores obtenidos son pasados por el clasificador, que decidirá si se trata de una palabra neutra, positiva o negativa. Si la palabra es clasificada como neutra, logicamente es descartada. Pero incluso si la palabra es clasificada como positiva o negativa, puede no ser seleccionada para ser incluida en el lexicón ampliado, si la probabilidad de pertenencia a la clase devuelta por el clasificador no supera un umbral mínimo fijado manualmente. Pretendemos de esta forma aumentar la precisión del método, a costa de una menor cobertura. Llevamos a cabo un estudio de este valor umbral en la sección 4.4.

4 Experimentación

En esta sección mostramos los distintos experimentos llevados a cabo para evaluar distintas variaciones de nuestro método. Trataremos de decidir si es más apropiado el uso de modelos de representación de las palabras entrenados sobre textos del dominio o modelos genéricos. También buscaremos el valor de probabilidad umbral más apropiado, y obtendremos una estimación final de la precisión de nuestro método.

4.1 Recursos utilizados

Hemos realizado los experimentos para el idioma español, en tres dominios distintos: alojamientos (principalmente, opiniones sobre hoteles), gastronomía (principalmente, opiniones sobre restaurantes) y actividades culturales (opiniones sobre museos, monumentos, conciertos, teatro, etc.). Los documentos de los que partimos para cada uno de estos dominios fueron extraídos de TripAdvisor[2]. En la Tabla 1 se muestran el número de documentos y el de palabras de cada uno de estos corpus.

Dominio	Opiniones	Palabras
Alojamiento	279.531	38.194.969
Gastronomía	284.782	17.321.589
Cultura	91.263	4.271.481

Tabla 1: Datos de los corpus utilizados.

En la Tabla 2 se muestran el número de términos que conforman cada uno de los lexicones de partida. También se indica el número de términos cuya probabilidad de opinión es igual a 1, es decir, aquellos que serán utilizados para el entrenamiento de los clasificadores. El resto de términos de los lexicones, para los que conocemos la polaridad y que no son usados en el entrenamiento, serán utilizados para evaluar el método.

Dominio	Total	Entren.	Eval.
Alojamiento	1.066	483	447
Gastronomía	1.852	885	462
Cultura	2.310	900	470

Tabla 2: Número de términos en los lexicones iniciales.

Para obtener las palabras neutras necesarias para el entrenamiento y la evaluación,

trabajo, sino que fueron obtenidos en el contexto del proyecto AORESCU (P11-TIC-7684 MO).

[2] `www.tripadvisor.com`

se extrajeron 500 palabras aleatoriamente de entre los tres corpus de documentos. Una vez filtradas las palabras con posibles usos de opinión quedaron 448, de las cuales 200 palabras fueron usadas para entrenamiento y 248 para evaluación.

Finalmente, para la representación continua de las palabras, se construyó un modelo para cada dominio a partir de los textos disponibles. También se dispone del modelo *Spanish Billion Words Corpus and Embeddings* (Cardellino, 2016). Se trata de un modelo entrenado con textos de diversa procedencia, y que podemos considerar por tanto genérico[3], frente a los modelos específicos del dominio anteriores.

4.2 Modelos de representación continua de palabras

Antes de proceder al entrenamiento de los clasificadores tomando como entrada las representaciones continuas de las palabras candidatas, llevamos a cabo una representación bidimensional de dichas representaciones. En la Figura 1 se muestra una proyección bidimensional de los vectores obtenidos mediante la herramienta *word2vec* para las palabras positivas, negativas y neutras del dominio alojamiento. Se empleó el modelo entrenado a partir de los textos del dominio. Parece observarse que existe una correlación entre la posición espacial de dichos vectores y la polaridad semántica de las palabras representadas, lo que refuerza nuestra idea de partida.

4.3 Entrenamiento Clasificadores

Para la obtención de los clasificadores se utilizó la herramienta *scikit-learn* (Pedregosa et al., 2011). El tipo de clasificador empleado fue *C-Support Vector Classification*, entrenado con parámetro de penalización $C = 100$ y un criterio de parada de $0,01$. Se entrenaron por un lado clasificadores usando los modelos de representación de palabras específicos del dominio y por otro el modelo genérico.

En las tablas 3 y 4 se muestran los resultados obtenidos al evaluar los clasificadores. En general, se observan mejores resultados utilizando los modelos de representación de palabras entrenados con los textos del dominio, frente al modelo genérico, aunque esta diferencia es menos pronunciada en el dominio de cultura. Esto puede ser debido al menor tamaño del corpus de documentos de dicho dominio. En cualquier caso, nuestro objetivo prioritario es garantizar una alta precisión, por lo que continuamos considerando ambos modelos en la experimentación restante.

[3]Lo llamaremos modelo genérico en el resto del trabajo.

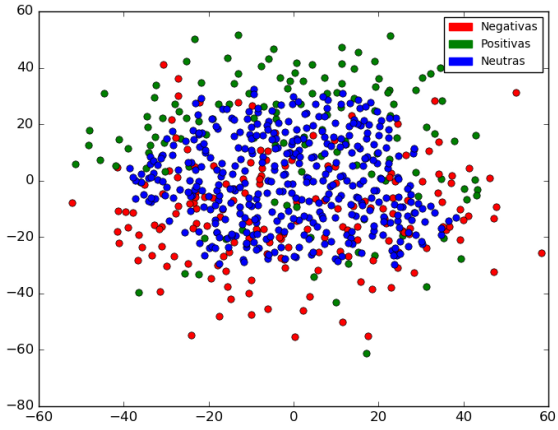


Figura 1: Proyección bidimensional de los vectores representantes de las palabras positivas, negativas y neutras para el dominio alojamiento.

Dominio	clase	p	r	F_1
Alojamiento	-1	0,794	0,761	0,777
	0	0,68	0,923	0,783
	1	0,955	0,675	0,791
Gastronomía	-1	0,738	0,842	0,786
	0	0,803	0,823	0,813
	1	0,874	0,779	0,824
Cultura	-1	0,677	0,625	0,650
	0	0,741	0,762	0,752
	1	0,785	0,801	0,793

Tabla 3: Evaluación de los clasificadores ternarios (modelos específicos del dominio).

4.4 Estimación del umbral

Dado que estamos interesados en primar la precisión frente a la cobertura, planteamos un

Dominio	clase	p	r	F_1
Alojamiento	-1	0,795	0,743	0,768
	0	0,644	0,919	0,568
	1	0,86	0,568	0,684
Gastronomía	-1	0,597	0,644	0,62
	0	0,641	0,762	0,696
	1	0,772	0,607	0,68
Cultura	-1	0,723	0,685	0,703
	0	0,636	0,790	0,705
	1	0,8	0,665	0,726

Tabla 4: Evaluación de los clasificadores ternarios (modelos genéricos).

método de selección de los candidatos más estricto: establecemos un valor umbral de probabilidad, de manera que sólo las palabras clasificadas como positivas o negativas con una probabilidad mayor o igual al umbral serán seleccionadas.

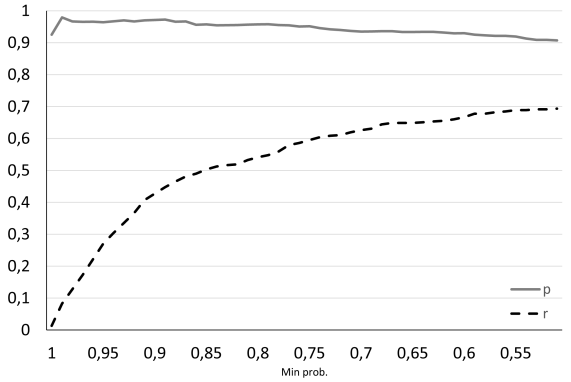


Figura 2: Precisión (p) y cobertura (r) del método para el dominio alojamiento en función del valor umbral de probabilidad.

En la Figura 2 se muestra cómo varían la precisión y la cobertura del clasificador basado en el modelo de palabras del dominio alojamiento, para todo el rango de valores del umbral mínimo de probabilidades. Nos fijamos únicamente en las clases positiva y negativa, correspondientes a las palabras seleccionadas por el método para ser incluidas en el lexicón. Seleccionaremos un valor de umbral mínimo buscando en esta gráfica el punto de la curva donde la cobertura está más próxima a 0.5 (valor que asegura que entorno a la mitad de las palabras de opinión presentes en el corpus serán seleccionadas para ser introducidas en el lexicón ampliado). Aplicando este procedimiento a los distintos clasificadores y dominios, se obtienen los valores de precisión y de umbral que se muestran en las tablas 5 y 6. En los dominios alojamiento y gastronomía los mejores valores de precisión se obtienen usando el modelo de representación de palabras del dominio, mientras que en el dominio cultura se obtiene usando el modelo genérico. Este comportamiento es coherente con lo observado al evaluar los clasificadores en la sección anterior, y puede ser achacado al menor tamaño del corpus del dominio cultura.

4.5 Método combinado

De los resultados anteriores se puede concluir que la decisión de usar un modelo de representación de palabras basado en el dominio o uno genérico puede influir sensiblemente a los resultados obtenidos. En lugar de establecer un mecanismo para tomar dicha decisión, proponemos un método de combinación de ambas opciones, basándonos de nuevo en la idea de primar la precisión del proceso: únicamente incluiremos en los lexicones ampliados aquellas palabras que sean igualmente clasificadas como positivas o negativas por ambos clasificadores (el basado en el modelo de representación de palabras del dominio y el basado en el modelo genérico), y siempre que la probabilidad de pertenencia a la clase en ambos casos sea mayor al umbral establecido. Aplicando este método y repitiendo el proceso de selección del umbral mínimo, obtenemos los resultados mostrados en la Tabla 7.

Dominio	$prob_{min}$	p
Alojamiento	0,841	0,957
Gastronomía	0,842	0,936
Cultura	0,67	0,836

Tabla 5: Umbrales de probabilidad y precisión media de las clases positiva y negativa (modelos específicos del dominio).

Dominio	$prob_{min}$	p
Alojamiento	0,662	0,909
Gastronomía	0,622	0,835
Cultura	0,671	0,911

Tabla 6: Umbrales de probabilidad y precisión media de las clases positiva y negativa (modelos genéricos).

Dominio	$prob_{min}$	p
Alojamiento	0,33	0,959
Gastronomía	0,53	0,933
Cultura	0,49	0,935

Tabla 7: Umbrales de probabilidad y precisión media de las clases positiva y negativa (método combinado).

En los dominios alojamiento y gastronomía se obtienen ahora precisiones similares a los mejores casos anteriores, sin necesidad de decidirnos por el uso del modelo de representación de palabras específico del dominio o por el modelo genérico. En el dominio cultura obtenemos además una mejora de los resultados anteriores de dos puntos porcentuales. Concluimos por tanto que el método combinado es el más apropiado para la tarea abordada.

5 Conclusiones

En este trabajo hemos expuesto un método de ampliación de lexicones de opinión, que a partir de un conjunto de documentos sin anotar extrae nuevos términos a incluir en el lexicón, y estima la polaridad de dichos términos. El método está diseñado para asegurar una alta precisión, de manera que podamos estar seguros de que la inmensa mayoría de los términos añadidos y sus polaridades son correctos. Los resultados obtenidos indican que nuestro método es capaz de ampliar los lexicones iniciales capturando la mitad del total de las palabras de opinión contenidas en los textos de partida, y con una muy alta precisión: en torno al 94 % de las palabras seleccionadas son efectivamente palabras de opinión y su polaridad, positiva o negativa, está correctamente asignada. Queda pendiente en nuestro trabajo la evaluación extrínseca del método, mediante la utilización de los lexicones ampliados para la resolución de tareas específicas (por ejemplo, la clasificación de la polaridad de documentos de opinión) para medir qué mejoras se obtienen frente al uso de los lexicones iniciales. Tampoco hemos abordado la selección de multipalabras, lo cual precisaría de un preprocesado de los textos de partida que detectara las colocaciones más frecuentes en el dominio. Una vez representadas como un único token, el método propuesto en este trabajo podría ser aplicado sin variaciones.

Agradecimientos

Este trabajo ha sido financiado a través del proyecto de investigación AORESCU (P11-TIC-7684 MO).

Bibliografía

Baccianella, Stefano, Andrea Esuli, y Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *LREC*, Valletta, Malta. European Language Resources Association (ELRA).

Cardellino, Cristian. 2016. Spanish Billion Words Corpus and Embeddings, March.

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, y G. Gandini, 2007. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics.* Franco Angeli Editore, Milano, IT.

Cruz, Fermín L, José A Troyano, Fernando Enríquez, F Javier Ortega, y Carlos G Vallejo. 2013. Long autonomy or long delay? the importance of domain in opinion mining. *Expert Systems with Applications*, 40(8):3174–3184.

Cruz, Fermín L, José A Troyano, F Javier Ortega, y Fernando Enríquez. 2011. Automatic expansion of feature-level opinion lexicons. En *Proceedings of WASSA*, páginas 125–131. Association for Computational Linguistics.

Cruz, Fermín L, Carlos G Vallejo, Fernando Enrı, José A Troyano, y others. 2012. Polarityrank: Finding an equilibrium between followers and contraries in a network. *Information Processing & Management*, 48(2):271–282.

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.

Dumais, Susan T. 1995. Latent semantic indexing (lsi): Trec-3 report. *Nist Special Publication SP*, páginas 219–219.

Dumais, Susan T. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Esuli, Andrea y Fabrizio Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. En *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Hatzivassiloglou, Vasileios y Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. En *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, páginas 174–181, Morristown, NJ, USA. Association for Computational Linguistics.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, páginas 168–177.

Huang, Eric H, Richard Socher, Christopher D Manning, y Andrew Y Ng. 2012.

Improving word representations via global context and multiple word prototypes. En *Proceedings of the ACL: Long Papers-Volume 1*, páginas 873–882. Association for Computational Linguistics.

Kamps, Jaap, Maarten Marx, Robert J. Mokken, y Maarten De Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. En *National Institute for*, volumen 26, páginas 1115–1118.

Kanayama, Hiroshi y Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. En *EMNLP*, páginas 355–363, Sydney, Australia, July. Association for Computational Linguistics.

Kim, Joo-Kyung y Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. En *EMNLP*, páginas 1625–1630.

Kim, Soo-Min y Eduard Hovy. 2004. Determining the sentiment of opinions. En *COLING*.

Mikolov, Tomas, Kai Chen, Greg Corrado, y Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, y Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

Molina-González, M Dolores, Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, y L Alfonso Ureña-López. 2015. A spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

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.

Pablos, Aitor García, Montse Cuadros, y German Rigau. 2015. Unsupervised word polarity tagging by exploiting continuous word representations. *Procesamiento del Lenguaje Natural*, 55:127–134.

Pavlopoulos, John y Ion Androutsopoulos. 2014. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. *Proceedings of LASMEACL*, páginas 44–52.

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

Qiu, Guang, Bing Liu, Jiajun Bu, y Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1).

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, y Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. En *EMNLP*, volumen 1631, página 1642. Citeseer.

Stone, Philip J. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, y Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. En *ACL (1)*, páginas 1555–1565.

Turian, Joseph, Lev Ratinov, y Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. En *Proceedings of the ACL*, páginas 384–394. Association for Computational Linguistics.

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.

Yu, Hong y Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. En *Proceedings of EMNLP*.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 57-64 recibido 29-03-2016 revisado 18-04-2016 aceptado 03-05-2016

Semantics-Driven Collocation Discovery

Descubrimiento de Colocaciones Utilizando Semántica

Sara Rodríguez-Fernández[1], Luis Espinosa-Anke[1], Roberto Carlini[1], Leo Wanner[1,2]
[1]NLP Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018 Barcelona (Spain)
[2]Catalan Institute for Research and Advanced Studies (ICREA)
sara.rodriguez.fernandez|luis.espinosa|roberto.carlini|leo.wanner@upf.edu

Abstract: *Collocations* are combinations of two lexically dependent elements, of which one (the *base*) is freely chosen because of its meaning, and the choice of the other (the *collocate*) depends on the base. Collocations are difficult to master by language learners. This difficulty becomes evident in that even when learners know the meaning they want to express, they often struggle to choose the right collocate. Collocation dictionaries, in which collocates are grouped into semantic categories, are useful tools. However, they are scarce since they are the result of cost-intensive manual elaboration. In this paper, we present for Spanish an algorithm that automatically retrieves for a given base and a given semantic category the corresponding collocates.
Keywords: collocations, collocation recognition, collocation semantic classification, second language learning, word embeddings, transformation matrix

Resumen: Las *colocaciones*, entendidas como combinaciones de dos elementos entre los cuales existe una dependencia léxica, es decir, donde uno de los elementos (la *base*) se escoge libremente por su significado, pero el otro (*colocativo*) depende de la base, suelen ser difíciles de utilizar por los hablantes no nativos de una lengua. Esta dificultad se hace visible en que estos, a menudo, aún sabiendo el significado que quieren expresar, tienen problemas a la hora de elegir el colocativo adecuado. Los diccionarios de colocaciones, donde los colocativos son agrupados en categorías semánticas son una herramienta muy útil, pero son recursos escasos y de costosa elaboración. En este artículo se presenta, para el español, un algoritmo que proporciona, dada una base y una categoría semántica, colocativos pertinentes a dicha categoría.
Palabras clave: colocaciones, reconocimiento de colocaciones, clasificación semántica de colocaciones, aprendizaje de lenguas, word embeddings, matriz de transformación

1 Introduction

Collocations such as *do* [*a*] *favour*, *take advice*, *take* [*a*] *picture*, *deep breath*, *close examination*, etc., are idiosyncratic co-occurrences of two lexical items with a direct syntactic dependency between them. One of the items (the *base*) is freely chosen by the speaker, while the selection of the other item (the *collocate*) is restricted by the base (Hausmann, 1984; Cowie, 1994; Mel'čuk, 1995). For instance, in *do* [*a*] *favour*, the choice of *favour* is free, while the choice of *do* is restricted; cf., e.g., **make* [*a*] *favour*, **take* [*a*] *favour*. The idiosyncratic nature of collocations makes them language-specific. Thus, while in English and French a picture is 'taken' (*take* [*a*] *picture*, *prendre* [*une*] *photo*), in Spanish it is 'made' (*hacer* [*una*] *foto*); in English you *spend time*, but you 'pass' it in Spanish and French (*pasar tiempo*, *passer du temps*).

Collocations are a key element in foreign language learning. They are difficult to master even for advanced students due to their idiosyncrasy (Hausmann, 1984; Bahns and Eldaw, 1993; Granger, 1998; Lewis and Conzett, 2000; Wible et al., 2003; Nesselhauf, 2005; Alonso Ramos et al., 2010). Wible et al. (2003) show that collocation errors are the most frequent of all errors in students' writings. Even when learners know the meaning they want to express, they often fail to do it by means of collocations, which, as a rule, means that they fail to select the col-

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

locate that expresses the intended meaning. For this reason, collocation resources that group collocations semantically can significantly contribute to second language learning. A few dictionaries of this kind already exist, see, among others, the Oxford Collocations Dictionary, MacMillan Collocations Dictionary, BBI (Benson, Benson, and Ilson, 2010), *Lexique actif du français* (LAF) (Mel'čuk and Polguère, 2007), and *Diccionario de Colocaciones del Español* (DiCE, http://dicesp.com). Some of them use explicit semantic glosses; cf., e.g., the MacMillan Collocations Dictionary for English, the LAF for French, or the DiCE for Spanish. However, since they are hand-crafted resources, the cost of their compilation is high, which explains why collocation dictionaries are usually of a limited coverage and are available only for a few languages.[1].

In this paper, we describe a *word embeddings*-based approach (Mikolov, Yih, and Zweig, 2013; Levy, Goldberg, and Ramat-Gan, 2014) to automatic compilation of semantically motivated collocation resources for Spanish. We build on the intuition that there is a linear relation between semantically similar words in embedding spaces. We exploit this linear relation to train a *function* (or *transition matrix*) that learns a semantic relation between bases and collocates (i.e., types or glosses of collocations).

The remainder of the paper is structured as follows. Section 2 presents a brief review of related work. In Section 3, the semantic glosses that are used to typify the collocations are presented. Section 4 describes the methodology for the acquisition of the resources, while Section 5 presents the performed experiments and the evaluation of their outcome. In Section 6, we discuss then the performance of the implementation of our approach. Finally, Section 7 draws some conclusions and outlines possible future work in the context of semantically-motivated automatic collocation classification.

2 Related work

In the last decades, a large body of work on automatic retrieval of collocations has been produced. Some approaches exploit statistical evidence to measure word distribution in corpora, both in isolation and in combination with other words (Choueka, 1988; Church and Hanks, 1989; Evert, 2007; Pecina, 2008). Other works combine statistical measures with syntactic information, under the assumption that only those co-occurring words that form a syntactic structure can also form a collocation (Smadja, 1993). More recently, contexts in which a pair of words co-occurs have been taken into account Bouma (2009).

With regard to the semantic classification of collocations, there seems to be a common trend to use supervised machine learning techniques for the classification of collocations against their corresponding target semantic categories, leveraging as training data lists of collocations (Wanner, Bohnet, and Giereth, 2006; Gelbukh and Kolesnikova, 2012; Moreno, Ferraro, and Wanner, 2013; Wanner, Ferraro, and Moreno, 2016).

In our previous work (Rodríguez-Fernández et al., 2016), we developed an approach for automatic extraction of collocations, which accounted for the underlying semantics of each word by means of their distributional representation. This allowed us to perform a joint process of extraction and semantic typification of collocations. The approach is based on the representation of individual words as word vectors and takes in an unsupervised setting advantage of semantic properties of word embeddings (Mikolov, Yih, and Zweig, 2013), which may be defined in terms of the well-known vector operations such as summation and subtraction. Specifically, given a particular meaning and a base, the algorithm retrieves collocates that have in combination with the given base this particular meaning. The discovery of new collocates is thus done by means of an analogy, e.g., it would attempt to discover $x = vec(\text{deafening})$ in the analogy $vec(\text{strong}) - vec(\text{wind}) + vec(\text{noise}) = x$.

As already, in (Rodríguez-Fernández et al., 2016), in our current work, we retrieve and classify collocations in semantic terms simultaneously. However, while in (Rodríguez-Fernández et al., 2016) this was done using an unsupervised learning model, here we draw upon a supervised model.

[1]To the best of our knowledge, collocation dictionaries of a reasonable coverage are only available for English

3 Semantic collocation typology

It is common that different bases prompt for different collocates to express a given meaning. For instance, to express that a *disease* is 'intense', the collocates *serious* or *dangerous* can be used. To express that a person 'is affected by' a *disease*, *suffer* or *have* is used. If someone 'starts having' a disease, *catch, get* or *contract* are preferable, while when there is a person 'causing' a disease in someone else, *give, transmit* or *pass on* will be used, and so on. In Spanish, an 'intense' *disease* (*enfermedad*) is *grave* 'grave'. *Sufrir* 'suffer', *padecer* 'endure' or *tener* 'have' can be used to express that a person 'is affected by' a *disease*. *Contraer* 'contract' is preferred for 'start having', while for 'cause' *contagiar* 'pass on' or *transmitir* 'transmit' should be used instead.

As already mentioned above, collocation dictionaries, such as the Oxford Collocations Dictionary or the MacMillan Collocations Dictionary for English, or the *Diccionario de Colocaciones del Español* (DiCE) for Spanish classify collocations into semantic categories such that language learners can find more easily the collocate that communicate the meaning they intend to express. Categories of different granularity are used in each case. Similarly, different works on automatic classification of collocations use as target classes categories of different granularity. For instance, Wanner, Ferraro, and Moreno (2016) use 16 categories to classify verb+noun collocations and 5 for adj+noun collocations; Moreno, Ferraro, and Wanner (2013) and Chung-Chi et al. (2009) classify collocations into broader categories; Wanner, Bohnet, and Giereth (2006), Gelbukh and Kolesnikova (2012) and also Moreno, Ferraro, and Wanner (2013) in their second run of experiments use the semantic typology of *Lexical Functions* (LFs) (Mel'čuk, 1996), also used in DiCE.

In our experiments, we use a subset of ten LFs . For all of these LFs, we define semantic glosses similar to those used in the MacMillan Collocations Dictionary, in order to make the LFs more transparent to users. Some examples are 'intense', 'perform', 'increase' and 'show'; cf. Table 1 for the list of glosses and examples that illustrate them.

4 Methodology

As argued by Mel'čuk (1996), the meaning of collocates across collocations can be captured in a generic semantic (lexical function, LF) typology. For convenience, Mel'čuk defines for each LF a Latin acronym (such as 'Oper', 'Func', 'Magn', etc.), but, in general, for each LF also a semantic gloss is available. For instance, *absolute*, *deep*, *strong*, *heavy* in *absolute certainty*, *deep thought*, *strong wind*, and *heavy storm* can all be glossed as 'intense'; *make*, *take*, *give*, *carry out* in *make* [*a*] *proposal*, *take* [*a*] *step*, *give* [*a*] *hint*, *carry out* [*an*] *operation* can be glossed as 'do'/'perform'; etc. Similarly, in Spanish, *ensordecedor* 'deafening' in *ruido ensordecedor* 'deafening noise', *alta* 'high' en *alta estima* 'high esteem' or *fuerte* 'strong' in *fuerte golpe* 'strong blow', can be glossed as 'intense', and so on. Our goal is to capture the relation that holds between the training bases and collocates that share the same gloss, such that, given a new base and a gloss, we can retrieve the corresponding collocate(s) of this new base pertinent to this gloss. Thus, given *absolute certainty*, *deep thought*, and *strong wind* as training examples, *storm* as input base and 'intense' as gloss, we aim at retrieving the collocate *heavy*. Our approach is based on Mikolov, Le, and Sutskever (2013)'s translation matrix, where word vector representations between two analogous spaces are found to be linearly related. In Mikolov et al.'s original work, which describes the potential of this property for Machine Translation, one space captures words in language L_1 and the other space words in language L_2, such that the found relations are between translation equivalents. In our case, we define a base space $\mathcal{B}$ and a collocate space $\mathcal{C}$ in order to relate bases with their collocates that have the same meaning, and in the same language. To obtain the word vector representations in $\mathcal{B}$ and $\mathcal{C}$, we use Mikolov, Yih, and Zweig (2013)'s *word2vec*.[2]

Let $\mathbf{T}$ be a set of collocations whose collocates share the semantic gloss τ, and let b_{t_i} and c_{t_i} be the corresponding base and collocate of a collocation $t_i \in \mathbf{T}$. Then, we may denote a *base* matrix as $B_\tau = [b_{t_1}, b_{t_2} \dots b_{t_n}]$, and a *collocate* matrix as $C_\tau = [c_{t_1}, c_{t_2} \dots c_{t_n}]$, given by the corresponding vector representations of each collo-

[2] https://code.google.com/archive/p/word2vec/

Semantic gloss	Example	# instances
'intense'	*sumo cuidado* 'extreme care'	174
'weak'	*cantidad insignificante* 'negligible amount'	23
'perform'	*dar* [*un*] *abrazo* 'to give [a] hug'	319
'begin to perform'	*tomar posesión* 'to take possession'	67
'stop performing'	*renunciar* [*a un*] *papel* 'to abandon [a] role'	3
'increase'	*fortalecer* [*el*] *control* 'to strengthen control'	22
'decrease'	*bajar* [*un*] *impuesto* 'to lower [a] tax'	16
'create', 'cause'	*escribir* [*una*] *carta* 'to write [a] letter'	181
'put an end'	*apagar* [*un*] *fuego* 'to extinguish [a] fire'	31
'show'	*expresar disconformidad* 'to express disagreement'	5

Table 1: Semantic glosses and size of training set

cation component. Together, they constitute a set of training examples Φ_τ composed by vector pairs $\{b_{t_i}, c_{t_i}\}_{i=1}^{n}$.

We learn a linear transformation matrix from Φ_τ, denoted as $\Psi_\tau \in \mathbb{R}^{\mathcal{B}\times\mathcal{C}}$. Specifically, and following the notation in (Tan et al., 2015), this transformation may be depicted as:

$$B_\tau \Psi_\tau = C_\tau$$

We follow Mikolov et al.'s original approach and compute Ψ_τ as follows:

$$\min_{\Psi_\tau} \sum_{i=1}^{|\Phi_\tau|} \|\Psi_\tau b_{t_i} - c_{t_i}\|^2$$

At the end of this procedure, each time our algorithm observes a novel base b_{j_τ}, a novel list of ranked collocate candidates is retrieved by applying $\Psi_\tau b_{j_\tau}$. The obtained list of candidates is filtered in terms of part of speech (only plausible PoS patterns are admitted as candidates) and in terms of the $NPMI$ metric. The $NPMI$ metric is an association measure based on pointwise mutual information that factors in the semantic asymmetry between the base and the collocate (Carlini, Codina-Filba, and Wanner, 2014):

$$NPMI = \frac{PMI(collocate, base)}{-log(p(collocate))}$$

Such a combination of heterogeneous models has been used before and proved to be effective to discover other types of relationships between word pairs (Zhila et al., 2013).

5 Experiments

In what follows, we first describe the setup of our experiments and then present their output.

5.1 Experimental setup

We carried out our experiments on the ten semantic collocate glosses listed in the first column of Table 1: eight verbal collocate glosses in verb+noun collocations and two property glosses in adj+noun collocations, first without filtering the obtained candidate list and then applying the PoS and $NPMI$ filters. The training examples for each of the glosses in our experiments were taken from a three thousand sentence corpus in which collocations were manually annotated and classified with respect to LFs.[3] Duplicates were removed. However it was common to find more than one collocate for each base. Ten instances for each gloss were set apart for testing. Since the distribution of collocations with different glosses is not homogeneous (e.g., collocations conveying the idea of 'intense' are used more often than those conveying the idea of 'weak', and those meaning 'perform' are more used than those meaning 'stop performing'), in our data, the number of instances per gloss also varies significantly (see Table 1 for the number of training instances for each gloss).

Both bases and collocates were modeled by training their word vectors over a 2014 dump of the Spanish Wikipedia. For the calculation of $NPMI$ during the postprocessing stage, a seven million sentence newspaper corpus was used.

5.2 Evaluation

To assess the outcome of the experiments, the correctness of each candidate from the top-10 that were retrieved for each test base was verified. Given that a base can have different collocates to express a meaning, the evaluation was not performed automatically against

[3] Recall that our glosses correspond to lexical functions.

the collocates found in the corpus; instead, each candidate was manually judged as correct or incorrect. For the outcome of each experiment, we computed both *Precision* (P) as the ratio of collocates with the targeted gloss retrieved for each base, and *Mean Reciprocal Rank* (MRR), which rewards the position of the first correct result in a ranked list of outcomes:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where Q is a sample of experiment runs and $rank_i$ refers to the rank position of the *first* relevant outcome for the ith run. MRR is commonly used in Information Retrieval and Question Answering, but has also shown to be well suited for collocation discovery; see, e.g., (Wu et al., 2010).

We compared the performance of our setup to the accuracy achieved in our previous work (Rodríguez-Fernández et al., 2016) (see also Section 2 above), which serves us as baseline, also in two variants: without and with PoS+$NPMI$ filters. The results of our experiments are shown in Table 2 ('S1' stands for the baseline configuration in which all top-10 retrieved candidates are kept; 'S2' for the baseline configuration with PoS+$NPMI$ filtering; 'S3' for our current transformation matrix-based setup without filtering; and 'S4' for the matrix-based setup with PoS+$NPMI$ filtering.

6 Discussion

As we can observe from the number of instances in Table 1, certain glosses seem to possess less linguistic variability, requiring a lower number of instances for building the transformation matrix from bases to collocates. For example, the transformation function of 'stop performing', trained with only 3 instances, achieves the second best results both for P and MRR.

Comparing the unfiltered configurations of both the baseline and our approach to their filtered counterparts, an evident increase of precision can be seen. This means that the incorporation of a filtering module, especially the $NPMI$, improves the performance of the algorithms substantially. For example, *suscitar* 'raise', as candidate for the base *infección* 'infection', was discarded by the $NPMI$, while *provocar* 'provoke' and *producir* 'produce', with $NPMI$ 0.30 and 0.31, were kept. Similarly, for the base *velocidad* 'speed', *amplio* 'wide' was discarded, while *máxima* 'maximum', *gran* 'great', *vertiginosa* 'vertiginous' and *alta* 'high', with $NPMI$ 0.46, 0.51, 0.51 and 0.77, were kept.

After close examination of the candidates, we found that a great number of the candidates retrieved and filtered by our system were actually correct collocates. However, their meaning was somewhat different to that of the semantic gloss. A very common source of error are antonym words, since our approach is based on word embeddings, i.e. vector representations of words based on their contexts. Antonyms often share the same linguistic context and are therefore considered as similar words by the model. Consider the following examples as illustration:

(1) *voz tenue*, 'faint voice' (belongs to 'weak' instead of 'intense')

(2) *fuerte tensión*, 'strong tension' (belongs to 'intense' instead of 'weak')

(3) *aumentar* [*una*] *tasa*, 'to increase [a] rate' (belongs to 'increase' instead of 'decrease')

(4) *derribar* [*un*] *templo*, 'to demolish [a] temple' (belongs to 'put an end' instead of 'create', 'cause')

(5) *plantear* [*una*] *duda*, 'to raise [a] question' (belongs to 'create', 'cause', instead of 'put an end')

However, the fact that we are able to obtain such 'intense' collocations as *velocidad vertiginosa* 'vertiginous speed', such 'put an end' collocations as *resolver* [*una*] *duda* 'solve [a] doubt', or such 'increase' collocations as *encarecer* [*un*] *precio* 'to increase [a] price' shows the potential of our approach.

A look at Table 2 may furthermore give the impression that the overall numbers are still rather low (e.g., for 'begin to perform' we achieve only 0.15 of precision, for 'decrease' only 0.19, etc.). In this respect, it should be noted that in our evaluation, the retrieved collocate candidates were considered correct only if they were both correct collocates and belonged to the target semantic gloss. In other words, for a candidate to be correct it was required not only to *collocate* with the base, but also to belong to the target semantic category. However, it is well-known that it is by far not always clear whether a given co-occurrence forms a collocation or a free word combination. If we relax our evaluation in the

	Precision				Mean Reciprocal Rank			
Semantic gloss	S1	S2	S3	S4	S1	S2	S3	S4
'intense'	0.25	0.12	0.17	**0.44**	0.52	0.10	0.31	**0.42**
'weak'	0.0	0.0	0.10	**0.45**	0.00	0.00	**0.75**	0.60
'perform'	0.09	0.00	**0.20**	0.16	0.19	0.00	**0.44**	0.25
'begin to perform'	**0.15**	0.00	0.03	**0.15**	**0.29**	0.00	0.04	0.08
'stop performing'	0.04	0.04	0.06	**0.44**	0.1	0.07	0.35	**0.53**
'increase'	0.20	0.08	0.25	**0.50**	0.51	0.17	0.63	**0.67**
'decrease'	0.04	0.09	0.08	**0.19**	0.07	0.10	0.35	**0.43**
'create', 'cause'	0.07	0.13	**0.21**	0.20	0.38	0.13	**0.57**	0.38
'put an end'	0.06	0.05	0.16	**0.23**	0.26	0.02	0.32	**0.43**
'show'	0.02	0.00	0.31	**0.33**	0.20	0.00	**0.85**	0.55

Table 2: Precision and MRR for the baselines (S1 and S2) and the two configurations of our approach (S3 and S4)

Semantic gloss	**Base**	**Retrieved candidates**
'intense'	*velocidad*	*alto, máximo, constante, gran, considerable, vertiginoso*
	'speed'	'high, maximum, constant, great, considerable, vertiginous'
'weak'	*plazo*	*breve, corto, largo, prorrogable*
	'period'	'brief, short, long, extendable'
'perform'	*viaje*	*hacer, embarcar, efectuar, realizar, iniciar, preparar, topar*
	'trip'	'make, load, carry out, make, initiate, prepare, bump into'
'begin to perform'	*éxito*	*alcanzar, medir, suponer, rebasar, propiciar, presumir, presagiar*
	'success'	'attain, measure, suppose, overflow, propiciate, boast, foretell'
'stop performing'	*escondite*	*abandonar*
	'hiding place'	'abandon'
'increase'	*producción*	*incentivar, fomentar, promover, alentar, potenciar, fortalecer*
	'production'	'incentive, foster, promote, encourage, improve, strenghten'
'decrease'	*pérdida*	*reducir, moderar, frenar, compensar, disminuir, elevar*
	'loss'	'reduce, moderate, brake, compensate, decrease, increase'
'create', 'cause'	*templo*	*construir, erigir, levantar, edificar, derribar*
	'temple'	'build, erect, raise, build, demolish'
'put an end'	*duda*	*resolver, solventar, plantear, zanjar*
	'doubt'	'solve, resolve, set out, settle'
'show'	*opinión*	*expresar, manifestar, reflejar, resumir, plasmar, exponer*
	'opinion'	'express, manifest, reflect, summarize, express, expound'

Table 3: Examples of retrieved collocations

sense that a candidate is judged to be correct if it belongs to the target semantic category, no matter whether it is considered to form with the base a collocation in the strict sense or not,[4] the precision is likely to increase. For instance, for English, we observe that for 'intense', 'put an end' and 'show', it increases 0.1, 0.18 and 0.15 points, respectively. For other glosses, the increase is minor, as, e.g., in the case of 'begin to perform' or 'stop performing', for which the increase is only 0.04 and 0.02.

7 *Conclusions and future work*

We have presented an approach to automatic compilation of semantically-motivated collocation resources. Our technique is grounded in Mikolov, Yih, and Zweig (2013)'s word embeddings and the assumption that semantically related words in two different vector representations are related by linear transformation (Mikolov, Le, and Sutskever, 2013). This property has also been exploited for other tasks, such as word-based translation (Mikolov, Le, and Sutskever, 2013), learning

[4]Thus, combinations such as *gran* in *gran tamaño*, *hacer* in *hacer* [*un*] *movimiento* or *bajo* in *salario bajo* would be considered correct collocates of the glosses 'intense', 'perform' and 'weak', respectively, while *amplia* in *amplia velocidad*, *hacer* in *hacer* [*una*] *decisión*, or *suscitar* in *suscitar* [*una*] *infección* would be rejected as collocates of the glosses 'intense', 'perform' and 'cause', respectively.

semantic hierarchies (hyponym-hypernym relations) in Chinese (Fu et al., 2014), or modeling linguistic similarities between standard and non-standard language (Tan et al., 2015). For our task of collocation discovery, we learn a series of *transition matrices* (one for each target semantic gloss) over a handful of collocation examples, where collocates share the same gloss, and then apply these matrices to discover, for any previously unseen base, new collocates that belong to the same semantic type. In the paper, we discussed the outcome of the experiments with ten different glosses such as 'do / perform', 'increase' or 'intense', and show that for most glosses, an approach that combines a stage of the application of a gloss-specific *transition matrix* and a pruning stage based on statistical evidence outperforms baselines which exploit only one of these stages or a baseline that is based on the embeddings property for drawing analogies (Rodríguez-Fernández et al., 2016).

Here, we focused on Spanish and only on a small amount of collocations. However, since our approach requires only big unannotated corpora, it is highly scalable and portable to other languages. Given the lack of semantically tagged collocation resources for most languages, our work has the potential to become influential in the context of second language learning.

In the future, we plan to investigate whether increasing the number of training instances, and using word embeddings trained on a corpus richer in collocations may affect the performance of the system. We also plan to extend our work by increasing the number of semantic glosses, thus generating more complete resources.

8 Acknowledgements

The present work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO), through a predoctoral grant (BES-2012-057036) in the framework of the project HARenES (FFI2011-30219-C02-02) and the Maria de Maeztu Excellence Program (MDM-2015-0502).

References

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, and S. Prieto. 2010. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214.

Bahns, J. and M. Eldaw. 1993. Should we teach EFL students collocations? *System*, 21(1):101–114.

Benson, M., E. Benson, and R. Ilson. 2010. *The BBI Combinatory Dictionary of English: Your guide to collocations and grammar, Third Edition*. Benjamins Academic Publishers, Amsterdam.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, R. Eckart de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference*. Gunter Narr Verlag, Tübingen, pages 31–40.

Carlini, R., J. Codina-Filba, and L. Wanner. 2014. Improving Collocation Correction by ranking suggestions using linguistic knowledge. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, Uppsala, Sweden.

Choueka, Y. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pages 34–38.

Chung-Chi, H., K. H. Kao, C. H. Tseng, and J. S. Chang. 2009. A thesaurus-based semantic classification of English collocations. *Computational Linguistics and Chinese Language Processing*, 14(3):257–280.

Church, K. and P. Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pages 76–83.

Cowie, A. 1994. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*. Pergamon, Oxford, pages 3168–3171.

Evert, S. 2007. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Fu, R., J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. 2014. Learning semantic hierarchies via word embeddings. In *ACL (1)*, pages 1199–1209.

Gelbukh, A. and O. Kolesnikova. 2012. *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer, Heidelberg.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*. Oxford University Press, Oxford, pages 145–160.

Hausmann, F.-J. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

Levy, O., Y. Goldberg, and I. Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

Lewis, M. and J. Conzett. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

Mel'čuk, I. 1995. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Hillsdale, pages 167–232.

Mel'čuk, I. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102.

Mel'čuk, I. and A Polguère. 2007. *Lexique actif du français*. de boeck, Brussels.

Mikolov, T., Q.V. Le, and I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., W.-T. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Moreno, P., G. Ferraro, and L. Wanner. 2013. Can we determine the semantics of collocations without using semantics? In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik, editors, *Proceedings of the eLex 2013 conference*.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam.

Pecina, P. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57.

Rodríguez-Fernández, S., R. Carlini, L. Espinosa-Anke, and L. Wanner. 2016. Example-based acquisition of fine-grained collocation resources. In *Proceedings of LREC*, Portoroz, Slovenia.

Smadja, F. 1993. Retrieving collocations from text: X-tract. *Computational Linguistics*, 19(1):143–177.

Tan, L., H. Zhang, C.LA. Clarke, and M.D. Smucker. 2015. Lexical comparison between wikipedia and twitter corpora by using word embeddings. *Volume 2: Short Papers*, page 657.

Wanner, L., B. Bohnet, and M. Giereth. 2006. Making sense of collocations. *Computer Speech and Language*, 20(4):609–624.

Wanner, L., G. Ferraro, and P. Moreno. 2016. Towards distributional semantics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography*, doi:10.1093/ijl/ecw002.

Wible, D., C.H. Kuo, N.L. Tsao, A. Liu, and H.L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1):90–102.

Wu, J.-C., Y.-C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.

Zhila, A., W.-T. Yih, C. Meek, G. Zweig, and T. Mikolov. 2013. Combining heterogeneous models for measuring relational similarity. In *HLT-NAACL*, pages 1000–1009.

Aprendizaje Automático en PLN

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 67-74 recibido 29-03-2016 revisado 21-04-2016 aceptado 02-05-2016

Una aproximación al uso de *word embeddings* en una tarea de similitud de textos en español

An approach to the use of word embeddings in a textual similarity task for Spanish texts

Tomás López-Solaz Universidad de Sevilla tlopez2@us.es
José A. Troyano Universidad de Sevilla troyano@us.es
F. Javier Ortega Universidad de Sevilla javierortega@us.es
Fernando Enríquez Universidad de Sevilla fenros@us.es

Resumen: En este trabajo mostramos cómo una representación vectorial de palabras basada en *word embeddings* puede ayudar a mejorar los resultados en una tarea de similitud semántica de textos. Para ello hemos experimentado con dos métodos que se apoyan en la representación vectorial de palabras para calcular el grado de similitud de dos textos, uno basado en la agregación de vectores y otro basado en el cálculo de alineamientos. El método de alineamiento se apoya en la similitud de vectores de palabras para determinar la vinculación entre las mismas. El método de agregación nos permite construir representaciones vectoriales de los textos a partir de los vectores individuales de palabras. Estas representaciones son comparadas mediante dos distancias clásicas como son la euclídea y la del coseno. Hemos evaluado nuestros sistemas con el corpus basado en Wikipedia distribuido en la competición de similitud de textos en español de SemEval-2015. Nuestros experimentos muestran que el método basado en alineamiento se comporta mucho mejor, obteniendo resultados muy cercanos al mejor sistema de SemEval. El método basado en agregación de vectores se comporta sensiblemente peor. No obstante, esta segunda aproximación parece capturar aspectos de similitud no recogidos por la primera, ya que cuando se combinan las salidas de ambos sistemas se mejoran los resultados del método de alineamiento, superando incluso los resultados del mejor sistema de SemEval.
Palabras clave: Similitud semántica, *word embedding*, alineamiento de textos

Abstract: In this paper we show how a vector representation of words based on word embeddings can help to improve the results in tasks focused on the semantic similarity of texts. Thus we have experimented with two methods that rely on the vector representation of words to calculate the degree of similarity of two texts, one based on the aggregation of vectors and the other one based on the calculation of alignments. The alignment method relies on the similarity of word vectors to determine the semantic link between them. The aggregation method allows us to construct vector representations of the texts from the individual vectors of each word. These representations are compared by means of two classic distance measures: Euclidean distance and cosine similarity. We have evaluated our systems with the corpus based on Wikipedia distributed in the competition of similarity of texts in Spanish of SemEval-2015. Our experiments show that the method based on the alignment of words performs much better, obtaining results that are very close to the best system at SemEval. The method based on vector representations of texts behaves substantially worse. However, this second approach seems to capture aspects of similarity not detected by the first one, as when the outputs of both systems are combined the results of the alignment method are surpassed, even exceeding the results of the best system at SemEval.
Keywords: Semantic similarity, *word embedding*, text alignment

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introducción

La medición fiable de la similitud de textos es una aplicación de gran ayuda para distintas tareas relacionadas con el procesamiento de textos en lenguaje natural. Sólo por citar algunos ejemplos, los sistemas de clasificación de documentos, de resúmenes de textos, o de traducción automática, pueden beneficiarse de un módulo de similitud que establezca el grado de parecido entre dos unidades textuales. En general, podemos distinguir entre dos tipos de similitud: a nivel de palabras y a nivel de textos.

Las distintas aproximaciones de similitud a nivel de palabras se agrupan en dos categorías: similitud léxica de palabras y similitud semántica de palabras. Dos palabras son similares a nivel léxico si están compuestas por secuencias parecidas de caracteres. Para determinar la similitud léxica de palabras se suelen utilizar distintas métricas basadas en comparación de cadenas de caracteres. La similitud semántica de palabras, por su parte, nos permite medir si dos palabras tienen significados parecidos o se usan en contextos parecidos. Hay dos grandes grupos de técnicas para calcular la similitud semántica de palabras: técnicas basadas en conocimiento y técnicas basadas en corpus.

Las técnicas de similitud de palabras basadas en conocimiento miden el grado de parecido entre palabras apoyándose en algún tipo de recurso lingüístico que proporcione información sobre el significado de las palabras. El recurso por excelencia es WordNet (Miller, 1995), una base de datos léxica organizada en torno a varias relaciones entre palabras. La relación de sinonimia es la más importante y en ella se apoya el concepto de *synset* (grupo de sinónimos) que permite definir de forma implícita el significado de las palabras a través del conjunto de *synsets* en los que aparece. En función de las relaciones entre palabras, se han definido varias métricas de similitud entre las que destacan las de Resnik (Resnik, 1995), Lin (Lin, 1998) y Jian & Conrath (Jiang y Conrath, 1997).

Las técnicas de similitud de palabras basadas en corpus determinan el parecido semántico de dos palabras en función de los usos de esas palabras en una gran colección de textos. Por lo general, estas técnicas se basan en algún tipo de representación vectorial de las palabras en función de los distintos contextos en los que dichas palabras aparecen. Dentro de las técnicas basadas en corpus destacan las de *latent semantic analysis* (LSA) y las de *word embedding*. LSA analiza las relaciones entre un conjunto de documentos y los términos que contienen, estableciendo que dos palabras son similares si ocurren en fragmentos similares de textos. LSA parte de una matriz de palabras frente a documentos y aplica una técnica matemática denominada *singular value decomposition* que permite reducir el número de filas (documentos) preservando la similitud entre columnas (palabras). Por su parte, las técnicas de *word embedding* parten de representaciones BOW (*bag of words*) de los distintos contextos de las palabras para obtener representaciones vectoriales de las palabras de dimensiones mucho más reducidas que capturan el significado y las relaciones entre palabras. Hay diversas técnicas para calcular estas representaciones, una de las más empleadas se basa en redes neuronales de una sola capa oculta que predicen la palabra dado el contexto o viceversa, adaptando así una de las piezas básicas de los modelos de aprendizaje profundo, los autocodificadores. Las técnicas de *word embedding* han demostrado ser muy útiles en múltiples tareas del Procesamiento del Lenguaje Natural aparte de la similitud de textos (Collobert et al., 2011), (Zou et al., 2013), y en la actualidad gozan de gran popularidad. Esto se debe en gran medida a la existencia de herramientas como Word2vec (Mikolov et al., 2013) o GloVe (Pennington, Socher, y Manning, 2014) que han facilitado mucho el acceso a este tipo de técnicas a la comunidad investigadora.

Las técnicas de similitud a nivel de palabras proporcionan una información básica para afrontar la tarea de similitud a nivel de textos. Muchas de las aproximaciones de similitud de textos se basan en la idea de alineamiento, que básicamente consiste en un emparejamiento entre palabras de los dos textos a comparar.

El alineamiento entre dos textos proporciona un marco sobre el que se pueden evaluar distintas heurísticas para determinar el grado de similitud entre los textos. Por ejemplo, usando las conexiones propuestas en el alineamiento para calcular métricas de similitud semántica entre las palabras emparejadas y agregando estas métricas individuales para obtener una métrica de similitud global entre los dos textos.

En este trabajo estamos interesados en analizar vías que permitan integrar el conocimiento obtenido mediante *word embeddings* a la hora de determinar el grado de similitud de dos textos. El uso de *word embeddings* no es nuevo en sistemas de similitud de textos, pero siempre como complemento a otras técnicas o recursos. Nuestra intención es evaluar el comportamiento de un sistema que se base exclusivamente en el conocimiento proporcionado por un modelo de *word embedding*.

Hemos identificado dos maneras en las que la representación de palabras en el espacio continuo puede ser de utilidad para esta tarea: con un método de alineamiento basado exclusivamente en *embeddings* y con una representación vectorial de textos basada en los vectores de palabras.

El método de alineamiento propuesto usa el grado de similitud de palabras como mecanismo para identificar posibles emparejamientos de palabras, construyéndose alineamientos bidireccionales en los que las palabras son emparejadas parcialmente en función de su similitud.

La segunda aproximación consiste en el uso de los vectores de palabras para construir representaciones vectoriales de los textos. Estas representaciones son comparadas mediante dos distancias clásicas como son la euclídea y la del coseno para obtener así un grado de similitud entre dos textos.

Hemos evaluado nuestros sistemas con el corpus basado en wikipedia distribuido en la competición de similitud de textos en español de SemEval-2015. De nuestras dos aproximaciones, la que mejor se comporta es la del método de alineamiento, para la que se obtienen resultados muy cercanos al mejor sistema de SemEval. La idea de usar los vectores de palabras para construir vectores para los textos se comporta sensiblemente peor. No obstante esta segunda aproximación parece capturar aspectos de similitud no recogidos por la primera, ya que cuando se combinan las salidas de ambos sistemas se mejoran los resultados del método de alineamiento, superando incluso los resultados del mejor sistema de SemEval.

El resto del artículo se organiza de la siguiente forma: la sección 2 describe brevemente algunos trabajos relacionados, la sección 3 describe tanto el método basado en representación vectorial de textos como el basado en alineamiento, así como el método de combinación aplicado, la sección 4 incluye los resultados experimentales y, por último, la sección 5 incluye las conclusiones y plantea algunas líneas de trabajo futuro.

2 Trabajos relacionados

La mayoría de las contribuciones recientes en el ámbito de la similitud semántica de textos provienen de los participantes en las tareas que se proponen anualmente en SemEval[1]. De estas participaciones han surgido muchas técnicas, ideas, e incluso frameworks que ofrecen componentes que pueden ser utilizados para desarrollar sistemas de similitud de textos, como por ejemplo DKPro (Bär, Zesch, y Gurevych, 2013). Aparte de los trabajos presentados en SemEval, otro referente clásico en este campo es el trabajo de (Mihalcea, Corley, y Strapparava, 2006) centrado en el cálculo de similitud semántica para textos cortos, y en el que se resumen las métricas de similitud más importantes que posteriormente se han venido utilizando como parte de la mayoría de los sistemas de similitud de textos.

Las distintas propuestas que han participado en las ediciones recientes de las tareas de similitud de SemEval son combinaciones de métricas clásicas con ideas originales. En muchos casos, los participantes van mejorando sus sistemas año a año para incorporar nuevas ideas o para adaptarse a los distintos cambios en la definición de las tareas por parte de los organizadores. A modo de muestra del tipo de técnicas usadas en estos sistemas, resumimos a continuación las características más significativas de los tres mejores sistemas en la tarea en español de la competición de SemEval de 2015 (Agirre et al., 2015), cuyo corpus hemos usado en nuestros experimentos.

El mejor grupo fue (Hänig, Remus, y Puente, 2015), su solución integra tres técnicas: representación vectorial de textos (BOW y distancia del coseno), alineamiento mediante métricas de similitud y uso de *machine learning* para la combinación de las distintas métricas calculadas. Usa un alineamiento secuencial y emplea Word2vec en una última fase para intentar emparejar palabras residuales que no han sido emparejadas por su técnica de alineamiento.

El segundo grupo de la competición (Ka-

[1] https://en.wikipedia.org/wiki/SemEval

rumuri, Vuggumudi, y Chitirala, 2015) se apoya en un sistema previo para inglés y utiliza el traductor de Google para adaptar las entradas al español a su sistema. Integra un sistema de alineamiento con distintas métricas basadas en proporcionalidad, número de sustantivos, número de adjetivos, tamaño de los textos, etc.

El tercer grupo de la competición (Biçici, 2015) presentó un sistema basado en máquinas de traducción referencial. La idea principal es comparar cómo se parecen los textos originales cuando son traducidos a otros idiomas.

3 Método propuesto

El método propuesto para el cálculo de la similitud entre textos se basa en la combinación de diversos indicadores en los que el factor común es el uso de *word embeddings* para representar las palabras.

3.1 Agregando *word embeddings*

La primera aproximación que forma parte de nuestro sistema consiste en la aplicación de alguna medida de similitud entre los vectores que proporciona la representación de *word embeddings*.

Dado que estos vectores están asociados a palabras individuales, para cada texto es necesario aplicar algún método que unifique los vectores de cada palabra generando un único vector. Tras probar distintas funciones de agregación, se eligió como representación la media aritmética de los vectores, obtenida sumando todos los vectores de palabras y dividiendo entre el número de palabras que componen el texto (ecuación 1).

$$\vec{V_d} = \frac{\sum_{i=0}^{n} \vec{v_i}}{n} \quad (1)$$

Una vez obtenido el vector agregado para cada texto se aplican medidas de similitud 'tradicionales', como son la distancia euclídea y la similitud del coseno.

3.2 Alineamiento

Otro indicador que obtenemos sobre el grado de similitud entre los textos se basa en la idea general del alineamiento de palabras. En este caso se lleva a cabo el emparejamiento de las palabras que, perteneciendo cada una a un texto distinto, guardan alguna relación semántica entre ellas. Una vez completada la fase de alineamiento de palabras se utiliza la distancia entre los *word embeddings* de cada palabra para finalmente aplicar una función de agregación que nos proporcione el grado de similitud global entre los textos.

El alineamiento se realiza habitualmente en un único sentido, buscando para cada palabra de la oración de menor tamaño aquella que debe formar pareja con ella de entre todas las que forman la oración más larga. Sin embargo, en nuestro método realizamos un alineamiento bidireccional, tras el cual se descartan los pares de palabras repetidos antes de tomar la decisión final. De este modo contamos con un mayor número de medidas parciales mejorando los resultados finales.

El proceso comienza con la construcción de un vector en el que existe una posición asociada a cada palabra del conjunto formado por la unión de los dos textos a analizar. Para cada palabra del primer texto se buscará una palabra del segundo texto con la que alinearla. Si la misma palabra existe en los dos textos se alineará consigo misma, y en caso contrario se calculará la distancia entre las representaciones vectoriales de la palabra de referencia del primer texto y todas las del segundo, con el fin de buscar la más cercana. El valor a introducir en la posición del vector asociada a la palabra de referencia será un 1 en el primer caso, y 1 menos la distancia en el segundo. Si el alineamiento no se puede llevar a cabo se introducirá un 0. Esto puede suceder si la palabra no está presente en el vocabulario del modelo de *word embeddings* utilizado.

El vector resultante representa indicadores individuales de similitud para las palabras que aparecen en los textos, por lo que es necesario aplicar una función de agregación.

En el algoritmo 1 se describe el proceso de forma más detallada, siendo posible aplicar distintas configuraciones para obtener las *selected_words* que se utilizan para generar el valor agregado de similitud (número total de palabras, palabras con valor distinto de cero, etc.).

En la figura 1 se muestra un ejemplo de alineamiento entre dos textos, incluyendo los valores de similitud obtenidos para cada par de palabras. El vector resultante tendrá once posiciones (el número de palabras distintas presentes en ambos textos) y el grado de similitud se calcula en este caso con las posiciones cuyo valor es distinto de cero, siendo el resultado $0,718$ sobre 1. La figura refleja la capa-

4	Dos oraciones son completamente equivalentes al significar la misma cosa.
3	Dos oraciones son prácticamente equivalentes pero algunos detalles difieren.
2	Dos oraciones son aproximadamente equivalentes pero alguna información importante difiere o no está.
1	Dos oraciones no son equivalentes pero son del mismo tema.
0	Dos oraciones son de diferentes temas.

Tabla 1: Descripción de los niveles de similitud para la tarea de SemEval 2015

Algoritmo 1 Alineamiento en pseudocódigo

Require: t_1, t_2 sets de tokens, m = modelo Word2Vec
Ensure: sim = similitud métrica

$vocab = t_1 + t_2$
$bow = [0] * vocab.size()$
for all w **in** t_1 **do**
  if w **in** t_2 **then**
    $bow[w] = 1$
  else if w **in** m **then**
    $bow[w] = max(m.similarity(w, t_2))$
  else
    $continue$
  end if
end for
for all w **in** t_2 **do**
  if w **in** m **then**
    $bow[w] = max(m.similarity(w, t_1))$
  else
    $continue$
  end if
end for
$values = (w\ for\ w\ in\ selected\_words)$
$sim = sum(values)/values.size()$
return sim

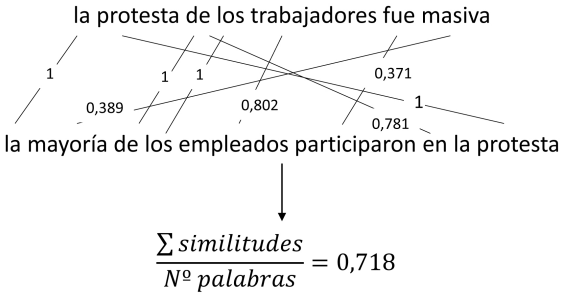


Figura 1: Ejemplo de alineamiento

cidad de establecer relaciones semánticas entre palabras distintas a través del modelo de *word embedding* ('trabajadores'-'empleados') y la posibilidad de tener una palabra alineada con más de una palabra distinta ('de').

3.3 Combinación

Con el objetivo de aprovechar los diferentes enfoques aportados por las soluciones propuestas en las secciones anteriores, hemos experimentado con un método de combinación basado en un algoritmo de aprendizaje automático supervisado para regresión. En este caso se construye un conjunto de datos de entrenamiento formado por los valores devueltos por los sistemas anteriormente descritos, generando un nuevo modelo que se pueda aplicar a nuevos pares de textos que se deseen analizar.

4 Experimentación

Como marco de referencia para evaluar el rendimiento de nuestra propuesta se recurrió a la última edición de la conferencia internacional SemEval, celebrada en 2015 (Agirre et al., 2015).

Entre los diferentes retos que se presentaron en este evento, hemos seleccionado la tarea para el español que consiste en, dadas dos oraciones, devolver un valor de similitud continuo entre 0 y 4. Para entender la diferencia de matices que hay entre los distintos valores, mostramos en la tabla 1 una pequeña descripción de cada uno de los cinco niveles en que se ha dividido el grado de similitud a obtener.

Teniendo en cuenta el alto poder de convocatoria de este evento y el nivel de los participantes, consideramos la tabla de resultados de esta tarea el mejor marco de referencia posible para validar nuestra propuesta. En la tabla 2 se muestran únicamente los mejores resultados de SemEval 2015.

Sistema	*Pearson*	Δ
Baseline-tokencos	0,529	0,0 %
RTM-DCU	0,582	5,4 %
UMDuluth	0,594	6,5 %
ExBThemis	**0,706**	**17,7 %**

Tabla 2: Resultados SemEval 2015

Valor	Frases
4	El espécimen es excepcional por las partes conservadas: un cráneo y mandíbula y un molde interno de la caja craneal. El espécimen comprende la mayor parte de la cara y mandíbula con los dientes y un molde interno de la caja craneal.
3	"Time 100" es una lista de las 100 personas más influyentes según la revista Time. La primera lista fue publicada en 1999 con las 100 personas más influyentes del siglo 20.
2	La "marinera" es un baile de pareja suelto, el más conocido de la costa del Perú. La marinera es el baile nacional del Perú, y su ejecución busca hacerse con derroche de gracia, picardía y destreza.
1	La "cripta de Santa Leocadia" está situada en el interior de la catedral de Oviedo, Asturias. Esteban Báthory fue sepultado en la cripta de la catedral de Wawel en Cracovia.
0	El río atraviesa la importante ciudad de Puebla de Zaragoza, la cuarta más poblada del país. El "Grêmio Esportivo Bagé" es un club de fútbol brasileño, de la ciudad de Bagé en el estado de Rio Grande do Sul.

Tabla 3: Ejemplos de pares de frases para cada nivel de similitud

4.1 Conjunto de datos

El conjunto de datos contiene pares de oraciones extraídas de los artículos de la Wikipedia y consta como es habitual de dos partes, una de *train* y otra de *test*. Cada parte tiene 324 y 251 pares de oraciones respectivamente. En la tabla 3 se muestra un ejemplo extraído del conjunto de datos para cada uno de los niveles de similitud establecidos para la tarea. Los valores de similitud de las oraciones fueron asignados calculando la media de las asignaciones manuales realizadas por cinco jueces humanos. Las oraciones en general están bien construidas y usan un lenguaje formal.

Se realizó un pequeño pre-procesado que consistió únicamente en eliminar los signos de puntuación. Aunque es una práctica común, la eliminación de palabras huecas o *stop words* no se llevó a cabo al considerarse que aportaban información relevante al proceso de alineamiento.

4.2 Modelos Word2Vec

El sistema de *word embedding* que hemos empleado en esta propuesta para obtener la representación vectorial de las palabras está basado en la implementación de Word2Vec (Mikolov et al., 2013) que encontramos en la herramienta Gensim (Řehůřek y Sojka, 2010). A través de ella hemos generado un primer modelo (Modelo-1) con textos en español provenientes de artículos de la Wikipedia[2], manteniendo así el mismo dominio que encontramos en los datos de la tarea de SemEval que nos sirve de referencia.

Este modelo se creó con vectores de 300 dimensiones, haciendo uso de la opción *negative sample* para eliminar palabras de ruido y ejecutando un total de 20 iteraciones para reforzar el entrenamiento y mejorar los resultados, siendo cada vez más estricto el factor de aprendizaje en cada una de estas etapas.

Además de generar este nuevo modelo también se han llevado a cabo experimentos utilizando un modelo extendido (Cardellino, 2016) (Modelo-2). Para su construcción se utilizaron, además de textos de la Wikipedia, otras fuentes de datos como el corpus AnCora-ES [3] o Europarl [4].

La experimentación con este modelo extendido nos permitirá comprobar los efectos de introducir en el modelo palabras extraídas de otros tipos de documentos con sus respectivos contextos, lo cual afectará a las distintas métricas aquí expuestas que hacen uso de dicho modelo.

4.3 Resultados

A la hora de evaluar los resultados, se ha utilizado la medida estadística que se aplicó en la tarea original del SemEval, el coeficiente de correlación de Pearson.

En las tablas 4 y 5 se muestran los resultados obtenidos por los métodos aquí propuestos, así como el mejor resultado registrado en la tarea de SemEval 2015 (ExBThemis). Aparecen tanto los métodos individuales como el resultado de combinar dichos métodos junto a la diferencia del resultado respecto al

[2]https://dumps.wikimedia.org/eswiki/latest/

[3]http://clic.ub.edu/corpus/

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

sistema ExBThemis. En la tabla 4 vemos los resultados obtenidos con el modelo de *word embeddings* creado exclusivamente con datos de la Wikipedia (Modelo-1), mientras que en la tabla 5 vemos los resultados obtenidos utilizando el modelo extendido con datos adicionales que no provienen de la Wikipedia (Modelo-2).

Sistema	*Pearson*	Δ
ExBThemis	0,706	-
Euclidea (E)	0,509	-19,7 %
Coseno (C)	0,467	-23,9 %
Alineamiento (A)	0,692	-1,4 %
Combinado (E+C+A)	**0,713**	**0,7 %**

Tabla 4: Resultados con el Modelo-1

Sistema	*Pearson*	Δ
ExBThemis	0,706	-
Euclidea (E)	0,642	-6,4 %
Coseno (C)	0,646	-5,9 %
Alineamiento (A)	0,687	-1,8 %
Combinado (E+C+A)	**0,723**	**1,8 %**

Tabla 5: Resultados con el Modelo-2

En ambos casos las métricas empleadas de forma individual no son capaces de obtener resultados que superen los obtenidos por el sistema ExBThemis, aunque el método de alineamiento aquí planteado se queda a menos de dos puntos porcentuales. Teniendo en cuenta que nos comparamos con el mejor sistema de entre todos los que se presentaron para resolver esta tarea en la edición 2015 de SemEval, dicho resultado puede considerarse relevante por sí mismo. Sin embargo, es la combinación de las diferentes métricas la que arroja las mejores cifras superando significativamente nuestra referencia, especialmente en el caso de utilizar el Modelo-2.

Precisamente en el mejor de los casos, utilizando el Modelo-2, la inserción de textos de diferentes dominios a la hora de generar el modelo de *word embedding* afecta negativamente al método de alineamiento, quizás por las diferencias en cuanto a las relaciones semánticas entre palabras que son aprendidas partiendo de dichos textos. Sin embargo, ese enriquecimiento del vocabulario favorece enormemente a las métricas más tradicionales, que consiguen resultados considerablemente mejores beneficiando con ello a la combinación, que compensa así la ligera pérdida sufrida por el alineamiento.

5 Conclusiones y trabajo futuro

En este trabajo hemos explorado la forma de aprovechar un modelo de *word embedding* para mejorar los resultados en una tarea de similitud semántica de textos. Nuestro principal objetivo era evaluar la mejora que se puede obtener en esta tarea sin hacer uso de otros recursos que no sean la representación vectorial en el espacio continuo. Para ello hemos definido dos maneras de calcular la similitud semántica de textos. Por un lado, mediante un alineamiento entre textos bidireccional y ponderado en función del parecido de las palabras emparejadas. La otra técnica se apoya en los vectores de palabras para construir representaciones vectoriales de los textos que son comparadas para determinar el grado de similitud entre ellos. Los experimentos sobre un corpus de la competición SemEval de 2015 para español muestran que el método de alineamiento se comporta de manera muy satisfactoria, quedando muy cerca del mejor sistema de la competición. En el caso de la técnica basada en la representación vectorial de textos, los resultados son peores pero aportan conocimiento complementario. Esto se demuestra con el hecho de que cuando se combinan las salidas de ambas técnicas se consiguen mejorar ambos resultados, superando incluso al mejor de los sistemas de la competición. También hemos experimentado con dos modelos distintos de *word embedding*, uno de ellos entrenado exclusivamente con textos de Wikipedia y otro, más extenso, también con textos de la Wikipedia más textos de otras fuentes. Dado que el corpus de SemEval está creado con textos de Wikipedia, este experimento nos ha permitido evaluar la sensibilidad de las técnicas a los textos usados para el entrenamiento de los modelos. Los experimentos han mostrado que la técnica basada en la representación vectorial de textos se beneficia de un modelo de *word embedding* más extenso aunque entrenado con textos de un dominio diferente. Por su parte, la técnica de alineamiento es penalizada levemente por el cambio de dominio aunque sus resultados siguen siendo mejores que los de la técnica de representación vectorial. En cualquier caso, la combinación sigue mejorando los resultados de ambas técnicas, quedando también por encima del mejor sistema de la competición. Como trabajo futuro estamos especialmente interesados en analizar el comportamiento de nuestros sistemas en un con-

texto multilingüe. El buen comportamiento de las técnicas de *word embedding* a la hora de comparar palabras de distintos idiomas (Mikolov, Le, y Sutskever, 2013) y el hecho de que nuestra aproximación sólo se base en la información de los modelos de *word embedding* nos hace pensar que puede comportarse bien para calcular la similitud de textos en distintos idiomas.

Agradecimientos

Este trabajo ha sido financiado a través del proyecto de investigación AORESCU (P11-TIC-7684 MO).

Bibliografía

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, y R. Mihalcea. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, páginas 252–263.

Bär, D., T. Zesch, y I. Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. En *ACL (Conference System Demonstrations)*, páginas 121–126.

Biçici, E. 2015. Rtm-dcu: Predicting semantic similarity with referential translation machines. *SemEval-2015*.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, y P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Noviembre.

Hänig, C., R. Remus, y X. D. L. Puente. 2015. Exb themis: Extensive feature extraction from word alignments for semantic textual similarity. *SemEval-2015*, página 264.

Jiang, J. y D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Karumuri, S., V. Vuggumudi, y S. Chitirala. 2015. Umduluth-blueteam: Svcsts-a multilingual and chunk level semantic similarity system. *SemEval-2015*, página 107.

Lin, D. 1998. Extracting collocations from text corpora. En *First workshop on computational terminology*, páginas 57–63. Citeseer.

Mihalcea, R., C. Corley, y C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. En *AAAI*, volumen 6, páginas 775–780.

Mikolov, T., Q. Le, y I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Pennington, J., R. Socher, y C. Manning. 2014. Glove: Global vectors for word representation. En *EMNLP*, volumen 14, páginas 1532–1543.

Řehůřek, R. y P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta, Mayo. ELRA. `http://is.muni.cz/publication/884893/en`.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Zou, W., R. Socher, D. Cer, y C. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. En *EMNLP*, páginas 1393–1398.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 75-82 recibido 29-03-2016 revisado 28-04-2016 aceptado 03-05-2016

Segmentación de palabras en español mediante modelos del lenguaje basados en redes neuronales*

Spanish word segmentation through neural language models

Yerai Doval
Grupo COLE, Dpto. de Informática
E.S. de Enxeñaría Informática
Universidade de Vigo
Campus de As Lagoas
32004 – Ourense (España)
yerai.doval@uvigo.es

Carlos Gómez-Rodríguez, Jesús Vilares
Grupo LYS, Dpto. de Computación
Facultade de Informática
Universidade da Coruña
Campus de A Coruña
15071 – A Coruña (España)
{cgomezr, jvilares}@udc.es

Resumen: En las plataformas de *microblogging* abundan ciertos *tokens* especiales como los *hashtags* o las menciones en los que un grupo de palabras se escriben juntas sin espaciado entre ellas; p.ej.: `#añobisiesto` o `@ryanreynoldsnet`. Debido a la forma en que se escriben este tipo de textos, este fenómeno de ensamblado de palabras puede aparecer junto a su opuesto, la segmentación de palabras, afectando a cualquier elemento del texto y dificultando su análisis. En este trabajo se muestra un enfoque algorítmico que utiliza como base un modelo del lenguaje —en nuestro caso concreto uno basado en redes neuronales— para resolver el problema de la segmentación y ensamblado de palabras, en el que se trata de recuperar el espaciado estándar de las palabras que han sufrido alguna de estas transformaciones añadiendo o quitando espacios donde corresponda. Los resultados obtenidos son prometedores e indican que tras un mayor refinamiento del modelo del lenguaje se podrá sobrepasar al estado del arte.
Palabras clave: segmentación de palabras, ensamblado de palabras, español, modelos del lenguaje basados en redes neuronales

Abstract: In social media platforms special tokens abound such as hashtags and mentions in which multiple words are written together without spacing between them; e.g. `#leapyear` or `@ryanreynoldsnet`. Due to the way this kind of texts are written, this word assembly phenomenon can appear with its opposite, word segmentation, affecting any token of the text and making it more difficult to perform analysis on them. In this work we show an algorithmic approach based on a language model —in this case a neural model— to solve the problem of the segmentation and assembly of words, in which we try to recover the standard spacing of the words that have suffered one of these transformations by adding or deleting spaces when necessary. The promising results indicate that after some further refinement of the language model it will be possible to surpass the state of the art.
Keywords: word segmentation, word assembly, spanish, neural language models

1 Introducción

La popularización de la Web 2.0 y las redes sociales en los últimos años ha dado como resultado una fuente de datos prácticamente inagotable. Millones de usuarios crean contenidos cada minuto sobre temáticas muy diversas y en formatos de lo más variado: desde textos manifestando la opinión de un usuario ante un producto determinado hasta vídeos en los que se relata la vida cotidiana de un usuario. Tal y como ya se ha demostrado, el procesamiento y análisis de esta información puede resultar de gran utilidad en un amplio abanico de ámbitos: inteligencia de negocio (Gallinucci, Golfarelli, y Rizzi, 2013), lucha antiterrorista y contra el acoso (Alonso et al., 2015), predicción de la valoración de

* Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad español a través de los proyectos FFI2014-51978-C2-1-R y FFI2014-51978-C2-2-R, y por la Xunta de Galicia a través del programa Oportunius. We gratefully acknowledge NVIDIA Corporation for the donation of a GTX Titan X GPU used for this research.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

líderes políticos por la ciudadanía (Alonso y Vilares, 2016) o incluso predicción de brotes epidémicos (Cebrian, 2012). Este hecho pone de manifiesto la importancia del desarrollo de sistemas que permitan explotar de forma efectiva este tipo de fuentes.

Con respecto a las fuentes de datos en forma de texto, las plataformas de *microblogging* como Twitter gozan de gran popularidad actualmente. En éstas, abundan ciertos fenómenos de escritura no estándar, englobados en lo que se conoce como *texting*[1], así como *tokens*[2] especiales como los *hashtags*, muy empleados por los usuarios para etiquetar una entrada (p.ej.: Feliz 29 #añobisiesto) o usados en lugar de una o varias palabras del cuerpo del mensaje en que aparecen (p.ej.: Este año es #añobisiesto). En diversas aplicaciones, sin embargo, los *hashtags* que contienen varias palabras son tratados como una sola, reduciendo así la profundidad del análisis efectuado sobre estos *tokens* (p.ej.: en el caso de #añobisiesto, no se tendrían en cuenta las palabras `año` y `bisiesto` que conforman el *hashtag*). De igual forma, bien debido a errores de escritura o bien a la influencia del *texting*, cualquier conjunto consecutivo de palabras podrían aparecer reunidas en un mismo *token* (p.ej.: `nopuedeser!`) o incluso fragmentadas algunas de ellas en múltiples *tokens* (p.ej.: `im posible!`). En cualquier caso, estos fenómenos plantean serios problemas para todos aquellos sistemas que en algún momento trabajen a nivel de palabra, como todos los que se apoyan en diccionarios y lexicones, ya que no dispondrán de las entradas correctas; p.ej.: el sistema puede reconocer las palabras `año` y `bisiesto` pero no `añobisiesto`, ni tampoco `bi` y `siesto` por separado.

En este trabajo haremos frente a la incorrecta segmentación de palabras de un texto a través de un sistema formado por dos componentes. En primer lugar, un algoritmo de búsqueda basado en el *beam search* que tratará de encontrar la mejor segmentación posible para un texto de entrada. En segundo lugar, un modelo del lenguaje basado en redes neuronales que guiará al algoritmo en la toma de decisiones. El principal motivo de elegir una aproximación basada en redes neuronales es su probada eficacia en la tarea de modelado del lenguaje (Mikolov y Zweig, 2012; Józefowicz, Zaremba, y Sutskever, 2015). En cualquier caso, se ha probado también un modelo del lenguaje basado en *n-gramas* para confirmar este punto y se ha enfrentado nuestro sistema contra el conocido Word Breaker del Oxford Project de Microsoft[3], el cual también posee como base un modelo de *n-gramas*, aunque mucho más maduro.

Durante la evaluación se ha estudiado el impacto en el rendimiento general del sistema de los parámetros ajustables del algoritmo así como del modelo del lenguaje empleado. Los resultados obtenidos muestran el gran potencial de nuestra aproximación, debido en gran medida al modelo del lenguaje basado en redes neuronales. Así, con el refinamiento de este componente en el trabajo futuro se espera superar el excelente rendimiento del sistema Word Breaker, al cual nos hemos aproximado en gran medida a pesar de disponer de un corpus de entrenamiento mucho más reducido (Wikipedia v. *Web crawl*).

El resto del documento se estructura como sigue. En la Sección 2 se repasa el estado del arte. La Sección 3 muestra la aproximación propuesta para lidiar con el problema introducido en la presente sección. En la Sección 4 mostramos los resultados obtenidos al evaluar dicha aproximación. Por último, la Sección 5 expone las conclusiones obtenidas así como algunas líneas de trabajo futuro.

2 Trabajo relacionado

La segmentación de palabras es un paso previo muy importante en diversos tipos de sistemas: traducción automática (Koehn y Knight, 2003), recuperación de información (Alfonseca, Bilac, y Pharies, 2008) o reconocimiento del habla (Adda-decker, Adda, y Lamel, 2000) son algunos ejemplos. La segmentación de palabras para idiomas asiáticos cuenta con gran atención en la comunidad investigadora, debido principalmente a la inexistencia de caracteres de separación de palabras en estos idiomas (Suzuki, Brockett, y Kacmarcik, 2000; Huang y Zhao, 2007; Wu y Jiang, 1998). Por otra parte, los idiomas con una morfología compleja como el alemán, el noruego o el griego, se benefician también en gran medida de este tipo de sistemas de segmentación de palabras (Alfonseca, Bilac, y Pharies, 2008; Koehn y Knight, 2003).

[1] Modo de escritura en el que prima la tendencia a escribir como se habla en entornos informales.

[2] En este trabajo, secuencia de caracteres exceptuando los de espaciado delimitada por éstos.

[3] `https://www.projectoxford.ai/demo/weblm`

Con respecto a la problemática específica de la Web, en primer lugar aparecieron los sistemas que trataban de analizar URLs. Dado que estos elementos tampoco permiten el uso de espacios, las palabras que pudieran contener deberían ser obtenidas mediante un proceso de segmentación (Chi, Ding, y Lim, 1999; Wang, Thrasher, y Hsu, 2011). Más tarde, con la llegada de la Web 2.0, el uso de los *hashtags* se hizo muy popular, requiriendo de nuevo un proceso de segmentación para poder ser analizados convenientemente (Srinivasan, Bhattacharya, y Chakraborty, 2012; Maynard y Greenwood, 2014).

En general, el proceso seguido por la mayor parte de estos trabajos puede resumirse en (1) obtención de todas las posibles segmentaciones, para lo que pueden emplearse conjuntos de reglas (Koehn y Knight, 2003) o (también en conjunción con) recursos como diccionarios y lexicones de palabras o morfemas (Kacmarcik, Brockett, y Suzuki, 2000), y (2) selección de la segmentación más probable según alguna función de puntuación, para lo que puede usarse el análisis sintáctico de las posibles segmentaciones (Wu y Jiang, 1998) o la secuencia de palabras más probable según algún modelo del lenguaje (Wang, Thrasher, y Hsu, 2011). Otras aproximaciones, muy empleadas por ejemplo para el idioma chino, consideran el problema de la segmentación de palabras como un problema de etiquetación. Bajo este enfoque, el sistema ha de asociar a cada carácter una etiqueta de entre las siguientes: comienzo de palabra, mitad de palabra, fin de palabra o palabra de carácter único. Recientemente, a los métodos tradicionales de entrenamiento supervisado como *Maximum Entropy* (Berger, Pietra, y Pietra, 1996) o *Conditional Random Fields* (Lafferty, McCallum, y Pereira, 2001) se han unido los basados en redes neuronales artificales debido al auge del denominado *aprendizaje profundo* (*deep learning*) (Bengio, 2009).

De los trabajos vistos en este apartado cabe destacar tres puntos que diferencian a nuestra propuesta: (1) No tienen en cuenta el ensamblado de palabras, al menos de forma explícita, generalmente porque los casos de uso considerados no lo requieren. (2) Los experimentos iniciales en este trabajo se han realizado sobre textos en castellano. (3) Los sistemas que basan buena parte de su procesamiento en el uso de lexicones y diccionarios pueden verse limitados en contextos en los que abunde el ruido —como la propia Web 2.0—, el cual no es nuestro caso.

3 Aproximación propuesta

El sistema que proponemos está formado por dos componentes bien diferenciados y que trabajan de forma conjunta: el modelo del lenguaje basado en redes neuronales y el algoritmo de segmentación/ensamblado. El primero sirve como proveedor de información sobre el lenguaje al segundo, que empleará este conocimiento para realizar las segmentaciones y ensamblados que correspondan para recuperar el espaciado entre palabras en un texto de entrada.

3.1 Modelo del lenguaje basado en redes neuronales

Tradicionalmente, el enfoque empleado para el modelado del lenguaje era el conteo de *n-gramas* (Brown et al., 1992) (secuencia de n caracteres o palabras) junto con el uso de técnicas de suavizado que permitieran considerar secuencias desconocidas durante el entrenamiento (Chen, 1998). Sin embargo, posteriormente se demostró que mediante el uso de técnicas basadas en redes neuronales y representaciones continuas de los *tokens* de entrenamiento es posible el desarrollo de modelos del lenguaje que superan en rendimiento a los tradicionales (Mikolov y Zweig, 2012).

Uno de los tipos de redes neuronales más empleados para el modelado del lenguaje son las redes neuronales recurrentes (Mikolov y Zweig, 2012), en las cuales los elementos de procesado (neuronas) no sólo poseen información sobre su entrada actual (en un instante de tiempo t) sino también información sobre su operación con su entrada anterior (en el instante de tiempo $t-1$). Esta propiedad convierte a este tipo de redes en una herramienta muy adecuada para el modelado de series temporales, en general, y de secuencias de unidades de texto en particular, donde dada una secuencia de datos (caracteres) de entrada se obtenga una predicción para el dato (carácter) siguiente en la secuencia. De entre todas las redes de la familia de las recurrentes, concretamente las LSTMs (*Long Short Term Memory*) (Hochreiter y Schmidhuber, 1997) han demostrado un buen rendimiento en la tarea de modelado del lenguaje natural (Józefowicz, Zaremba, y Sutskever, 2015). Esto es debido en gran parte a su capacidad para retener la información de las

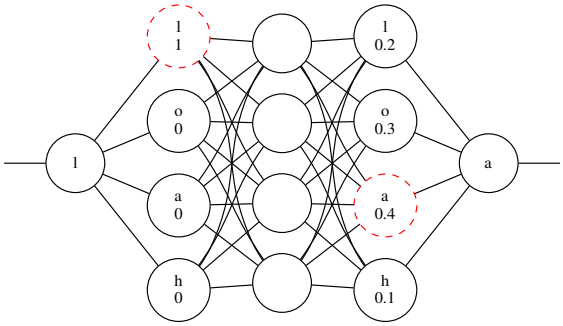


Figura 1: Predicción del siguiente carácter usando una red neuronal. El conjunto de caracteres considerados es $C = \{a, h, l, o\}$.

entradas más alejadas en el pasado, considerando así un contexto más amplio a la hora de procesar la entrada actual. Los mecanismos que facilitan dicha capacidad se encuentran ubicados en el diseño de cada uno de los elementos de procesamiento de este tipo de redes: las denominadas *puertas* (*gates*) de entrada, salida y olvido, junto con la *celda* de memoria. Estos elementos se encargan de regular el flujo de información dentro de cada unidad de procesamiento de la red, determinando así cuándo un dato es importante y debe ser recordado y cuándo debería ser olvidado por no ser relevante.

Por los motivos anteriormente expuestos, en nuestro trabajo hemos hecho uso de modelos del lenguaje basados en redes neuronales obtenidos mediante el entrenamiento de LSTMs a nivel de carácter. El objetivo de estas redes es el de estimar la distribución de probabilidad del carácter siguiente a uno dado. En la Figura 1 se muestra una representación gráfica de una red neuronal realizando el proceso de predicción. A nivel de implementación hemos empleado el *framework* de aprendizaje profundo `Torch`[4] junto con la implementación de LSTM realizada por Andrej Karpathy,[5] configurada en 3 capas ocultas de 1 000 neuronas cada una.

3.2 Algoritmo de segmentación y ensamblado de palabras

Asumiendo que se dispone de un modelo del lenguaje a nivel de caracteres entrenado sobre textos similares a los que nuestro sistema recibirá como entrada, el algoritmo de segmentación y ensamblado de palabras se muestra

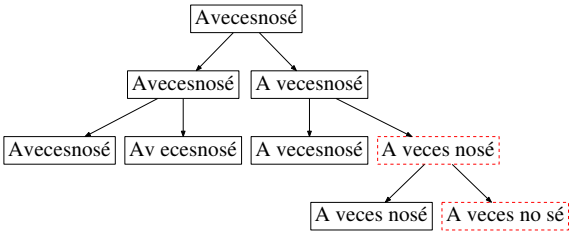


Figura 2: Ejecución del algoritmo con nivel máximo de recursión 3 y selección de un solo candidato al alcanzar resultados parciales.

en el Algoritmo 1.[6] La idea fundamental es la siguiente: primero, se eliminan los caracteres de espaciado del texto de entrada (primera sentencia de la línea 18). Luego, para cada carácter del texto resultante se obtiene mediante el modelo del lenguaje la distribución de probabilidad para el siguiente carácter, teniendo en cuenta además los caracteres anteriores al actual (línea 4). Se introducirá entonces un carácter de espacio justo después del actual siempre que aquél posea una mayor probabilidad que el carácter siguiente del texto o que su probabilidad sea alta de acuerdo con el parámetro *threshold* (condición en línea 6, inserción en línea 11) explicado más abajo. En cualquier caso, se ramifica la ejecución mediante una llamada recursiva que considere el no añadir o eliminar el carácter de espacio para una mayor exhaustividad en la búsqueda de la solución óptima (línea 8). Con el fin de mantener el tiempo de ejecución del algoritmo dentro de unos límites aceptables así como dirigir este proceso de búsqueda, la profundidad de recursión se limita terminando la ejecución al alcanzar un determinado nivel en las llamadas recursivas (línea 13). En este momento se dispone de un conjunto de resultados parciales de entre los que se escogen los mejores según la media de sus probabilidades logarítmicas (línea 24) y con los que se reinicia el algoritmo (bucle en la línea 19). Este proceso se repite hasta que no quede más texto de entrada que procesar (línea 25). Una representación gráfica de este proceso puede observarse en la Figura 2.

Varios aspectos de este algoritmo son ajustables por el usuario por medio de tres parámetros de entrada: *threshold*, *stop* y *prun*. El primero es empleado para permitir los casos de recursión siempre que el carácter actual considerado se encuentre entre los

[4] `http://torch.ch/`

[5] `https://github.com/karpathy/char-rnn`

[6] Código y recursos disponibles en `https://cloud.wyffy.com/index.php/s/YFgcHU1inp2WBMs`.

Algoritmo 1 Algoritmo de segmentación y ensamblado de palabras. *text* contiene el texto de entrada, *acctext* acumula el texto de salida y $\top_n(C)$ representa el conjunto de los n elementos más probables del conjunto C.

```
1: function ESPACIAR(text, model, threshold, stop, level, acctext)
2:     resultado ← ∅ ; b ← stop − level                    ▷ Límite para la recursión
3:     for i ∈ [1, |text|] do
4:         pred ← predecir(text_i, acctext, model)          ▷ Lista de tuplas (carácter, probabilidad)
5:         acctext ← acctext + text_i
6:         if pred_espc > pred_text_{i+1} ∨ espc ∈ ⊤_threshold(pred) then
7:             if pred_text_{i+1} ∈ ⊤_threshold(pred) then
8:                 recur ← espaciar(text_{i+1..|text|}, model, threshold, stop, level + 1, acctext)
9:                 resultado ← resultado ∪ recur ; b ← b − 1
10:            end if
11:            acctext ← acctext + espc
12:        end if
13:        if b =< 0 then break end if
14:    end for
15:    return resultado ∪ {(acctext, P(acctext, model))}   ▷ Lista de tuplas (cadena, probabilidad)
16: end function
17: function SEGMENTAR(text, model, threshold, stop, prun)
18:    text ← remove(text, espc) ; partials ← {([], 0)}
19:    repeat
20:        tmp ← ∅
21:        for part ∈ partials do
22:            tmp ← tmp ∪ espaciar(text_{l+1..|text|}, model, threshold, stop, 1, part_cadena)
23:        end for
24:        partials ← ⊤_prun(tmp)
25:    until ∀part_cadena, l ← |remove(part_cadena, espc)|, text_{l+1..|text|} = []
26:    return partials
27: end function
28: segmentar(text, model, threshold, stop, prun)        ▷ Llamada inicial a la función
```

threshold primeros más probables de la predicción realizada para el siguiente carácter. *stop* determina el nivel máximo de profundidad a la que puede llegar el algoritmo con las llamadas recursivas. Por último, *prun* determina cuántos de los mejores resultados parciales se escogen tras una parada del algoritmo (por llegar a la profundidad máxima *stop*) y con los que se reiniciará la ejecución. El caso particular de $stop = 2$ es equivalente a una búsqueda tipo *beam search*, ya que se seleccionan los mejores nodos después de expandir cada nivel.

La ejecución del algoritmo descrito dará como resultado una lista con las posibles segmentaciones para una entrada *texto* y un modelo del lenguaje *model*.

4 Evaluación

Siendo el sistema descrito nuestra primera propuesta para resolver el problema de la segmentación y ensamblado de palabras, era de nuestro interés el realizar una evaluación que nos permitiese observar sus puntos fuertes y débiles. Así, el análisis del rendimiento realizado trata de considerar todos aquellos factores que podrían tener un mayor impacto sobre el rendimiento final del sistema, cuantificado en términos de *precisión*, medida como el número de ocasiones en las que el sistema da el resultado correcto sobre el número total de instancias del conjunto de evaluación, y *tiempo de ejecución* en segundos.

Con respecto al modelo del lenguaje, se ha comparado la puntuación de validación (*negative log-likelihood*, *NLL*) obtenida en el entrenamiento de cada modelo con la precisión y tiempo de ejecución final del sistema. Cabe esperar que a menor puntuación de validación mayor precisión, quedando como incógnita el comportamiento del tiempo de ejecución. Durante el entrenamiento se han empleado los artículos de parte del volcado de la Wikipedia en castellano del 28/02/2015, filtrado mediante el programa `wikiextractor`[7] y con las expresiones propias del lenguaje de marcado de la Wikipedia eliminadas. Del resultado se extrajo el primer 25 % de líneas de texto, totalizando 4 000 000 de líneas, 5 666 491 de oraciones, 134 360 571 de palabras, 1 245 884 de palabras distintas, 815 525 127 de caracteres y 5 128 caracteres distintos. Para entrenar el mejor modelo del lenguaje aquí presentado fueron necesarios aproximadamente dos días.

En el caso del algoritmo propuesto, se ha estudiado el impacto de los parámetros de entrada *threshold*, *stop* y *prun*. Dado que a ma-

[7] `https://github.com/attardi/wikiextractor`

	NLL	*P*	*t* (s)
M. neuronas	3.1178	0.0	3799
	2.9974	0.0	8259
	1.5485	0.52	6888
	1.3198	0.63	6442
	0.9898	**0.71**	**5913**
		P	*t* (s)
threshold	10	0.66	2258
	20	**0.70**	**4401**
	50	0.71	6396
prun	1	0.41	3669
	2	0.70	5471
	3	0.77	7160
	4	**0.79**	**8794**
	5	0.82	21478
stop	2	0.65	3655
	3	**0.71**	**5913**
	4	0.75	12714
	5	0.77	36594
M. *n-gramas*	–	0.51	15722

Tabla 1: Resultados de precisión y tiempo de ejecución para cada modelo del lenguaje y valor de los parámetros del algoritmo probados.

yor valor asignado a estos parámetros mayor sería la exhaustividad de la búsqueda, cabría esperar encontrar curvas ascendentes de precisión y tiempo de ejecución.

Para la evaluación se ha tomado el último 25 % de líneas de texto del volcado procesado de la Wikipedia. Además, se ha dividido el texto resultante en secuencias de caracteres de longitud mínima 20, realizando los cortes en el siguiente carácter de espaciado, y para cada secuencia se han eliminado los caracteres de espaciado. Del resultado se escogen de forma aleatoria 1 000 líneas de texto (5 065 palabras, 2 722 palabras distintas, 42 562 caracteres y 94 caracteres distintos).

Los datos reflejados en la Tabla 1 se muestran coherentes con nuestra intuición inicial. Los mejores modelos del lenguaje permiten obtener los mejores resultados además de reducir el tiempo de ejecución. Este último dato surge como consecuencia de una menor ramificación en la ejecución del algoritmo con estos modelos, los cuales son capaces de descartar con mayor seguridad opciones de segmentación poco probables. Únicamente nos encontramos con un dato anómalo para el peor modelo en el que el tiempo de ejecución alcanza un valor mínimo. En este caso, de nuevo el modelo descarta con mucha seguridad la mayoría de las opciones de segmentación aunque, en vista de los resultados de precisión, lo hace de forma incorrecta.

En cuanto a los parámetros del algoritmo, cuanto mayor sean los valores asignados mayor será el tiempo de ejecución, con ciertos valores estableciendo límites entre lo que sería una ejecución más o menos rápida y otra mucho más lenta. Así, pueden considerarse valores adecuados $threshold = 20$, $prun = 4$ y $stop = 3$, considerando incrementar ligeramente cada uno de ellos si se desea maximizar el rendimiento a costa del tiempo de ejecución.

En conjunto, asignando a los parámetros del algoritmo los valores $threshold = 30$, $prun = 5$ y $stop = 4$, la precisión alcanzada es de 0.82. En estos momentos, el factor limitante en cuanto a rendimiento y tiempo de ejecución parece ser el modelo del lenguaje, tal y como cabría esperar. La dificultad en este punto es la mejora de este componente, lo cual requiere en cualquier caso de mayores cantidades de tiempo para el proceso de entrenamiento. Sin embargo, los datos obtenidos nos indican que invirtiendo más recursos en este proceso se podrá obtener un sistema muy competitivo, dado que no hemos encontrado todavía evidencias de que la tendencia creciente del rendimiento del sistema con respecto al ajuste del modelo del lenguaje vaya a detenerse próximamente.

Para contrastar los resultados obtenidos, en primer lugar hemos probado a sustituir el modelo del lenguaje empleado por uno basado en *n-gramas*, con el objetivo de validar el enfoque neuronal. Se construyó pues, mediante el paquete *software* `kenlm`,[8] un modelo de orden 5 en el que se dejó el resto de parámetros del paquete en su valor por defecto. Empleando de nuevo los valores $threshold = 30$, $prun = 5$ y $stop = 4$ en el algoritmo, la precisión obtenida es de 0.51, claramente inferior a la conseguida por nuestro modelo basado en redes neuronales.

En segundo lugar, hemos comparado el rendimiento de nuestro sistema contra el de uno bien conocido en el ámbito de la segmentación: el Word Breaker de Microsoft. Realizando una ejecución con el *Bing body model* de orden 5, este sistema obtiene una precisión de 0.85, posicionándose ligeramente por encima de nuestra propuesta. Llegados a este punto, es conveniente resaltar que la ejecución del Word Breaker no se realizó sobre el mismo conjunto de evaluación empleado hasta ahora, ya que debido a limitaciones de este sistema fue necesario procesar las instan-

[8] `http://kheafield.com/code/kenlm.tar.gz`

cias de evaluación disponibles de modo que sólo contuvieran caracteres alfanuméricos del conjunto de minúsculas de ASCII. Estas limitaciones surgen muy probablemente de la necesidad de reducir efectos indeseados como la dispersión en el modelo del lenguaje, basado en *n-gramas*. Además, el sistema de Microsoft cuenta con la ventaja añadida de haber sido entrenado con un *Web crawl*, un corpus mucho más extenso y completo que el empleado para entrenar nuestras redes neuronales. En definitiva, aun operando en un dominio mucho más restringido y alimentado por mejores recursos lingüísticos, el Word Breaker no consigue resultados notablemente mejores, lo cual indica la viabilidad y prometedor futuro de nuestra propuesta. De cualquier modo, las razones para usar este sistema para la comparación han sido su nivel de madurez, fácil accesibilidad a través de un API REST y su disponibilidad para trabajar con textos en castellano.

5 Conclusiones y trabajo futuro

En este trabajo hemos presentado nuestra primera propuesta para la segmentación y ensamblado de palabras en textos escritos en español. Nuestra aproximación hace uso de un algoritmo parametrizable que aprovecha la información condensada en un modelo del lenguaje a nivel de caracteres basado en redes neuronales, implementado como una red neuronal recurrente tipo LSTM. Los experimentos realizados nos han permitido observar el rendimiento y tiempo de ejecución del sistema en función de los valores asignados a los parámetros del algoritmo y el modelo del lenguaje empleado. Se han observado así los valores más convenientes para dichos parámetros de modo que se contenga el tiempo de ejecución del sistema y no se obstaculice su rendimiento. Por otra parte, los resultados obtenidos muestran claramente la importancia de continuar entrenando modelos del lenguaje que posean un mayor ajuste con los textos sobre los que se emplee el sistema para mantener la tendiencia ascendente del rendimiento, lo cual en estos momentos constituye el siguiente desafío. Por último, el modelo del lenguaje basado en redes neuronales se ha mostrado muy superior al de *n-gramas* empleado como *baseline*, y un sistema bien conocido como el Word Breaker, a pesar de trabajar en un ámbito más restringido, no ha destacado contra nuestra propuesta.

Como línea futura de trabajo nos centraremos en la parte de modelado del lenguaje, pieza fundamental del sistema y que presenta múltiples oportunidades de mejora. Por una parte consideramos la inclusión de un modelo del lenguaje entrenado en orden inverso. De esta forma no sólo poseeríamos información sobre los caracteres anteriores para realizar las predicciones de los siguientes, sino también sobre los posteriores. Por otra parte, la mejora del modelo del lenguaje empleado: bien mediante una búsqueda más exhaustiva del mejor conjunto de hiperparámetros para la arquitectura actual (LSTM) o bien a través de otros tipos de arquitectura como podrían ser las redes convolucionales. En cualquier caso, el proceso de entrenamiento con la arquitectura actual continúa su curso y se espera seguir obteniendo resultados cada vez mejores. También se considerará el uso de otros corpus de entrenamiento más extensos que el utilizado actualmente, como por ejemplo un mayor porcentaje del volcado de la Wikipedia o un *Web crawl* obtenido del proyecto Common Crawl.[9] Finalmente, tenemos intención de entrenar modelos para los idiomas en los que la segmentación es un problema bien conocido, como el chino o el alemán.

Bibliografía

Adda-decker, M., G. Adda, y L. Lamel. 2000. Investigating text normalization and pronunciation variants for german broadcast transcription. En *ICSLP'2000*, páginas 266–269.

Alfonseca, E., S. Bilac, y S. Pharies. 2008. Decompounding query keywords from compounding languages. En *Proc. of the 46th Annual Meeting of the ACL: Short Papers*, HLT-Short '08, páginas 253–256, Stroudsburg, PA, USA. ACL.

Alonso, M. A., C. Gómez-Rodríguez, D. Vilares, Y. Doval, y J. Vilares. 2015. Seguimiento y análisis automático de contenidos en redes sociales. En *Actas: III Congreso Nacional de i+d en Defensa y Seguridad, DESEi+d 2015*, páginas 899–906.

Alonso, M. A. y D. Vilares. 2016. A review on political analysis and social media. *Procesamiento del Lenguaje Natural*, 56:13–24.

[9] `https://commoncrawl.org/`

Bengio, Y. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Berger, A. L., S. D. Pietra, y V. J. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Brown, P. F., P. V. deSouza, R. L. Mercer, V. J. D. Pietra, y J. C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, Diciembre.

Cebrian, M. 2012. Using friends as sensors to detect planetary-scale contagious outbreaks. En *Proc. of the 1st International Workshop on Multimodal Crowd Sensing*, CrowdSens '12, páginas 15–16, New York, NY, USA. ACM.

Chen, S. F. 1998. An empirical study of smoothing techniques for language modeling. Informe técnico.

Chi, C.-H., C. Ding, y A. Lim. 1999. Word segmentation and recognition for web document framework. En *Proc. of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, páginas 458–465, New York, NY, USA. ACM.

Gallinucci, E., M. Golfarelli, y S. Rizzi. 2013. Meta-stars: Multidimensional modeling for social business intelligence. En *Proc. of the Sixteenth International Workshop on Data Warehousing and OLAP*, DOLAP '13, páginas 11–18, New York, NY, USA. ACM.

Hochreiter, S. y J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, C. y H. Zhao. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*, 21(3):8–20.

Józefowicz, R., W. Zaremba, y I. Sutskever. 2015. An empirical exploration of recurrent network architectures. En *Proc. of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France*, páginas 2342–2350.

Kacmarcik, G., C. Brockett, y H. Suzuki. 2000. Robust segmentation of japanese text into a lattice for parsing. En *Proc. of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, páginas 390–396, Stroudsburg, PA, USA. ACL.

Koehn, P. y K. Knight. 2003. Empirical methods for compound splitting. En *Proc. of the Tenth Conference on European Chapter of the ACL - Volume 1*, EACL '03, páginas 187–193, Stroudsburg, PA, USA. ACL.

Lafferty, J. D., A. McCallum, y F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. En *Proc. of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA*, páginas 282–289.

Maynard, D. y M. A. Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. En *LREC*, páginas 4238–4243.

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

Srinivasan, S., S. Bhattacharya, y R. Chakraborty. 2012. Segmenting web-domains and hashtags using length specific models. En *Proc. of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, páginas 1113–1122, New York, NY, USA. ACM.

Suzuki, H., C. Brockett, y G. Kacmarcik. 2000. Using a broad-coverage parser for word-breaking in japanese. En *Proc. of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, páginas 822–828, Stroudsburg, PA, USA. ACL.

Wang, K., C. Thrasher, y B.-J. P. Hsu. 2011. Web scale nlp: A case study on url word breaking. En *Proc. of the 20th International Conference on World Wide Web*, WWW '11, páginas 357–366, New York, NY, USA. ACM.

Wu, A. y Z. Jiang. 1998. Word segmentation in sentence analysis. En *Proc. of the 1998 International Conference on Chinese Information Processing*, páginas 169–180.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 83-90 recibido 28-03-2016 revisado 20-04-2016 aceptado 01-05-2016

COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques

COPOS: Corpus de Opiniones de Pacientes en Español. Aplicación de Técnicas de Análisis de Sentimientos

Flor Miriam Plaza-del-Arco, M. Teresa Martín-Valdivia, Salud María Jiménez-Zafra, M. Dolores Molina-González, Eugenio Martínez-Cámara
SINAI Research Group
Universidad de Jaén
E-23071 Jaén, Spain
fmpa0002@red.ujaen.es, {maite, sjzafra, mdmolina, emcamara}@ujaen.es

Abstract: Every day more users are interested in the opinion that other patients have about a physician or about health topics in general. According to a study in 2015, 62% of Spanish people access the Internet in order to be informed about topics related to health. This paper is focused on Spanish Sentiment Analysis in the medical domain. Although Sentiment Analysis has been studied for different domains, health issues have hardly been examined in Opinion Mining and even less with Spanish comments or opinions. Thus we have generated a corpus by crawling the website Masquemedicos with Spanish opinions about medical entities written by patients. We present this new resource, called COPOS (Corpus Of Patient Opinions in Spanish). To the best of our knowledge, this is the first attempt to deal with Spanish opinions written by patients about medical attention. In order to demonstrate the validity of the corpus presented, we have also carried out different experiments with the main methodologies applied in polarity classification (Semantic Orientation and Machine Learning). The results obtained encourage us to continue analysing and researching Opinion Mining in the medical domain.
Keywords: Corpus, patient opinions, medical domain, Spanish, Sentiment Analysis, polarity classification

Resumen: Cada día son más los usuarios interesados en la opinión que otros pacientes tienen sobre un médico o sobre temas de salud en general. De acuerdo con un estudio de 2015, el 62% de la población española consulta información en Internet acerca de temas relacionados con la salud. Este trabajo está centrado en el Análisis de Sentimientos en español aplicado al dominio médico. Aunque el Análisis de Sentimientos ha sido estudiado en diferentes dominios, el dominio de la salud apenas ha sido investigado, especialmente en opiniones escritas en español. Por ello, hemos generado un corpus en español con opiniones de pacientes sobre médicos a partir de la extracción de las mismas del portal web Masquemedicos. Este corpus ha sido denominado COPOS (Corpus Of Patient Opinions in Spanish - Corpus de Opiniones de Pacientes en Español). Hasta donde sabemos, es la primera vez que se intenta trabajar con opiniones en español sobre atención médica escritas por pacientes. Para demostrar la validez de este recurso, hemos realizado diferentes experimentos con las principales metodologías aplicadas en la tarea de clasificación de polaridad (Orientación Semántica y Aprendizaje Automático). Los resultados obtenidos nos animan a seguir investigando en el Análisis de Sentimientos en este dominio.
Palabras clave: Corpus, opiniones de pacientes, dominio médico, español, Análisis de Sentimientos, clasificación de polaridad

1 Introduction

The growth of medical documents available on the Internet in the last decade requires the development of more efficient systems to access this kind of information. A 2011 survey of the US population estimated that 59% of

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

all adults have looked online for information about health topics such as a specific disease or treatment (Fox, 2011). Several forums such as Biocreative (Wei et al., 2015), ImageCLEFMed (Müller et al., 2012) or CLEF eHealth (Palotti et al., 2015) have attracted the attention of Natural Language Processing (NLP) researchers (Friedman, Rindflesch, and Corn, 2013). A number of different NLP tasks have been studied in the health domain, from question answering (Lee et al., 2006) to multimodal information retrieval (Martín-Valdivia et al., 2008). Moreover, a number of techniques have been applied to improve the different systems, from basic Machine Learning (ML) algorithms (Chapman et al., 2011) to knowledge integration from medical ontologies (Díaz-Galiano, Martín-Valdivia, and Ureña López, 2009).

Nevertheless, in Sentiment Analysis (SA) it is very difficult to find out about research in the medical field. Although lately the development of SA methods and systems has vastly increased (Liu, 2012), the application of these technologies in the health domain is rather scarce. We can find some recent interesting work which mainly focuses on mining biomedical literature or processing medical web content (Denecke and Deng, 2015). However, most studies only deal with documents written in English, perhaps because platforms for expressing emotions, opinions or comments related to health issues are mainly oriented towards Anglophones. For example, PatientsLikeMe[1] is an online web platform that connect patients with one another, improving their outcomes and enabling research. Other example can be found at the website Patient Opinion[2], in which patients post their point of views after using a health service in United Kingdom, Ireland and Australia.

We consider this is clearly a topic of growing interest not only for people speaking English but also for people who speak a different language, such as Spanish. Our main goal is to launch research in the health domain by mining Spanish patient opinions extracted from the medical web Masquemedicos[3]. Actually, a 2012 survey of the Spanish population estimated that 29.9% of adults have looked online for information about health topics[4]. Nowadays, this number has increased exponentially. According to a study in 2015, 62% of the Spanish people consult Internet to be informed about topics related to the health[5].

In this paper we present the first Corpus Of Patient Opinions in Spanish (COPOS). In addition, we assess the validity of COPOS in implementing two basic polarity classification systems: one based on Semantic Orientation (SO) and another based on ML.

The present paper is structured on the following way: Section 2 describes briefly other studies related to the medical domain. In Section 3 the different resources used and the methodology employed to generate the corpus of opinions of patients are explained. Section 4 shows the experiments carried out and the discussion of the results obtained. Finally, conclusions and future work are presented.

2 Background

As we already stated, research in medical SA is very limited, although we can find some preliminary papers. Perhaps one of the first approaches is that presented by Niu et al. (2005). They manually annotate a corpus of medical abstracts extracted from MEDLINE (1,509 sentences). Then they apply ML (SVM) to classify the polarity of the sentences and the final result is about 79% in recall and precision. Sarker, Molla, and Paris (2011) follow a similar approach but they study the polarity classification at document level over another manually annotated corpus of 520 documents with a total of 9,221 sentences. The system also deals with the detection of negation cues. A comparison with several ML algorithms including SVM, Naïve Bayes, Bayes Net and C4.5 Decision Tree, was carried out and the results are near to 75% in accuracy. Chew and Eysenbach (2010) focus on extracting a corpus from Twitter containing references to the pandemic H1N1 and classify tweets into 16 different categories of opinions and sentiments. Bobicev et al. (2012) also build a corpus of tweets containing Personal Health Information (PHI). They manually annotate the corpus in order to apply ML algorithms to

[1] https://www.patientslikeme.com/

[2] https://www.patientopinion.org.uk/

[3] http://masquemedicos.com/

[4] http://www.ontsi.red.es/ontsi/sites/default/files/informe_ciudadanos_esanidad.pdf

[5] http://insights.doctoralia.es/informe-doctoralia-sobre-salud-e-internet-2015/

classify into positive, negative or neutral the sentiments expressed in the tweets. A similar ML methodology was presented in (Sokolova and Bobicev, 2013), but in this case the corpus used was extracted from a medical forum with messages related to In Vitro Fertilization (IVF). The documents were classified into 5 classes: encouragement, gratitude, confusion, facts, and facts+sentiments. A very interesting point in this paper is the generation of a specific lexicon for the health domain, the HealthAffect Lexicon (HAL). Authors show that the results obtained using HAL are better than applying other general lexicons and features. In a later paper, Bobicev, Sokolova, and Oakes (2015) continue studying the effect of applying HAL over the IVF medical forum, but in this case they focus on analysing sequences of sentiments in online discussions instead of considering only individual posts. This represents a more difficult task oriented towards discourse analysis.

Greaves et al. (2013) apply ML techniques to classifying opinions from patients related to their experience in a hospital of the English National Health Service. They collect a total of 6,412 online comments from patients, also rating the opinions using a scale from 0-5 points. The main goal of the authors was to predict automatically from the textual information in the comment whether the patient would recommend a hospital, whether the hospital was clean and whether he/she was treated with dignity. Another interesting study (Deng, Stoehr, and Denecke, 2014) compares and analyses different lexical and linguistic features in medical documents with sentiments and subjective non-medical texts. The aim is to study the applicability of typical SA methods in clinical narratives. The main conclusion of this study is that a simple method of SA is not suitable for analysing sentiment in clinical documents. Finally, we can find a very good literature review of SA for the medical domain in (Denecke and Deng, 2015).

All the described studies only deal with English documents (Personal Health Information in records or tweets, opinions in blogs or forums). In this paper we focus on Spanish SA in the medical domain. To the best of our knowledge, this is the first attempt to deal with opinions written by patients in Spanish about medical attention. Our approach is similar to the work of (Greaves et al., 2013), but oriented towards applying approaches of SA in the medical forum Masquemedicos. In this site people, mainly without technical or medical knowledge, post opinions and give a ranking for medical entities based on their own experience. Our approach not only applies ML in order to evaluate the viability of our corpus, but we also present a Semantic Orientation method for determining the polarity of the opinions.

3 *COPOS: Corpus Of Patient Opinions in Spanish* [6]

Due to the growing interest in online patient reviews, we have tried to find a forum or website where opinions are extracted from patients in order to analyze them. Within the medical domain, we have focused on opinions of patients about physicians who they have visited. In choosing the source of information from which to extract the corpus, the following factors were taken into account:

- There must be a reasonable number of opinions and these must be written by patients.
- Each opinion must be assessed by the owner of that opinion.
- The web portal should be a reliable portal in the domain of medicine.
- It must be an internationally prestigious site in search of information about medical entities.

In order to find a source that met all these requirements we conducted an exhaustive study, exploring all possible medical forums containing relevant patient opinions. This task was not easy because there are not too many web sites of patient opinions written in Spanish.

After studying some web sites, our final choice was the medical forum Masquemedicos. The generated corpus is a collection of patient opinions about medical entities that come from six countries (Chile, Colombia, Ecuador, Spain, Mexico, Venezuela). This forum only contains a maximum of 100 opinions per speciality. However, most of the specialities have less than 100 opinions. Moreover, we discarded those opinions that have some empty field. Therefore, the corpus is

[6] http://sinai.ujaen.es/copos-2/

composed of 743 reviews about 34 medical specialities taken on December 3, 2015. Each review contains information about the patient, the medical entity and the textual opinion. About the patient, we obtain his user name or Anonymous (in case of the patient does not show his identity) and his evaluation about the medical entity tagged with stars. In relation to the medical entity, the name and the speciality of the doctor, clinical or hospital are extracted with the city where the consultation was performed. Finally, the textual opinion is composed of positive and negative text parts and the date when the opinion was written. An example of a review of this medical forum can be found in Figure 1.

The reviews are rated on a scale from 0 to 5 stars. A value of 0 means that the patient expresses a very negative opinion about the medical entity, while a score of 5 means that the author has a very good opinion. The number of reviews per rating is shown in Table 1.

Rating	**#Reviews**
0	3
1	88
2	18
3	35
4	51
5	548
Total	**743**

Table 1: Distribution of reviews per rating.

Table 2 shows some interesting features of the corpus. It can be noted that the opinions have an average of 3 sentences, 44 words, 4 adjectives, 3 adverbs, 8 verbs and 10 nouns. We can check that the corpus is completely unbalanced with a portion of positive opinions much higher than negative ones. We have now retrieved all the reviews provided in the Masquemedicos forum. Thus, it seems patients are more interested in writing good comments than bad opinions.

4 Polarity classification with COPOS

Polarity classification is one of the most widely studied tasks in SA. This task aims to determine the category of opinion that can be assigned to a text. The category can be

	Positive	**Negative**	**Total**
#Reviews	634	109	743
#Sentences	1,603	406	2,009
#Words	24,244	8,121	32,365
#Adjectives	2,408	594	3,002
#Adverbs	1,695	587	2,282

Table 2: Statistics of COPOS corpus.

binary (positive, negative) or otherwise, and it can be made up of different levels of intensity. In our experiments we consider a binary classification of the reviews of the COPOS corpus (Table 3). In this way, opinions are classified as positive if they have 3, 4 or 5 stars, and negative if their rating is of 0, 1 or 2 stars.

Classes	**#Reviews**
Negative	109
Positive	634
Total	**743**

Table 3: Binary classification of COPOS corpus.

Although different approaches have been applied by the research community to tackling the polarity classification task, the mainstream basically consists of two major methodologies: On the one hand, the supervised or ML approach is based on using a dataset to train the classifiers (Pang, Lee, and Vaithyanathan, 2002). On the other hand, the approach based on computing the Semantic Orientation (SO) of the words in the documents does not need prior training, but takes into account the orientation (positive or negative) of words (Turney, 2002). This method is also known as the unsupervised approach. Both methodologies have their advantages and drawbacks. For example, the ML approach depends on the availability of annotated collections of data (training data), and in many cases this is difficult to achieve. On the contrary, a huge amount of lexical resources like lists of opinion words, lexicons or dictionaries, often with dependency on the language, are required by the SO approach. In this paper we present experiments at the document level based on these two methodologies over the COPOS corpus.

In order to tackle the supervised experi-

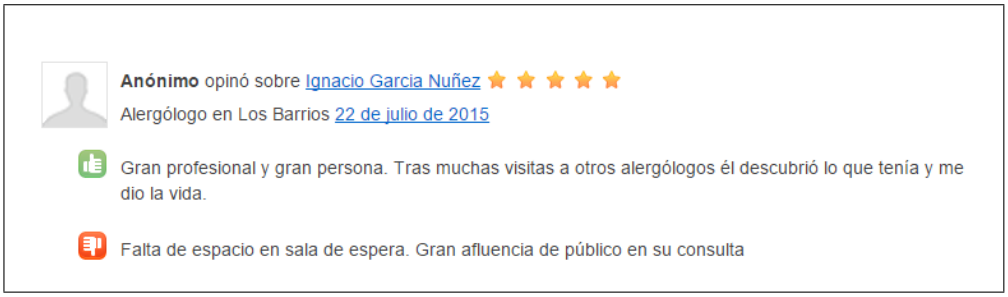


Figure 1: Example of an opinion on the Masquemedicos web portal.

ments the chosen algorithm is the Support Vector Machine (SVM) (Cortes and Vapnik, 1995) because it is one of the most successfully used in Opinion Mining (OM). On the other hand, in the unsupervised experiments we used the improved Spanish Opinion Lexicon (iSOL) (Molina-González et al., 2013), a well-known resource in the SA Spanish community. In order to calculate the polarity (p) of a review (r) with this lexicon, we took into account the total number of positive words (#positive) and the total number of negative words (#negative) within the review, according to the following strategy:

$$p(r) = \begin{cases} 1 & \text{if } \#positive \geq \#negatives \\ -1 & \text{if } \#positive < \#negative \end{cases} \quad (1)$$

where $p(r)$ is the polarity of the review r.

We used the typical evaluation measures employed in text classification: Accuracy (Acc.) and the macro-averaged version of the measures Precision (P), Recall (R) and F1.

4.1 Machine Learning Approach

In this paper we choose the data mining system RapidMiner as a tool for classifying the polarity in the COPOS corpus. The COPOS documents were preprocessed with different combinations of stopper and stemmer in each of the experiments, but in all of them the capital letters were changed to non-capital letters. In order to carry out the supervised approach, the SVM algorithm was applied. The selected SVM algorithm, broadly known by the research community in NLP, uses a linear kernel and normalizes the feature vectors. In order to represent the document we used the TF-IDF weighting scheme. After calculating the features of the documents, a 10-fold cross-validation framework was applied in order to assess the performance of the classifier. The results obtained are shown in Table 4.

Afterward we extended the supervised experiments using a balanced version of the COPOS corpus, formed by 109 negative reviews and 109 positive reviews randomly selected, and the results are shown in Table 5.

4.2 Semantic Orientation approach

The experiments based on Semantic Orientation employed a lexicon-based method. This method involves finding out the presence of opinion words of the lexicon in the documents. As we have mentioned before, the chosen lexicon is iSOL. It is a lexical resource increasingly used for the polarity classification of Spanish reviews. Before carrying out the experiments we performed a preprocessing step on the COPOS corpus in order to apply the same criteria followed during the generation of the iSOL list. For example, for each review we changed capital letters for non-capital letters, accented letters for non-accented letters, and all special characters were deleted from the opinions. Moreover, stop words were discarded.

We first carried out experiments over the original COPOS corpus, classifying a total of 743 reviews. In addition, we also applied our SO approach to the balanced version of COPOS with 218 opinions. Table 6 shows the results achieved by our SO system.

4.3 Result analysis

We consider that the results obtained with the SO approach over the original COPOS corpus are very good, especially taking into account that the results with SVM are similar. In fact, the improvement achieved over the accuracy with the best ML approach is only 0.46%. Bearing in mind that iSOL is a general purpose lexicon, we think that the adaptation of iSOL to the medical domain could achieve very promising results.

With respect to the experiments with the

Stopper	Stemmer	Precision	Recall	F1	Accuracy
YES	YES	78.43%	55.98%	65.33%	86.40%
YES	NO	52.77%	50.83%	51.78%	85.46%
NO	YES	84.17%	**61.40%**	71.00%	**87.88%**
NO	NO	**91.27%**	58.57%	**71.35%**	87.61%

Table 4: Result of polarity classification using SVM over COPOS.

Stopper	Stemmer	Precision	Recall	F1	Accuracy
YES	YES	**90.03%**	**88.32%**	**89.17%**	**88.51%**
YES	NO	88.95%	87.95%	88.46%	88.05%
NO	YES	87.51%	86.09%	86.79%	86.21%
NO	NO	88.23%	86.55%	87.38%	86.69%

Table 5: Result of polarity classification using SVM over balanced COPOS.

COPOS	Precision	Recall	F1	Accuracy
Original	75.28%	70.25%	72.31%	87.48%
Balanced	77.70%	70.66%	74.01%	70.18%

Table 6: Results obtained with iSOL Lexicon over COPOS corpus.

ML method, we first carried out the classification of the reviews with the whole corpus, but due to the highly unbalanced number of opinions for each class (approximately 85% of the opinions are positive and only 15% are negative), we decided to perform a new experiment with a balanced version of the COPOS corpus in order to avoid bias in the classification. However, it is interesting to point out that the results obtained with the balanced corpus are not much better than those with the unbalanced one. It should be noted that with the unbalanced corpus, when a stopper is applied the results are worse while when a stemmer is used the results achieved are better than when it is not applied. On the other hand, when the corpus is balanced the best result is obtained when stemmer and stopper are applied. Experiments conducted on other domains over balanced corpus, such as the movie domain, also obtain the same conclusion (Martínez Cámara et al., 2011).

Regarding the two approaches followed to classify the opinions of COPOS it is noteworthy the difference between the values of F1 and Accuracy measures when the original version of the corpus is used. The main reason for this difference is that Accuracy is a measure that may be biased by the majority class of a dataset. As it was mentioned before, COPOS is a unbalance corpus whose majority class is Positive, which is composed of 634 documents, meanwhile there are 109 negative reviews. Besides, if a classifier reaches a good performance over the majority class of the dataset, there is a higher likelihood that the Accuracy value will be biased. Table 7 shows the confusion matrix of the best configuration when the original version of COPOS is used (Stopper: No; Stemmer: Yes), in which the Precision and the Recall values of the class Positive are very high, meanwhile the performance of the class Negative is not remarkable. Therefore, the difference between the Accuracy and the F1 values are due to the unbalanced nature of COPOS.

5 Conclusions and future work

To the best of our knowledge, we have presented the first Corpus Of Patient Opinions in Spanish (COPOS). In order to demonstrate the usefulness of the corpus, we have carried out experiments with the main methodologies employed in the task of polarity classification (Semantic Orientation and Machine Learning). The results achieved and the growing interest of users in knowing opin-

	True Positive	True Negative	Class Precission
Prediction 1	627	83	88.31%
Prediction -1	7	26	78.79%
Recall	98.90%	23.85%	

Table 7: Confusion matrix of the ML experiment SVM (Stopper: No; Stemmer: Yes) with the original version of COPOS.

ions in the medical domain encourages us to follow this line of research.

Regarding our future work, we plan to extract more patient opinions due to the fact that the corpus presented here is unbalanced. This is an arduous task because in Spanish there are few reliable medical forums with patient opinions. On the other hand, we consider that the adaptation of the general purpose lexicon iSOL to the medical domain would be very interesting research and could greatly improve the final result. Finally, the integration of external medical knowledge, for example extracted from SNOMED, should be investigated.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) from the Spanish Government and by a Grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983).

References

Bobicev, V., M. Sokolova, Y. Jafer, and D. Schramm. 2012. Learning sentiments from tweets with personal health information. In *Proceedings of the 25th Canadian Conference on Advances in Artificial Intelligence*, Canadian AI'12, pages 37–48, Berlin, Heidelberg. Springer-Verlag.

Bobicev, V., M. Sokolova, and M. Oakes. 2015. What goes around comes around: Learning sentiments in online medical forums. *Cognitive Computation*, 7(5):609–621.

Chapman, W. W., P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Chew, C. and G. Eysenbach. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118, 11.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

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

Deng, Y., M. Stoehr, and K. Denecke. 2014. Retrieving attitudes: Sentiment analysis from clinical narratives. pages 12–15.

Díaz-Galiano, M. C., M. Martín-Valdivia, and L. A. Ureña López. 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Comput. Biol. Med.*, 39(4):396–403, April.

Fox, S. 2011. The social life of health information, 2011. Technical report, PewResearchCenter, May.

Friedman, C., T. C. Rindflesch, and M. Corn. 2013. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765 – 773.

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 15.

Lee, M., J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu. 2006. Beyond information retrieval — medical question answering. In *Proceedings of*

the AMIA Annual Symposium, pages 469–473.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Martín-Valdivia, M. T., M. C. Díaz-Galiano, A. Montejo-Raez, and L. A. Ureña López. 2008. Using information gain to improve multi-modal information retrieval systems. *Inf. Process. Manage.*, 44(3):1146–1158, May.

Martínez Cámara, E., M. T. Martín Valdivia, J. M. Perea Ortega, and L. A. Ureña López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47(0):163–170.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Syst. Appl.*, 40(18):7250–7257.

Müller, H., A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, and I. Eggel. 2012. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September.

Niu, Y., X. Zhu, J. Li, and G. Hirst. 2005. Analysis of polarity information in medical text. In *AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005*.

Palotti, J. R. M., G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. F. Jones, M. Lupu, and P. Pecina. 2015. CLEF ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse*, September.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volume 10 of *EMNLP '02*, pages 79–86. ACL.

Sarker, A., D. Molla, and C. Paris. 2011. Outcome polarity identification of medical papers. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 105–114, Canberra, Australia, December.

Sokolova, M. and V. Bobicev. 2013. What sentiments can be found in medical forums? In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 633–639, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. ACL.

Wei, C.-H., Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu. 2015. Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 91-98 recibido 29-03-2016 revisado 20-04-2016 aceptado 02-02-2016

Universal Dependencies for the AnCora treebanks

Dependencias Universales para los treebanks AnCora

Héctor Martínez Alonso
Univ. Paris Diderot
Sorbonne Paris Cité,
Alpage (INRIA), France
hector.martinez-alonso@inria.fr

Daniel Zeman
Charles University in Prague
Faculty of Mathematics and Physics
Czech Republic
zeman@ufal.mff.cuni.cz

Resumen: Este artículo presenta la conversión de los treebanks AnCora del catalán y el castellano al formalismo de Dependencias Universales (UD). Describimos el proceso de conversión y estimamos la calidad de los treebanks resultantes en términos de sus resultados en análisis sintáctico automático en un esquema monolingüe, en un esquema trans-lingüístico y en un tercero trans-dominio. Los treebanks convertidos muestran un nivel de consistencia interna de anotación comparable a la de los datos originales de la distribución CoNLL09 de AnCora, e indican algunas diferencias en terminos del inventario de expresiones polilexemáticas con respecto al anterior treebank del castellano en UD. Los dos nuevos treebanks convertidos serán distribuidos con la versión 1.3 de Dependencias Universales.
Palabras clave: AnCora, treebank, catalán, castellano, Dependencias Universales

Abstract: The present article describes the conversion of the Catalan and Spanish AnCora treebanks to the Universal Dependencies formalism. We describe the conversion process and assess the quality of the resulting treebank in terms of parsing accuracy by means of monolingual, cross-lingual and cross-domain parsing evaluation. The converted treebanks show an internal consistency comparable to the one shown by the original CoNLL09 distribution of AnCora, and indicate some differences in terms of multiword expression inventory with regards to the already existing UD Spanish treebank. The two new converted treebanks will be released in version 1.3 of Universal Dependencies.
Keywords: AnCora, treebank, Catalan, Spanish, Universal Dependencies

1 Introduction

AnCora treebanks[1] (Taulé, Martí, y Recasens, 2008) are consolidated treebanks for Catalan and Spanish, and have indeed been the canonical treebanks of these languages. Their smaller, preliminary versions were used in the CoNLL 2006 and 2007 shared tasks in dependency parsing; the much larger and mature AnCora 2.0.0 was used in CoNLL 2009 (Hajič et al., 2009), henceforth CoNLL09. The native AnCora syntactic annotation is based on constituents but an in-house conversion to dependencies is also available (Civit, Martí, y Bufí, 2006).

In this article we present the conversion of the AnCora treebanks to Universal Dependencies. There is a UD Spanish treebank since release 1.0 (January 2015). This corpus is a legacy of an older universal treebank project (McDonald et al., 2013) and it is unrelated to AnCora. It is made up of web data and we refer to it as the Spanish Web UD treebank. However, according to the Web treebank documentation, its developers have made use of AnCora for lemmatization and morphological analysis. With our UD conversion of AnCora, appearing in UD release 1.3 (May 2016), the UD collection thus contains two treebanks for Spanish, as well as the first UD treebank for Catalan. Independent from our work, a Galician treebank is also scheduled to appear in UD 1.3. If we also consider the already existing Basque and Portuguese treebanks, UD 1.3 will allow parsing a great deal of the linguistic diversity of the Iberian peninsula.

Section 1.1 outlines the important characteristics of the UD formalism that we have taken into account for the conversion. Section 2 describes the conversion steps. Sec-

[1] http://clic.ub.edu/corpus/

ISSN 1135-5948
© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

tion 3 offers quantitative evaluation of the conversion results, and finally Section 4 offers conclusions and perspectives.

1.1 Universal dependencies

Universal Dependencies (UD)[2] (Nivre et al., 2016) is a project that seeks to define cross-linguistically applicable annotation guidelines for morphology and syntax of natural languages. An integral part of this effort is frequent releases of annotated data (UD treebanks) in multiple languages, that conform to the UD guidelines and that are freely available to the research community.

UD uses a set of 17 universal POS, a set of 40 universal dependency relations, and a set of universal features to give account for lexical or grammatical information in terms of key-value pairs like *Gender=Fem*. The POS inventory is fixed across all languages—even though not all languages use all tags—whereas the formalism allows language-specific extensions of dependency relations and features.

A key aspect of the UD dependency formalism is the primacy of content over function words. While other conventions make auxiliaries the head of periphrastic verb tenses, or even determiners the head of noun phrases, content words are the preferred heads in UD. Notably, this analysis also demotes copula verbs to dependent status, and makes the attribute the head of the copula construction. This decision aims at harmonizing the representation across languages, including those that have no explicit copula. Figure 1 shows an example sentence from the original Spanish AnCora CoNLL09 corpus and its UD conversion. We describe the example in more detail in Section 2.5.

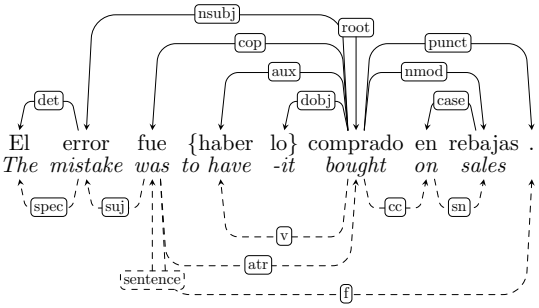


Figure 1: Example of dependency tree before and after conversion. Above: UD, below: CoNLL09 (dashed).

The main contribution of this conversion is to make the AnCora treebanks available for further research using the UD formalism, as they are treebanks that have been benchmarked for a decade. During conversion, we make the treebanks compliant with the specifications of the UD formalism, and we harmonize certain choices to increase compatibility with the other Romance languages, in particular aiming at making the choices of structure and part of speech as similar as possible between the already existing UD Spanish Web treebank and our conversion of Spanish AnCora.

2 Conversion

Recent work (Kolz, Badia, y Saurí, 2014) describes a dependency conversion of the Spanish AnCora treebank. However, it used a syntax-driven formalism, where function words are more likely to be the heads. In this present work we take the complementary stance.

Moreover, the AnCora treebanks have already appeared in another multi-lingual treebank collection, namely HamleDT (Zeman et al., 2014), which is one of the predecessors of UD.[3] HamleDT provides an automatic conversion of over 40 treebanks by first converting them all to the formalism of the Prague Dependency Treebank, and later exporting them to either the Stanford dependency formalism in the first releases of HamleDT, or to UD in the current release 3.0 (Zeman et al., 2015). However, HamleDT 3.0 does not follow some important UD guidelines, most notably those concerning tokenization. We take HamleDT as our starting point and extend the conversion to produce a fully UD-compliant release of the sibling corpora.

2.1 HamleDT conversion outline

We make use of the freely available HamletDT 3.0 conversion to incorporate the AnCora treebanks to UD. This section provides an overview of the main operations carried out during the HamletDT conversion. For further details, cf. (Zeman et al., 2014). First, HamleDT transforms the original CoNLL09 treebanks to Prague Dependencies (Böhmová et al., 2003) following these steps:

[2] `http://universaldependencies.org/`

[3] `http://ufal.mff.cuni.cz/hamledt/`

1. Convert POS and features from CoNLL09 to UD.[4]
2. Convert the CoNLL dependency relation labels to those of the Prague Dependency Treebank. Some labels are further adjusted when the tree structure is transformed. This step is not trivial because some "dependency" relations in fact just denote leaf nodes from the original constituents, whose (dependency) relation to the constituent head is not marked. Conversion of these cases often requires also transforming the tree structure in subsequent steps.
3. The original annotation has no explicit marking of coordination. Some dependency relations such as `grup.nom` are good indicators of coordination but there are structures such as coordinated clauses, that cannot be detected this way. HamleDT searchers for larger-scope coordinated constituents joined by coordinating conjunctions, and marks them as coordination.
4. Some phrases present idiosyncratic head choices, e.g. in the Catalan *una mica* ('a bit'), the noun *mica* is attached as a dependent of the article *una*. This analysis does not correspond with the Prague guidelines, and it is also inconsistent with how determiners and nouns are connected elsewhere in the treebank. This particular analysis in CoNLL09 is in fact similar to the UD convention for multiword expressions (cf. Section 2.3).

The main steps of the conversion from Prague to Universal Dependencies are:

1. Convert Prague dependency relation labels to UD relations.
2. Convert coordination from the Prague style (headed by conjunction) to the Stanford style (headed by the first conjunct). This effectively means reverting to the approach taken in the original CoNLL09 data, except that now all coordination is explicitly and uniformly marked. For more details on the complexities of coordination conversion and the function-structure tradeoff in dependency syntax, cf. (Popel et al., 2013).
3. Invert prepositional phrases so that the nominal is the head and the preposition is attached to it using the `case` relation.
4. Invert copula constructions so that the attribute (adjective or nominal predicate) is the head and the copula is attached to it using the `cop` relation. The subject and adverbial modifiers are also re-attached to the attribute.
5. Detect controlled verbs and treat them as non-finite subordinate clauses, i.e. infinitives attached to other verbs, e.g. in *va refusar donar més detalls* ('refused to elaborate'), *donar* is attached to *refusar* with the relation `xcomp`.

[4] http://universaldependencies.org/tagset-conversion/index.html

2.2 Tokenization

Tokenization makes up the most of our adaptation from HamleDT 3.0 to full UD compliance. Notice that tokenization changes in an already-annotated treebank are not trivial, because they also imply rewriting the dependency structure of the sentences with modified tokens. We also implement other operations like feature and empty-sentence cleanup. The UD stance on tokenization is that dependency relations hold between *syntactic words*, which do not have to be identical with *orthographic words*. Surface tokens are split if their parts perform independent syntactic functions. For instance, *del* is the preposition *de* fused with the definite article *el*; in UD, the preposition and the article are independent syntactic words corresponding to separate nodes in the dependency tree.

On the other hand, UD does not allow tokens to be made up of more than one orthographic word, i.e. "words with spaces" are disallowed. If a frozen expression of multiple orthographic words behaves as one syntactic unit (where the internal syntactic structure does not exist or has become vacant), each orthographic word will have its own node in the tree and technical relations such as `mwe` (for multiword expressions) or `name` (for proper names) will connect them. The head of these two kinds of structures is the leftmost token, and all other tokens are its dependents. Figure 2 provides an example of `mwe` in UD. The expression *pel que fa a* (lit. 'for that which does to', En. 'regarding') is a single token in the CoNLL09 data joined with underscores, and becomes a `mwe` subtree.

A key difference between `mwe` and `name` subtrees is that UD guidelines treat prepositions, articles and conjunctions in proper names as such, and not as tokens with the `PROPN` POS tag. This difference implies that

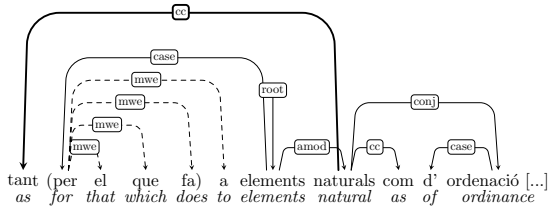


Figure 2: UD Catalan example with a multiword expression subtree (dashed) and a nonprojectivity caused by coordination (thick edge) crossing another edge.

`name` subtrees like *Ajuntament* ***de*** *Vilanova* ***i la*** *Geltrú* will have more than one depth level where the words in bold will be leaves, whereas the depth of `mwe` is always one.

The CoNLL09 treebanks make elliptic subjects explicit by giving them an empty token with pronoun POS and subject function, marked with an underscore as form. Elliptic subjects are not syntactic words, and we remove them from the sentences.

Moreover, UD allows providing the original tokens of certain constructions prior to being tokenized apart into syntactic words. These *multiword tokens* are provided as an additional annotation of the sentence and play no role in the syntax, but can be used for ease of reconstruction of the original text, to train tokenizers, etc. We apply language-specific modifications to identify two kinds of multiword tokens in Catalan and Spanish, namely fused article-preposition tokens, and verbs with clitics.

2.3 Catalan-specific tokenization

The Catalan inventory of fused adposition-articles is *al*, *del*, *pel*, and their plural forms *als*, *dels*, and *pels*, for a total of six. We split these tokens into their two forming syntactic words, and provide the original token as a multiword token.

Clitic tokenization in Catalan is simple because clitics are introduced by hyphens or apostrophes. Verbs with clitic pronouns are already tokenized apart in AnCora-Catalan. We identify verb-headed spans with clitics and add the multiword-token to the annotations of the sentence. The most common possessive determiner construction is a two-word periphrasis e.g. ***la meva*** *casa* (‘**my** house’), where the first word is the definite determiner (*la*, ‘the’) and the second one is the possessive adjective (*meva*, ‘mine’), placed prenominally and consecutive to the determiner. The original AnCora tokenization treats these two words together as a one, joined with underscores. We split these pretokenized possessive constructions into their two forming syntactic words, and make them both dependent of the noun they introduce. We inform of the possessiveness of the second word using the `Poss` feature.

2.4 Spanish-specific tokenization

The Spanish inventory of fused adposition-articles is a set of two, namely *al* and *del*. We split them and keep the original fused for as a multiword token.

AnCora-Spanish does not provide split verbs with clitics like *encontrándoselas* or *abridlo*, cf. Table 1. We split away the clitics for these verbs and insert them as pronouns, with a POS tag PRON and the corresponding case, gender and number features, making them dependents of the verb. Moreover, we normalize the spelling of the verb form without clitics by removing diacritics. e.g. *encontr**á**ndoselas* vs. *encontrando*. This change in spelling is a consequence of the change of stress pattern when removing clitics from the verb. We add the original multiword token verb as a sentence annotation.

encontrándoselas		
encontrando	se	las
FIND-gerund	3-refl	3-f-p-a
abridlo		
abrid	lo	
OPEN-imperative-plur	3-m-s-a	

Table 1: Spanish clitic-verb examples.

2.5 Conversion example

Figure 1 shows the original AnCora (dashed, below) tree and the converted UD tree for the Spanish sentence *El error fue haberlo comprado en rebajas’* (En. ‘The mistake was to have bought it on sales’). We can observe how the names of the relations are all different, and the relations above use the UD inventory. In terms of the structure, the original main node *fue* is demoted to leaf status as a `cop`, a copula auxiliary of the predicate *comprado*, which is the main node in the UD tree. Note that *fue* does not license a passive reading, which would be marked as `auxpass`. The preposition *en* also becomes a leaf, in

this case of the noun *rebajas*. Moreover, the verb-clitic multiword *haberlo* ('have+it') is split in two different tokens, and *lo* is a `dobj` dependent of the verb *comprado*.

3 Results

3.1 Treebank properties

There is of course the danger that each new conversion will lose more information or introduce errors. We dedicate this section to determining the consistency of the treebanks after UD conversion. Table 2 shows the properties of the AnCora treebanks at their three distributions, namely CoNLL09, HamleDT 3.0 and UD 1.3. The columns list the number of sentences (S) and of words (W), the POS ambiguity (PA), the average edge length (EL), the average root distance (RD), and the number of sentences that are not fully projective (NP).

In terms of number of words and sentences, we keep the official test data split from the CoNLL 2009 shared task, where the training data is 80% of the corpus, and development and test data are 10% each. We can see that the number of sentences diminishes slightly as a result of empty-sentence removal, while the number of words increases in 10% as a consequence of splitting multiword expressions, fused preposition-articles, and verbs with clitics.

We measure POS ambiguity as the proportion of words that have more than one possible POS and a frequency above one. The conversion steps make PA increase. While an increase in ambiguity makes POS prediction harder using these datasets, it also indicates that the conversion process successfully applies a disambiguation of the original CoNLL09 POS inventory of 12 tags into the UD inventory of 17 tags by means of structure and feature analysis, notably incorporating the lexical verb / auxiliary verb distinction (`VERB`/`AUX`) and the common noun / proper name (`NOUN`/`PROPN`) distinction.

The average edge length (EL) increases monotonically on each conversion step, as relations across predicates are pulled up in the decision tree when e.g. a subordinate clause becomes headed by its verb and not by its subordinating conjunction. Parallel to this change, the average distance to the root node (RD) decreases because trees become flatter.

The CoNLL09 distribution of both AnCora treebanks is fully projective. However, the conversion steps incorporate non projectivities into the structure. Upon manual inspection, we find that some non-projectivities are introduced by the splitting process of large proper-name multiwords that are internally connected by determiners and prepositions such as *Les Terres de l'Ebre*. Moreover, we also find legitimate case of non-projectivity such as the Catalan example on Figure 2. The conjunction *tant* has scope over *naturals com d'ordenació*, and the intermediate word *elements*, higher in the tree, issues a crossing edge. The expression *tant ... com* is a double conjunction is a manner similar to 'as well ... as'.

3.2 Monolingual dependency parsing

Dependency parsing evaluation allows estimating the consistency of a treebank's annotations. We use TurboParser (Martins et al., 2010) for all the parsing experiments in this section. We have trained all the models using the local, arc-factored feature model for the parser. While a richer feature model would improve performance, we use the arc-factored model for speed reasons. The goal of the parsing experiments in the following sections is to assess the relative consistency of the different versions of the treebanks, and not to benchmark the parser itself. Nevertheless, arc-factored TurboParser obtains scores comparable to the best systems for Catalan and Spanish in the CoNLL09 shared task.[5]

Table 3 shows the parsing results for the three steps in the conversion of the treebanks, namely the original CoNLL09 AnCora distribution, the HamleDT 3.0 UD-compatible conversion, and our UD conversion.

These scores are not strictly comparable, given that all treebanks have different tokens, part-of-speech tags, and dependency relations. However, we provide them as an indication of the general reliability of the conversion. This approach has also been used in previous conversion works, who also report scores in the 80-85% range (Johannsen, Martínez Alonso, y Plank, ; Pyysalo et al., 2015; Silveira y Manning, 2015).

In spite of these differences, some relations are straightforward to compare. The `sentence` relation in CoNLL09 maps to the `root` relation in UD. Both the Catalan and

[5] `https://ufal.mff.cuni.cz/CoNLL09-st/results/results.php`

		S	W	PA	EL	RD	NP
Ca	CoNLL09	16,786	497k	0.10	3.66	4.60	0
	HamleDT	16,786	497k	0.13	3.86	4.07	76
	UD	16,678	547k	0.16	4.00	4.12	468
Es	CoNLL09	17,709	528k	0.11	3.69	4.76	0
	HamleDT	17,709	528k	0.13	3.90	4.22	327
	UD	17,680	569k	0.15	3.99	4.19	624

Table 2: Treebank statistics

	C09		HDT		UD	
	LAS	UAS	LAS	UAS	LAS	UAS
Ca	86.7	89.6	85.4	87.7	85.4	87.9
Es	86.1	88.9	85.0	87.1	84.9	87.3

Table 3: Labeled and Unlabeled Attachment Scores (LAS and UAs, in grey) for the three different conversion stages of the treebanks.

Spanish CoNLL09 treebanks have an average `sentence` accuracy of 93% and, and their UD conterparts have a `root` accuracy of 90%. This degradation is a result of i.a. the promotion of lexical tokens to the head of the copulas, which penalizes the general tendency to make verbs the heads of clauses.

Complementarily, the LAS of prepositions, tagged `s` in CoNLL09 and `ADP` in UD, goes from 79% to 97% after the conversion. This improvement is a result of preposition attachment becoming more local in UD, and easier to resolve, while noun attachment becomes more difficult to predict, and goes from 90% to 76%. Indeed, UD transfers most of the important relations to relations between content words, and makes function words easier to attach: auxiliaries, determiners and prepositions have all attachment accuracies above 96% in both UD AnCora treebanks.

3.3 Dependency parsing between UD languages

We use cross-lingual parsing as a way to estimate the consistency with the other treebanks, and thereby with the current state of UD as a whole. In order to do so, we apply a delexicalized transfer scenario, removing form and lemma information from the training data, and we train Turbo parser in unlabeled mode. While it is customary to also remove the feature information, we decide to keep all the morphological features of the source and target data, because they belong to the harmonized UD inventory.

Table 4 shows the results on training on Catalan or Spanish and testing on itself (*Self*), on the other converted AnCora treebank (*Sibling*), i.e. training on Catalan and testing on Spanish. The three last columns show the results on delexicalized parsing the test section of whole set of UD1.2 languages. The *All* column provides the macro-average for all 32 languages, while the *Romance* column is the macro-average score for French, Italian, Portuguese and Romanian, as well as the pre-existing Spanish Web treebank. The *Other* column shows the average results for all the non-Romance UD languages. We compare the Spanish Web and the Spanish AnCora treebank in more detail in Section 3.4.

	Self	Sib.	Rom	Other	All
Ca	84.5	81.7	66.2	49.5	52.0
Es	83.5	82.6	62.5	46.9	49.6

Table 4: mean UAS for delexicalized transfer.

The drop in UAS from full lexicalized to delexicalized is only of 5%. The high delexicalized parsing scores for Self indicate that, in spite of the chain of conversions, the AnCora UD treebanks have a well-coupled mapping between the POS tags and features, and dependency structure. The AnCora UD treebanks are internally very consistent, and the parser achieves comparatively high UAS when trained on one sibling and applied to the other. However, the results are 20 points lower in average when parsing the other Romance languages. For Italian and Portuguese the score is around 71% for both sources, while for French it is around 53%. The internal variation between the Romance language

test bench can have a linguistic basis, but it is also caused by divergences in the treebank's annotation. If we compare with the *Other* set, we observe the differences are even larger, and the drop from *Romance* to *Other* is of 13 points. Regardless of the variation in dependency annotation, there is more consistency between the new AnCora treebanks and the Romance languages, which confirms the linguistic basis for better parsing scores for e.g. Italian and Portuguese.

3.4 Dependency parsing between AnCora and Web Spanish

We assess the similarity of the Ancora Spanish conversion to UD with the preexisting Spanish Web UD using dependency parsing. If the two treebanks were as similar as possible, the differences in parsing accuracy when e.g. parsing with AnCora and testing with Web would be due to dataset size and domain change, and not to differences in dependency convention. Table 5 shows the attachment accuracies when using one Spanish treebank to parse the other (*Other*), along with intra-treebank evaluation for comparison (*Self*).

	Self		Other	
	LAS	UAS	LAS	UAS
AnCora	84.9	87.3	64.0	71.8
Web	81.7	84.5	69.6	78.7

Table 5: Labeled and Unlabeled Attachment Scores (LAS and UAs, in grey) for the four possible source/target pairs for Spanish.

There is a severe drop in performance when changing training treebank, e.g. when using Web to parse AnCora, the performance drops in about 15 points, from 84.9% to 69.6% LAS. A similar drop of 17 points appears when using AnCora to predict Web.

These differences are large enough not to be exclusively an effect of domain change. UD treebanks encode some lexical information such as multiword expressions in the dependency structure (cf. Section 2), and the inventory of multiword expressions is different across treebanks.

In AnCora UD, the elements marked as multiwords are the tokens that were underscored together in the CoNLL09 data, and that were not proper names. AnCora UD has 521 word types that are attached with a `mwe` relation, whereas Web has 113. While AnCora has a larger `mwe` inventory, it is not a perfect superset of the `mwe` expressions in Web, because it only covers 80% of the expressions in the Web corpus. Some of the common expressions are treated in the same fashion in both treebanks, such as *sin embargo*, *ni siquiera*, or *a través de* ('nevertheless', 'not even', 'through'), but the difference in multiword inventory make the parser predict mistmatching structures across datasets.

Both treebanks have a lenient definition on auxiliary verbs. Besides the auxiliares that are used to form verb tenses, namely *haber* for compound tenses, *ser* for passive forms, and *estar* for gerund forms, the treebanks also license an `AUX` reading for verbs that indicate modality *(poder, querer),* aspect *(continuar, terminar),* and also other verbs used for support constructions like *llegar* in *llegar a causar* ('come to cause'). These choices are semantically motivated and aim at promoting the lexical verb of the construction to the head of the structure (cf. Section 1.1, instead of treating light verbs as syntactic heads. However, Romance languages do not have a family of modal verbs as distributionally well-defined as Germanic languages do, and the criteria for labeling a semantically impoverished verb as `AUX` can be revised in further releases of the treebanks.

Another major difference between the two Spanish treebanks is the interpretation of the word *que* as either a subordinating conjunction or a pronoun (`SCONJ` / `PRON`), which gives a very low attachment score to pronouns (35%) in the cross-treebank (*Other*) setup, in spite of the very high accuracy in the intra-treebank setup (85%).

The AnCora Spanish corpus is largely made up of newswire from Spain, whereas the Web Spanish corpus potentially holds any of the variants of Spanish. We find sentences like *Olvidate todo, seguí tu vida* ('Forget about everything, get on with your life'), with verb usage characteristic of Rioplatense Spanish. A more detailed study on the differences between both Spanish corpora would shed light on the relevance of regional specificity in corpus choice for Spanish processing.

4 Conclusion

We have presented the conversion of the Catalan and Spanish AnCora treebanks to the Universal Dependencies formalism. We use the freely available HamleDT 3.0 as

a starting point for POS and dependency conversion, and we apply a set of operations to tune tokenization to UD, tackling language-specific phenomena like fused article-prepositions like *del* and verb-clitic multiword tokens like *encontrársela*.

We have evaluated the consistency of the resulting converted treebanks by means of dependency parsing. We obtain parsing scores comparable to other converted UD treebanks (cf. Section 3.2), and we assess the variation between UD languages in terms of typological proximity and annotation convention using delexicalized transfer parsing 3.3. The fairly large loss of unlabeled attachment score for the other languages is much less dramatic for the Romance languages.

We also analyze the differences between the Spanish Ancora UD-converted treebank and the preexisting Spanish Web treebank. While the parsing between the two Spanish UD treebanks fares above the delexicalized transfer setup, the differences in multiword treatment and some POS particularities indicate that the treebanks need further harmonization.

4.1 Further work

Universal Dependencies is a constantly improving effort, and the guidelines are refined before each release. In further releases, we expect to harmonize the two AnCora UD treebanks with regards to the treatment of auxiliary verbs across all Romance languages, revise the non-projective sentences and keep the legitimate examples, and in particular improve the comparability of the two UD Spanish treebanks, namely AnCora and Web.

Bibliografía

Böhmová, A., J. Hajič, E. Hajičová, y B. Hladká. 2003. The prague dependency treebank. En *Treebanks*.

Civit, M., M. A. Martí, y N. Bufí. 2006. Cat3LB and Cast3LB: from constituents to dependencies. En *Advances in Natural Language Processing*.

Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, y Y. Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. En *CoNLL-2009*.

Johannsen, A., H. Martínez Alonso, y B. Plank. Universal dependencies for Danish. En *TLT14*.

Kolz, B., T. Badia, y R. Saurí. 2014. From constituents to syntax-oriented dependencies. *Procesamiento del Lenguaje Natural*.

Martins, A. F., N. A. Smith, E. P. Xing, P. M. Aguiar, y M. A. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. En *EMNLP 2010*.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström, y others. 2013. Universal dependency annotation for multilingual parsing. En *ACL*.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, y D. Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. En *LREC*.

Popel, M., D. Marecek, J. Stepánek, D. Zeman, y Z. Zabokrtskỳ. 2013. Coordination structures in dependency treebanks. En *ACL*.

Pyysalo, S., J. Kanerva, A. Missilä, V. Laippala, y F. Ginter. 2015. Universal dependencies for Finnish. En *NoDaLiDa*.

Silveira, N. y C. Manning. 2015. Does universal dependencies need a parsing representation? an investigation of English. *Depling 2015*.

Taulé, M., M. A. Martí, y M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. En *(LREC 2008)*.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, y J. Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*.

Zeman, D., D. Mareček, J. Mašek, M. Popel, L. Ramasamy, R. Rosa, J. Štěpánek, y Z. Žabokrtský. 2015. HamleDT 3.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Análisis del Contenido Textual

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 101-108 recibido 03-04-2016 revisado 21-04-2016 aceptado 03-05-2016

Traducción Automática usando conocimiento semántico en un dominio restringido

Automatic translation using semantic knowledge in a restricted domain

Lluís-F. Hurtado, Iván Costa, Encarna Segarra, Fernando García-Granada, Emilio Sanchis
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de Vera s/n
lhurtado@dsic.upv.es

Resumen: El propósito que sigue este trabajo es incorporar conocimiento semántico a la traducción automática con el objetivo de mejorar la calidad de ésta en dominios restringidos. Nos centraremos en la traducción entre inglés, francés y español en el contexto de consultas telefónicas a un servicio de información ferroviaria. Se han desarrollado varias estrategias para la incorporación de la semántica en el proceso de traducción. Algunas de estas aproximaciones incorporan la semántica directamente en los elementos al ser traducidos, mientras que otras utilizan una interlingua o lengua pivote que representa la semántica. Todas estas aproximaciones han sido comparadas experimentalmente con una traducción automática basada en segmentos léxicos que no incorpora conocimiento semántico.
Palabras clave: traducción automática, semántica, dominios restringidos

Abstract: The purpose of this work is to add semantic knowledge to a machine translation process in order to improve its quality for restricted domains. We will focus on the translation between English, French and Spanish, in a task of telephonic query to a railway information service. Many strategies have been developed for the incorporation of semantics into the translation process. Some of these approaches directly incorporate semantics into the elements to be translated and some others use an interlingua or pivot language that represents the semantics. All of these approaches have been experimentally compared to an automatic translation based on lexical segments that do not incorporate semantic knowledge.
Keywords: automatic translation, semantics, restricted domains

1 Introducción

En traducción automática, así como en la mayor parte de aplicaciones del procesamiento automático del lenguaje natural, existen dos tipos de tareas diferentes según su ámbito de aplicación. El primero, el ámbito de propósito general, en el que el sistema ha sido entrenado de manera genérica, de forma que se pueden procesar frases de cualquier contexto. Por ejemplo, Europarl (Koehn, 2005) es un corpus que recoge sesiones del parlamento europeo en las que se ha hablado de distintos temas. El segundo ámbito representa un dominio restringido, referente a un sistema que ha sido entrenado para procesar frases dentro de un contexto concreto. Este es el caso del corpus DIHANA (Benedí et al., 2006), que contiene diálogos referentes a consultas sobre información ferroviaria.

Entre las distintas aproximaciones a la traducción automática, en este trabajo nos centraremos en la traducción estadística, también referida como SMT (Statistical Machine Translation). El objetivo de esta aproximación es obtener buenos traductores a partir de estadísticas obtenidas de un corpus. Para entrenar un traductor de un idioma origen (o fuente) a otro destino mediante un sistema de traducción automática estadística es necesario disponer de un corpus paralelo. La calidad de un sistema de traducción automática estadística depende de las características del corpus paralelo de entrenamiento disponible. Hay dos factores importantes: el tamaño del corpus paralelo de

* Este trabajo ha sido financiado por el MINECO y fondos FEDER en el proyecto TIN2014-54288-C4-3-R

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

entrenamiento y el dominio. Un conjunto de datos de entrenamiento pequeño conduce a modelos de traducción pobremente estimados y, en consecuencia, a una mala calidad en la traducción. Por esta razón, la calidad de la traducción, empeora cuando no tenemos suficientes datos de entrenamiento para el dominio específico objetivo.

Los métodos de adaptación al dominio se pueden dividir en dos amplias categorías. La adaptación al dominio puede hacerse en el corpus, por ejemplo, mediante la ponderación y la selección de los datos de entrenamiento (Gao et al., 2002). La adaptación también puede hacerse mediante la adaptación de los modelos de traducción. Algunos trabajos apuestan por la incorporación de conocimiento semántico para mejorar los resultados de la traducción, por ejemplo en (Banchs y Costa-jussà, 2011) se propone el uso de ciertas características semánticas para la SMT basada en el uso de Latent Semantic Indexing.

También se obtienen malos resultados cuando se trabaja con idiomas para los que no se dispone de recursos suficientes. Para estos casos se ha propuesto el uso de una lengua pivote o interlingua como paso intermedio entre la traducción entre dos idiomas. En (Habash y Hu, 2009) se propone el uso del inglés como lengua pivote entre árabe y chino, en (Babych, Hartley, y Sharoff, 2007) se propone el uso de una lengua pivote cercana a la lengua fuente y se reporta una comparación con la traducción directa.

El objetivo de este trabajo es incorporar conocimiento semántico al proceso de traducción automática para mejorar su calidad en dominios restringidos. La incorporación de este conocimiento se lleva a cabo siguiendo dos estrategias: o bién añadimos etiquetas semánticas a las palabras o secuencias de palabras en el lenguaje fuente o destino, o bién utilizamos la representación semántica como lengua pivote para la traducción. En este trabajo se exploran diferentes etiquetados semánticos y se compararan los resultados con los correspondientes a la traducción automática sin conocimiento semántico usando el corpus DIHANA, que contiene frases procedentes de diálogos referentes a consultas sobre información ferroviaria.

Las diferentes estrategias empleadas para incorporar este conocimiento semántico son: uso de la representación semántica mediante etiquetas semánticas asociadas a cada palabra y el uso de una lengua pivote basada en la representación semántica.

Para estimar el traductor automático estadístico se ha usado el conjunto de software MOSES (Koehn y et al., 2007). En este trabajo se han seleccionado IRSTLM (Federico, Bertoldi, y Cettolo, 2008) para la generación de modelos de lenguaje, y GIZA++ (Och y Ney, 2003) para la generación de alineamientos entre frases en distintos idiomas como herramientas externas. Como resultado de la fase de entrenamiento se han estimado diferentes modelos de traducción, dependiendo de la calidad de los alineamientos entre frases que se obtuvieron mediante el software alineador. Posteriormente estos modelos se han mejorado haciendo uso de un conjunto de Desarrollo.

Para la evaluación de los resultados utilizaremos la métrica BLEU. El BLEU se calcula dependiendo de la precisión de n-gramas indicada, habitualmente se emplea una precisión de 4, y a tal métrica se la llama BLEU-4 (Koehn, 2010). Es importante saber que BLEU no tiene en cuenta la relevancia de las palabras que son sinónimas, por lo que sólo cuenta un acierto cuando las palabras que compara son iguales.

2 Descripción del corpus

El corpus utilizado en este trabajo ha sido el corpus DIHANA. Se compone de 900 diálogos en español en el ámbito de llamadas telefónicas realizadas a un servicio de información de trenes. Todos estos diálogos fueron adquiridos mediante la técnica del Mago de Oz. Para disponer de un corpus paralelo se han traducido esos mismos diálogos a francés y a inglés. El corpus se ha dividido en tres conjuntos: Entrenamiento, Desarrollo (para ajuste de parámetros) y Prueba. El conjunto de Entrenamiento se ha traducido a francés e inglés mediante traductores web de propósito general sin supervisión. Se optó por emplear cuatro traductores diferentes para realizar cada una de las traducciones, de manera que se mitigaran los errores de traducción que podía haber producido cada uno de ellos individualmente. Estos traductores web son: Google Translate, Bing, Lucy y OpenTrad. En el caso de las traducciones de las frases que conforman los corpus de Desarrollo y Prueba para inglés y para francés estas traducciones fueron hechas por hablantes nativos de ambas lenguas.

El corpus en español está etiquetado desde el punto de vista semántico. Todas las frases del corpus están segmentadas y cada segmento tienen asociado una etiqueta semántica asignada de forma manual, además toda la frase tiene asociada una representación semántica en términos de Frames (Tabla 1).

Frase
hola quiero saber el horario de ida de Palencia a Oviedo el viernes dieciocho de junio
Conceptos Semánticos
hola:cortesia quiero saber:consulta el horario de:<hora> ida:tipo_viaje de Palencia:ciudad_origen a Oviedo:ciudad_destino el viernes dieciocho de junio:fecha
Frame
TIPO-VIAJE:ida (HORA) CIUDAD-ORIGEN:Palencia CIUDAD-DESTINO:Oviedo FECHA:[viernes-18-06-??,viernes-18-06-??]

Tabla 1: Ejemplo frase corpus DIHANA

El conjunto original en castellano contiene 6226 frases, un total de 47186 palabras y un vocabulario compuesto de 719 palabras diferentes. A partir del corpus original se han definido los conjuntos de Entrenamiento, Desarrollo y Prueba. Para los corpus en inglés y francés se han utilizado las cuatro traducciones diferentes proporcionadas por los respectivos traductores web. Esto hace que el corpus original con las frases en castellano tenga 4887 frases para entrenamiento mientras que los corpus en Francés e Inglés contienen 19548. Para equiparar el número de frases se han replicado las frases en castellano, de manera que el corpus de Entrenamiento en castellano contiene también 19548 frases.

Las Tablas 2, 3 y 4 describen los conjuntos de Entrenamiento, Desarrollo y Prueba.

	ES	EN	FR
frases	19548	19548	19548
palabras	147720	152368	142642
vocabulario	638	1277	1927

Tabla 2: Conjuntos de Entrenamiento

Puede verse que el vocabulario tanto para inglés como para francés es mucho mayor que el original en español. Esto se debe a que las traducciones proporcionadas por los traductores web introducen nuevas palabras que además son diferentes entre los diferentes traductores.

	ES para			
	Inglés	Francés	EN	FR
frases	340	277	340	277
palabras	2619	2013	2744	2175
vocabulario	263	235	287	253

Tabla 3: Conjuntos de Desarrollo

El número de frases de los conjuntos de Desarrollo para francés e inglés son diferentes porque una vez adquirido el corpus multilingüe se hizo un control de calidad que descartó algunas frases, mayoritariamente en francés. Se debe tener en cuenta que para el ajuste de parámetros se necesita que el conjunto de frases para el lenguaje origen y para el lenguaje destino contengan el mismo número de frases, ya que uno se traducirá y se comparará con el otro, que será la referencia para ajustar los pesos de cada modelo de traducción. En consecuencia, en español hay dos conjuntos de Desarrollo distintos, uno para cuando el otro idioma es el inglés y otro para cuando es el francés.

	ES	EN	FR
frases	1000	1000	1000
palabras	7637	7701	7650
vocabulario	412	604	586

Tabla 4: Conjuntos de Prueba

A destacar que tanto los conjuntos de Desarrollo como los conjuntos de Prueba son muy pequeños, pero que sin embargo han sido traducidos por humanos.

3 Representación de la semántica y su uso en el modelo de traducción

Inicialmente disponemos de unas etiquetas semánticas y un Frame asociados a cada una de las frases del corpus en castellano. Para incluir esta información en las frases del conjunto de Entrenamiento hemos seguido diferentes aproximaciones. Recordemos que sólo el corpus en español está segmentado y etiquetado semánticamente.

3.1 Etiquetas semánticas asociadas a cada palabra

Esta primera aproximación consiste en añadir a cada palabra en cada frase del corpus de Entrenamiento en castellano la etiqueta

semántica que tiene asociado el segmento al que pertenece la palabra. Al usar esta representación en nuestros modelos conseguimos que las mismas palabras asociadas a distintos conceptos se cuenten como distintas y las probabilidades de aparición de cada palabra, tanto sola como en bigramas o trigramas, sea diferente. Aumentamos el vocabulario de la tarea. Obtenemos los nuevos conjuntos para Entrenamiento, Desarrollo y Prueba, ahora con información semántica.

La Tabla 5 muestra un ejemplo obtenido con este tipo de representación.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Representación
quiero#consulta ir#consulta a#ciudad_destino Segovia#ciudad_destino desde#ciudad_origen Valencia#ciudad_origen

Tabla 5: Representación mediante etiquetas semánticas asociadas a cada palabra.

3.2 Semántica en destino y en origen

Dado que los etiquetados semánticos sólo están disponibles en castellano, es posible usar éstos de dos formas distintas: donde el castellano es la lengua origen (semántica en origen), y cuando es la lengua destino (semántica en destino). En el caso de la semántica en destino, para entrenar el sistema de traducción usamos la aproximación del apartado anterior para modelar la lengua destino. Así obtenemos un modelo de lenguaje y traducción que traducirá desde un idioma fuente (con sólo palabras) a castellano con información semántica.

En el caso de la semántica en origen, el idioma fuente será el castellano enriquecido mediante información semántica, y se traducirá al francés o al inglés. Como resultado obtendremos una traducción al inglés o al francés en la que solamente habrá palabras.

3.3 Uso de interlingua basada en pares atributo-valor

La aproximación mediante interlingua consiste en definir un lenguaje pivote en el proceso de traducción. Una representación con interlingua implica entrenar sistemas que traduzcan desde un lenguaje origen a interlingua, más otros sistemas que traduzcan desde interlingua a un lenguaje destino.

Esta aproximación trata de traducir una frase en un idioma fuente por otra en un idioma destino que tenga un contenido semántico similar, en lugar de traducir palabra por palabra. De esta manera, se intenta conservar el significado de la frase original más que su literalidad. Esto brinda un abanico más grande de posibilidades, ya que tendremos más opciones a la hora de realizar una traducción. Sin embargo, también cabe destacar que la métrica BLEU no será tan representativa con esta aproximación. Una frase en el lenguaje destino, con significado equivalente pero distintas palabras en la frase de referencia obtendría con BLEU una puntuación baja, pese a poder ser una traducción válida.

En esta aproximación hemos empleado el Frame contenido en las frases del corpus DIHANA como lenguaje pivote. Cada frase de entrenamiento se traducirá en la secuencia de pares atributo-valor del frame correspondiente. Un ejemplo de este etiquetado se muestra en la Tabla 6. Nótese que cada par atributo-valor se considera un símbolo en el lenguaje pivote.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
TIPO-VIAJE-nil () CIUDAD-ORIGEN-Valencia CIUDAD-DESTINO-Segovia

Tabla 6: Representación mediante una interlingua basada en pares atributo-valor

Con un corpus de Entrenamiento para la interlingua que contenga este tipo de frases podremos entrenar sistemas de traducción que traduzcan de castellano, inglés o francés a interlingua, y después podremos entrenar otros sistemas de traducción que traduzcan de interlingua a castellano, inglés o francés.

3.4 Uso de interlingua basada en secuencias atributo-valor

Esta representación está basada en la anterior, pero considerando ahora el atributo y su valor como dos símbolos distintos. Un ejemplo de este etiquetado aparece en la Tabla 7.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
TIPO-VIAJE nil () CIUDAD-ORIGEN Valencia CIUDAD-DESTINO Segovia

Tabla 7: Representación mediante una interlingua basada en secuencias atributo-valor.

3.5 Uso de interlingua basada en secuencias de conceptos y valores

La tercera representación con interlingua está basada en todas las representaciones anteriores. En esta aproximación representaremos una frase mediante la secuencia de etiquetas semánticas de la segmentación de esa frase en el corpus DIHANA. Además, se tienen en cuenta las apariciones de valores en la representación de Frame, pero sustituyendo los valores concretos por valores genéricos para aumentar la cobertura del modelo. Esta representación persigue conseguir mejores resultados equiparando los tokens del lenguaje origen con una secuencia de atributos que normalmente es mayor a la secuencia de pares atributo-valor que hay en el Frame. En la Tabla 8 vemos un ejemplo de este tipo de representación.

4 Experimentos y Resultados

Se ha llevado a cabo una evaluación experimental para comprobar la idoneidad de las representaciones semánticas propuestas en el proceso de traducción automática estadística. Como punto de partida, experimento baseline, realizaremos experimentos en los cuales no se ha utilizado ninguna información semántica.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
consulta ciudad_destino#*ciudad*1 ciudad_origen#*ciudad*2

Tabla 8: Representación mediante una interlingua basada en secuencias de conceptos y valores.

4.1 Experimentos con etiquetas asociadas a cada palabra

Se han hechos experimentos tanto para cuando la representación semántica (el etiquetado semántico sólo está disponible para castellano en el corpus DIHANA) está en el destino como cuando está en el origen.

En el experimento con etiquetas asociadas a cada palabra en destino se han llevado a cabo a su vez otros dos tipos de experimentos. En el primer tipo hemos empleado en la fase de desarrollo un conjunto de datos sin etiquetas asociadas de manera que se maximiza el BLEU para traducir de palabras en un idioma origen a palabras en castellano. Sin embargo, para el segundo tipo de experimentos usamos un conjunto de datos con etiquetas asociadas para la fase de desarrollo, de manera que tenemos un sistema que maximiza el BLEU para traducir de palabras en un lenguaje origen a castellano con etiquetas asociadas.

En la Tabla 9 están los resultados para el baseline y para los experimentos con etiquetas asociadas a cada palabra en destino, haciendo uso en un caso de un conjunto de Desarrollo con etiquetas semánticas asociadas y otro conjunto de Desarrollo sin etiquetas semánticas asociadas. En el caso de traducir desde el inglés obtenemos mejores resultados ajustando los pesos mediante el uso de un conjunto de datos sin etiquetas asociadas y conseguimos mejorar al baseline. Sin embargo, al traducir desde el francés obtenemos mejores resultados usando un conjunto de datos para desarrollo con etiquetas asociadas y se mejora también el baseline.

	EN-ES	FR-ES
Baseline	39.06	37.91
Desarrollo=Etiquetas Sem	39.01	**39.38**
Desarrollo=Palabras	**39.49**	39.03

Tabla 9: BLEU de la representación semántica con etiquetas asociadas a cada palabra.

A continuación se muestra un ejemplo de traducción EN-ES y FR-ES:

(1) Hello, good morning. I'd like times for trains to Cuenca → hola buenos días quisiera horarios para trenes a Cuenca

(2) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → hola quisiera saber horarios de trenes para ir a Barcelona

En la Tabla 10 se han vuelto a realizar los dos experimentos para semántica en destino, pero esta vez sin eliminar las etiquetas asociadas a las palabras en la salida de traducción y empleando un conjunto de Prueba con etiquetas asociadas para que sirva de referencia. De este modo estamos midiendo no solo el acierto del traductor a nivel de palabras sino además el acierto considerando los conceptos semánticos asociados a ellas. Ésta es una tarea más compleja que la anterior, y justifica el hecho de que se obtenga un BLEU menor, pues el rango de palabras que tiene para acertar al traducir al castellano sin etiquetas asociadas es de 719 palabras, mientras que al castellano con etiquetas asociadas es de 1746 palabras.

	EN-ES	FR-ES
Desarrollo= Etiquetas Sem	33.44	**34.03**
Desarrollo=Palabras	**33.99**	33.61

Tabla 10: BLEU de la representación semántica con etiquetas asociadas a cada palabra, evaluando las etiquetas semánticas.

A continuación se muestra un ejemplo de traducción EN-ES y FR-ES:

(3) Hello, good morning. I'd like times for trains to Cuenca → hola#cortesia buenos#cortesia días#cortesia quisiera#consulta horarios# <hora> para# <hora> trenes# <hora> a#ciudad_destino Cuenca#ciudad_destino

(4) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → hola#cortesia quisiera#consulta saber#consulta horarios# <hora> de# <hora> trenes# <hora> para# <hora> ir# <hora> a#ciudad_destino Barcelona#ciudad_destino

Por último en la Tabla 11 se muestran los experimentos usando etiquetas asociadas a cada palabra en origen. En este caso el lenguaje origen usa siempre las etiquetas asociadas a cada palabra del castellano, que se traduce a francés o inglés.

Los resultados del baseline obtienen menos BLEU que cuando el castellano es la lengua destino. La razón puede estar en que las traducciones de inglés y francés han sido obtenidas por cuatro traductores web diferentes, generando un vocabulario mayor y que hace más difícil que se den aciertos.

	ES - EN	ES - FR
Baseline	**22.01**	26.91
Etiquetas Sem	21.72	**27.01**

Tabla 11: BLEU de la semántica en origen

A continuación se muestra un ejemplo de traducción ES-FR:

(5) sí# <afirmacion> me#consulta podrías#consulta decir#consulta el# <precio> precio# <precio> del#numero_relativo_orden primero#numero_relativo_orden del#numero_relativo_orden último#numero_relativo_orden → oui vous me dire le prix du premier du dernier

4.2 Experimentos usando una interlingua basada en pares atributo-valor

La Tabla 12 muestra los resultados obtenidos usando una interlingua basada en pares atributo-valor. Se hace uso de los pares atributo-valor del *Frame* de cada frase.

Se ha hecho un experimento para cada combinación de traducciones posible, cabe destacar los experimentos hechos para traducir de un idioma a interlingua y después otra vez a ese mismo idioma. En este caso lo que se hace es traducir por palabras o segmentos que expresen los mismos atributos semánticos. Dicho de otro modo, lo que hacemos al traducir al interlingua es "comprender" el significado de la frase, y al traducir desde un interlingua traducimos a una frase

con ese significado. En la traducción usando una interlingua, aunque generemos una frase con el mismo significado y las palabras a la salida del proceso de traducción sean sinónimos de las palabras en la referencia, la métrica BLEU las contará como un error.

	BLEU
ES-interlingua-ES	18.77
EN-interlingua-ES	8.06
FR-interlingua-ES	8.55
ES-interlingua-EN	6.34
EN-interlingua-EN	8.54
FR-interlingua-EN	4.60
ES-interlingua-FR	10.45
EN-interlingua-FR	7.86
FR-interlingua-FR	11.78

Tabla 12: BLEU de una representación mediante Interlingua de pares atributo-valor.

A continuación se muestra un ejemplo de traducción ES-IL e IL-ES mediante una interlingua basada en pares atributo-valor:

(6) hola buenos días mira quería saber horarios de trenes para ir de Castellón a Barcelona → TIPO-VIAJE-ida TIPO-VIAJE-nil HORA CIUDAD-ORIGEN-Castellón CIUDAD-DESTINO-Barcelona → me gustaría saber el horario de un viaje de ida para ir de Castellón a Barcelona

En el ejemplo se observa una traducción correcta que sin embargo proporciona un BLEU bajo.

4.3 Experimentos usando una interlingua basada en secuencias atributo-valor

Con esta representación se espera aumentar la cobertura del modelo dado que no hará falta que atributo y valor aparezcan juntos para poder ser traducidos. En la Tabla 13 se muestran los resultados obtenidos para esta aproximación realizando un experimento por cada par posible de idiomas. Como se esperaba, si lo comparamos con la tabla 12 los resultados en BLEU han aumentado. El vocabulario en esta aproximación es menor (370 tokens) que en la representación de interlingua anterior, debido a que no hay combinaciones concepto-valor sino que tenemos los conceptos por un lado y los valores por el otro, dando un número menor de tokens.

A continuación se muestra un ejemplo de traducción FR-IL e IL-EN mediante una interlingua basada en secuencias atributo-valor:

	BLEU
ES-interlingua-ES	19.96
EN-interlingua-ES	10.10
FR-interlingua-ES	11.75
ES-interlingua-EN	10.10
EN-interlingua-EN	11.33
FR-interlingua-EN	8.58
ES-interlingua-FR	10.87
EN-interlingua-FR	7.98
FR-interlingua-FR	10.25

Tabla 13: BLEU de la representación mediante Interlingua de secuencias atributo-valor.

(7) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → TIPO-VIAJE nil HORA CIUDAD-DESTINO barcelona → i'd like to know the schedules of trains to barcelona

4.4 Experimentos usando una interlingua basada en secuencias de conceptos y valores

Para esta última representación de interlingua, y al igual que para las anteriores se ha llevado a cabo un experimento por cada combinación de idiomas (ver Tabla 14). Los resultados en BLEU son generalmente mayores que las obtenidas por las representaciones de interlingua anteriores. Esto se debe a que al hacer uso de los conceptos en lugar de sólo los atributos del *Frame* podemos enriquecer la representación semántica, tal es el caso de la etiqueta *cortesía*. Además, las otras representaciones interlingua que hemos usado en este trabajo sólo usan el *Frame*, y este tiene los pares atributo-valor ordenados canónicamente, con lo que la tarea de reordenamiento en el proceso de traducción era también más difícil.

A continuación se muestra un ejemplo de traducción EN-IL e IL-ES mediante una interlingua basada en secuencias conceptos y valores:

(8) Well, I want leave on the thirtieth of July → afirmacion consulta m_salida fecha#30X07 tipo_viaje#vuelta july → sí me gustaría salir el próximo día treinta de vuelta july

	BLEU
ES-interlingua-ES	32.67
EN-interlingua-ES	17.44
FR-interlingua-ES	17.86
ES-interlingua-EN	8.62
EN-interlingua-EN	9.18
FR-interlingua-EN	7.60
ES-interlingua-FR	12.71
EN-interlingua-FR	8.32
FR-interlingua-FR	11.87

Tabla 14: BLEU de la representación mediante una interlingua de secuencias de conceptos y valores.

5 Conclusiones

En este trabajo hemos presentado diferentes alternativas para añadir conocimiento semántico a los sistemas de traducción automática en el ámbito de una tarea de dominio restringido. Estas alternativas se pueden agrupar en dos clases. En una de ellas se usan los conceptos semánticos que tiene el corpus DIHANA para enriquecer el proceso de traducción. En la otra se usa una lengua pivote como paso intermedio en el proceso de traducción entre idiomas.

Con la primera de las aproximaciones hemos obtenido resultados superiores a los del baseline, es decir, a la traducción cuando no se usa información semántica. Se puede inferir de todo esto que si trabajamos con un corpus de dominio cerrado, tal como DIHANA, y podemos asignar etiquetas semánticas a las palabras entonces es posible mejorar los resultados de la traducción con estas representaciones siempre y cuando el corpus sea lo suficientemente grande.

En cuanto al uso de una lengua pivote, si bien la puntuación en BLEU es menor que en la otra aproximación, la propia métrica BLEU tampoco refleja bien su calidad. Al traducir de una lengua origen a una interlingua se está haciendo una interpretación del significado de las palabras de la lengua origen, y al traducir de este interlingua a la lengua destino se está generando una frase que expresa el significado deseado. Es frecuente pues que en la traducción aparezcan palabras diferentes pero que sin embargo tienen el mismo signficado.

Los resultados obtenidos indican que es alentador continuar en esta dirección. Esto implicaría, entre otras medidas, desarrollar una nueva métrica para cuantificar la calidad de una traducción, de forma que se tengan en cuenta las similitudes semánticas además de las léxicas.

Bibliografía

Babych, B., A. Hartley, y S. Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. En *Proceedings of the MT Summit XI*, páginas 412–418.

Banchs, R. E. y M. R. Costa-jussà. 2011. A semantic feature for statistical machine translation. En *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, páginas 126–134. ACL.

Benedí, J.-M., E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, y A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. En *Proceedings of LREC 2006*, páginas 1636–1639.

Federico, M., N. Bertoldi, y M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. En *INTERSPEECH*, páginas 1618–1621.

Gao, J., J. Goodman, M. Li, y K.-F. Lee. 2002. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing*, 1(1):3–33.

Habash, N. y J. Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. En *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, páginas 173–181. ACL.

Koehn, P. y et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. En *Proc. of ACL demonstration session*, páginas 177–180.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Procs. of Machine Translation Summit X*, páginas 79–86.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edición.

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

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 109-116 recibido 20-03-2016 revisado 20-04-2016 aceptado 01-05-2016

Comparing Distributional Semantics Models for identifying groups of semantically related words

Comparación de dos modelos de semántica distribucional para identificar grupos de palabras semánticamente relacionadas

Venelin Kovatchev, Maria Salamó, M. Antònia Martí
Universitat de Barcelona
Gran Via 585, 08007 Barcelona, Spain
{vkovatchev, maria.salamo, amarti}@ub.edu

Abstract: Distributional Semantic Models (DSM) are growing in popularity in Computational Linguistics. DSM use corpora of language use to automatically induce formal representations of word meaning. This article focuses on one of the applications of DSM: identifying groups of semantically related words. We compare two models for obtaining formal representations: a well known approach (CLUTO) and a more recently introduced one (Word2Vec). We compare the two models with respect to the PoS coherence and the semantic relatedness of the words within the obtained groups. We also proposed a way to improve the results obtained by Word2Vec through corpus preprocessing. The results show that: a) CLUTO outperforms Word2Vec in both criteria for corpora of medium size; b) The preprocessing largely improves the results for Word2Vec with respect to both criteria.
Keywords: DSM, Word2Vec, CLUTO, semantic grouping

Resumen: Los Modelos de Semántica Distribucional (MSD) están siendo utilizados de manera extensiva en el área de la Lingüística Computacional. Los MSD utilizan corpus de uso de la lengua para inducir de manera automática diferentes tipos de representaciones sobre el significado de las palabras. Este artículo se centra en una de las aplicaciones de los MSD: la identificación de grupos de palabras semánticamente relacionadas. Se comparan dos modelos de obtención de representaciones formales: CLUTO, una herramienta estándar de clusterización y Word2Vec, una aproximación reciente al tema. Comparamos los resultados obtenidos con ambos modelos basándonos en dos criterios: la coherencia que presentan estas agrupaciones respecto de la categoría morfosintáctica y la cohesion semántica entre las palabras dentro de cada grupo. Se propone también como mejorar los resultados obtenidos con Word2Vec mediante su preprocesamiento morfosintáctico. Los resultados obtenidos demuestran que: a) CLUTO supera a Word2Vec en ambos criterios cuando se trata de corpus de tamaño medio: b) el preprocesamiento mejora de manera clara los resultados obtenidos con Word2Vec para ambos criterios.
Palabras clave: DSM, Word2Vec, CLUTO, agrupación semántica de palabras.

1 Introduction

In recent years, the availability of large corpora and the constantly increasing computational power of the modern computers have led to a growing interest in linguistic approaches that are automated and data-driven (Arppe et al., 2010). Distributional semantic models (DSM) (Turney and Pantel, 2010; Baroni and Lenci, 2010) and the vector representations (VR) they generate fit very well within this framework: the process of extracting vector representations is mostly automated and the content of the representations is data-driven.

The format of the vector is suitable for carrying out different mathematical manipulations. Vectors can be compared directly through an objective mathematical function. They can also be used as a dataset for various Machine Learning algorithms. VR are more often used on tasks related to lexical simi-

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

larity and relational similarity (Turney and Pantel, 2010). In such tasks, the emphasis is on pairwise comparisons between vectors.

This article focuses on another use of the Vector Representations: the grouping of vectors, based on their similarity in the Distributional space. This grouping can be used, among other things, as a methodology for identifying groups of semantically related words. High quality groupings can serve for many purposes: they are a semantic resource on their own, but can also be applied for syntactic disambiguation or pattern identification and generation (Martí et al., Submitted, 2016), for example.

We compare two different methodologies for obtaining groupings of semantically related words in English - a well known approach (CLUTO) and a more recently introduced one (Word2Vec). The two methodologies are evaluated in terms of the quality of the obtained groups. We consider two criteria: 1) the semantic relatedness between the words in the group; and 2) the PoS coherence of the group. We evaluate the role of the corpus size with both methodologies and in the case of Word2Vec, the role of the linguistic preprocessing (lemmatization and PoS tagging).

The rest of this paper is organized as follows: Section 2 presents the general framework and related work. Section 3 describes the available data and tools. Section 4 presents the experiments and the results obtained. Finally Section 5 gives conclusions and identifies directions for future work.

2 Related work

Distributional Semantics Models (DSM) are based on the Distributional Hypothesis, which states that the meaning of a word can be represented in terms of the contexts in which it appears (Harris, 1954; Firth, 1957). As opposed to semantic approaches based on primitives (Boleda and Erk, 2015), approaches based on distributional semantics can obtain formal representations of word meaning from actual linguistic productions. Additionally, this data-driven process for semantic representation can mostly be automated.

Within the framework of DSM, one of the most common ways to formalize the word meaning is a vector in a multi-dimensional distributional space (Lenci, 2008). For this purpose, a matrix with size **m** by **n** is extracted from the corpus, representing the distribution of **m** words over **n** contexts. The format of a vector allows for direct quantitative comparison between words using the apparatus of linear algebra. At the same time it is a format preferred by many Machine Learning algorithms.

The choice of the matrix is central for the implementation of a particular DSM. Turney and Pantel (2010) suggest a classification of the DSM based on the matrix used. They analyze three different matrices: term-document, word-context, and pair-pattern. The different matrices represent different types of relations in the corpus and the choice of the matrix depends on the goals of the particular research.

Baroni and Lenci (2010) present a different, sophisticated approach for extracting information from the corpus. They organize the information as a third order tensor, with the dimensions representing <'word', 'link', 'word' >. This third order tensor can then be used to generate different matrices, without the need of going back to the original corpus.

In this paper we focus on one of the classical vector representations - the one based on word-context relation. It measures what Turney and Pantel (2010) call "attributional similarity". In particular, we are interested in the possibility to group vectors together, based on their relations in the distributional space.

Erk (2012) offers a survey of possible applications of different DSM. She lists clustering as an approach that can be used with vectors, for word sense disambiguation. Moisl (2015) presents a theoretical analysis on the usage of clustering in computational linguistics and identifies key aspects of the mathematical and linguistic argumentation behind it.

Here we analyze and compare two approaches that induce vector representations from a corpus and apply algorithms to identify sets of semantically related words. We are interested in the quality of the obtained groups, as we believe that they can be a useful, empirical, linguistic resource.

Martí et al. (Submitted, 2016) present a methodology named DISCOveR for identifying candidates to be constructions from a corpus. As part of this methodology they use CLUTO (Karypis, 2002) for clustering words based on their vector representations. Their

approach uses a word-context matrix where the context is defined by combining a syntactic dependency with a lemma. After all the vectors are extracted, CLUTO is used in order to obtain clusters of semantically related words. Later on these clusters are used to generate a list of the candidates to be constructions.

Mikolov et al. (2013) suggest a different approach towards extracting vector representations and grouping. Their methodology is based on deep learning and is intended for quick processing of very large corpora. Word2Vec[1], the tool they present, includes an integrated algorithm for grouping words based on proximity in space. The context they use for vector extraction is simple co-occurrence within a specified window of tokens. Originally, they make no use of linguistic preprocessing such as lemmatization, part of speech tagging or syntactic tagging. As part of this paper we evaluate the effect of linguistic preprocessing on the obtained vectors and groups.

3 Data and tools

In this section we present the corpus that we use in the evaluation (Section 3.1) and the two methodologies (Section 3.2 and Section 3.3).

3.1 The corpus

For all of the experiments described in this paper, we use PukWaC (Baroni et al., 2009)[2]. It is a 2 billion word corpus of English, built up from sites in the .uk domain. It is available online and is already preprocessed: XML tags and other non-linguistic information have been removed, it is lemmatized, PoS tagged and syntactically parsed. The PoS tagset is an extended version of the Penn Treebank tagset. The syntactic dependencies follow the CONLL-2008 shared task format.

3.2 Grouping with CLUTO

DISCOveR (Martí et al., Submitted, 2016) is a methodology for identifying candidates to be construction from a corpus. It uses vector representations, extracted from a corpus. CLUTO (Karypis, 2002) is used on these representations in order to obtain clusters of semantically related words. CLUTO is a software package for clustering low and high dimensional data sets and for analysis of the characteristics of the various clusters. CLUTO provides three different classes of clustering algorithms, based on partitional, agglomerative and graph-partitioning paradigms. It computes clustering solution based on one of the different approaches.

For this article, we are interested only in the first three steps of the DISCOveR process. Step 1 is the linguistic preprocessing of the corpus. The raw text is cleared from non-linguistic data, it is PoS tagged and syntactically parsed. In Step 2, the DSM matrix is constructed. The rows of the matrix correspond to lemmas and the columns correspond to contexts. Contexts in this approach are defined as a triple of syntactic relation, direction of the relation and lemma in [direction:relation:lemma] format[3]. This matrix is used to generate vector representations for the 10,000 most frequent words in the corpus. Next, Step 3 uses CLUTO to create clusters of semantically related lemmas from the DSM matrix and the corresponding vectors. The clusters are created based on shared contexts.

Martí et al. (Submitted, 2016) start from a raw, unprocessed corpus and in Step 1 they clear the corpus and tag it with the linguistic data relevant to the matrix extraction. The format they use is shown in Table 1.

Token	sanitarios
Lemma	sanitario
PoS	NCMP
Short PoS	n
Sent ID	000
Token ID	0
Dep ID	2
Dep Type	suj

Tabla 1: Diana-Araknion Format

The original DISCOveR experiment is done with the Diana-Araknion corpus of Spanish. For the purpose of this article, we replicated the process for English, using the Puk-

[1] Available at: https://code.google.com/archive/p/word2vec/

[2] Available at: http://wacky.sslmit.unibo.it

[3] For example, from the sentence "El barbero afeita la larga barba de Jaime", three different contexts of the noun lemma barba are generated: [<:dobj:afeitar_v], [>:mod:largo_a] and [>:de_sp:pn_n]. The example is from (Martí et al., Submitted, 2016)

WaC corpus. For step 1 we had to make sure that our preprocessing is equivalent to the one of Diana-Araknion. The corpus PukWaC is already preprocessed and the format is similar to the one of Diana-Araknion. However, in order to make it fully compatible, we had to make several modifications of the format and linguistic decisions. Regarding the format, we removed any remaining XML tags, enumerated the sentences in the corpus, and generated "short PoS"[4]. From the linguistic side, we had to decide whether all PoS and Dependencies were relevant for the vector generation or some of them could be merged together or even discarded in order to optimize and speed up the process.

The process of generating vectors and clusters is based on analyzing the contexts where each word appears in. A word is identified by its lemma and its PoS tag. However, in the PukWac tagset there are many PoS tags which specify not only the PoS of the token, but also contain information about other grammatical features, such as person, number, and tense. If these tags are kept unchanged, a separate vector will be generated for different forms of the same word, based on different PoS tag. To avoid this problem and to generate only one vector for all of the different word forms, we have decided to merge certain PoS tags under one category.

We decided to simplify the POS tagset further. It is a common practice in DSM to focus the experiment on the relations between content words. Function words and punctuation are usually not considered relevant contexts. Because of that, we have put them under the common tag "other". All of the changes on the PoS tagset are summarized in Table 2.

The list of syntactic dependencies in PukWaC is also not fully relevant to the task of vector generation. While the unnecessary PoS tags may lead to multiple vectors for the same word, unnecessary dependencies generate additional contexts, increasing the dimensionality of the vectors and leading to a more complicated computational process. Therefore the modification of the dependencies is mostly related to the optimization of the computational process. After analyzing the tagset, we have decided to merge the

[4]short PoS is a one letter tag representing the generic PoS tag of the lemma. In this experiment, short PoS is the first letter of the full PoS

Tag	Original tag	Description
J	JJ JJR JJS	Adjective
M	MD	Modal verb
N	NN NNS	Noun (common)
NP	NP NPS	Noun (personal)
R	RB RBR RBS RP	Adverb
S	IN	Preposition
V	VB* VH* VV*	Verb (all)
O	CC CD DT PDT EX FW LS POS PP* SYM TO UH W* punctuation	Rest

Tabla 2: PoS tagset modifications

OBJ and **IOBJ** tags due to some inconsistencies of their usage. We have also decided to discard the following relations: **CC** (conjunction), **CLF** (be/have in a complex tense), **COORD** (coordination), **DEP** (unclassified relation), **EXP** (experiencer in few very specific cases), **P** (punctuation), **PRN** (parenthetical), **PRT** (particle), **ROOT** (root clause).The final list of dependencies is shown in Table 3.

Dependency	Description
ADV	Unclassified adv
AMOD	Modifier of adj or adv
LGS	Logical subj
NMOD	Modifier of nom
OBJ	Direct or indirect obj
PMOD	Preposition
PRD	Predicative compl
SBJ	Subject
VC	Verb chain
VMOD	Modifier of verb
empty	No dependency

Tabla 3: Syntactic Dependencies

Once the corpus is preprocessed, the process of matrix extraction is mostly automated. For the matrix, we have only generated vectors for words that appear at least 5 times in the corpus. Out of them we have used only the vectors of the 10,000 most frequent words for the clustering process.

For the clustering process, we configure CLUTO to use direct clustering, based on the H2 criterion function, with 25 features

per cluster. We have ran the clusterization multiple times, ranging from 100 to 1,000 clusters. We then used CLUTO's H2 metric to determine the optimal number of clusters, which has been 800 for all of the experiments.

3.3 Grouping with Word2Vec

Word2Vec is based on the methodology proposed by Mikolov et al. (2013). It takes a raw corpus and a set of parameters and generates vectors and groups. The algorithm of Word2Vec is based on a two layer neural network that are trained to reconstruct linguistic context of words. Word2Vec includes two different algorithms - Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW learns representations based on the context as a whole - all of the words that co-occur with the target word in a specific window. Skip-Gram learns representation based on each single other word within a specified window. When using Word2Vec usually the emphasis is put on the choice of the paremeters for the algorithm, and not on the specifications of corpus. However, we consider that the specifications of the corpus (size and linguistic preprocessing) can largely affect the quality of the obtained results.

By default Word2Vec works with a raw corpus. Neither of the two models makes explicit use of morpho-syntactic information. However, by modifying the corpus, some morphological information can be used implicitly. If the token is replaced by its corresponding lemma or by the lemma and part of speech tag in a "lemma_pos" format, the resulting vectors would be different: using the lemma would generate only one vector for the word as opposed to separate vector for every word form; using PoS can make a distinction between homonyms with same spelling and different PoS. As part of our work we wanted to examine how linguistic preprocessing can affect the quality of the vectors. For that reason we created three separate corpus samples - one raw corpus, one where each token was replaced by its lemma, and one where each token was replaced by "lemma_pos". We generated vectors separately for each of the corpora. Unfortunately, there was no trivial way to introduce syntactic information implicitly in the models of Word2Vec.

4 Experiments

In this section we present the setup for the different experiments (Section 4.1), the evaluation criteria (Section 4.2), and the obtained results (Section 4.3).

4.1 Setup

We carried out a total of 15 experiments - 3 experiments using CLUTO and 12 experiments using Word2Vec. For the experiments with CLUTO, the only variation between the experiments was the size of the corpus: 4M tokens, 20M tokens, and 40M tokens[5]. In all the experiments we used the preprocessing described at Section 3.2, we generated vectors for the 10,000 most frequent words and we split them into 800 clusters. For the experiments with Word2Vec, we changed three parameters of the experiments: (1) the algorithm (CBOW and Skip-Gram), (2) the linguistic preprocessing of the corpus (raw, lemma, lemma and PoS), and (3) the size of the corpus (4M, 20M, and 40M). We carried out 9 experiments with CBOW (all size and preprocessing combinations) and 3 experiments with Skip-Gram (the three variants of the 40M corpus). Mikolov et al. (2013) identify two important parameters to be set up when using Word2Vec: the vector size and the window size. For the window size, we used 8, which is the recommended value. For the vector size, Mikolov et al. (2013) show that increasing vector size from 100 to 300 leads to significant improvement of the results, however further increase does not have big impact. For that reason we have chosen vector size of 400, which is above the recommended minimum. For the number of groups we used 800: the same number that was determined optimal for CLUTO. For the number of lemmas, we used the 10,000 most frequent ones, the same setup as with CLUTO.

4.2 Evaluation

The two methodologies and all of the different setups are evaluated based on the quality of the obtained groups. We consider two criteria: 1) The semantic relatedness between the words in each group; and 2) The PoS coherence of the groups. The PoS coherence is a secondary criterion which should be

[5]The 40M corpus contains in itself the 20M corpus. The 20M corpus contains in itself the 4M corpus. The same corpora has been used for the experiments with both CLUTO and with Word2Vec.

considered in addition to the semantic relatedness. Our intuition is that groups that are semantically related and PoS coherent are a better resource than groups that are only semantically related. For evaluating the semantic relations of the words in the groups, we present two methodologies - an automated method based on WordNet distances and a manual evaluation done by experts on a subset of the groups in each experiment. The PoS coherence is calculated automatically.

There is no universal widely accepted criteria for determining the semantic relations between two words. Two of the most common approaches are calculating WordNet distances and expert intuitions. We used both when evaluating the quality of the obtained groups.

For the WordNet similarity evaluation, we use the WordNet interface built in NLTK (Bird, Klein, and Loper, 2009). We calculate the Leacock-Chodorow Similarity[6] between each two words[7] in every group. We then sum all the obtained scores and divide them by the number of pairs to obtain average WordNet similarity for each method.

For the expert evaluation, we selected a subset of groups, generated in each experiment[8]. Three experts were asked to rate each group on a scale from 1 (unrelated) to 4 (strongly related)[9]. We calculate the average between all of the scores they gave on the groups of each experiment.

We define PoS coherence as the percent of words that belong to the most common PoS tag in each group. In order to calculate it, all obtained groups are automatically PoS tagged[10]. Then for each group, we count the percent of words that belong to each PoS and identify the most common tag.

4.3 Results

Table 4 shows the WordNet similarity evaluation. The average similarity score obtained by CLUTO is higher than the score obtained by Word2Vec (0.81-0.96 against 0.67-0.81). This indicates that the distances between the words in the CLUTO groups are shorter and the semantic relations are stronger. Increasing the corpus size improves the results for both CLUTO and Word2Vec. Preprocessing (specifically PoS tagging) improves the obtained results for all of the Word2Vec experiments. The groups obtained using Skip-Gram get lower scores in the evaluation compared with the groups obtained using CBOW.

Methodology	Corpus	Similarity
W2V-CBOW	4M (raw)	0.67
W2V-CBOW	4M (lemma)	0.67
W2V-CBOW	4M (pos)	0.72
W2V-CBOW	20M (raw)	0.74
W2V-CBOW	20M (lemma)	0.75
W2V-CBOW	20M (pos)	0.77
W2V-CBOW	40M (raw)	0.77
W2V-CBOW	40M (lemma)	0.78
W2V-CBOW	40M (pos)	0.81
W2V-SG	40M (raw)	0.69
W2V-SG	40M (lemma)	73
W2V-SG	40M (pos)	0.74
CLUTO	4M	0.81
CLUTO	20M	0.92
CLUTO	40M	0.96

Tabla 4: Wordnet Similarity

Table 5 shows the results from the expert evaluation of the semantic relations in the groups. The data is similar to the results with WordNet distances. The groups obtained by CLUTO show higher degree of semantic relatedness (2.8-3.4) compared to the groups obtained by Word2Vec (1.6-2.7). The CLUTO groups at 20M and 40M obtain average above 3, meaning that the experts consider all of the groups to be strongly related. For the experiments with Word2Vec, linguistic preprocessing improves the results, especially at biger corpus size (2.5 against 1.8 for 20M and 2.7 against 2 for 40M). The groups obtained using Skip-Gram algorithm are rated lower than the groups obtained using CBOW. The

[6] It calculates word similarity, based on the shortest path that connects the senses and the maximum depth of the taxonomy in which the senses occur.

[7] The calculation is based on the first sense of every word

[8] We selected the groups based on a word they contain - three verb groups (the ones that contain "say", "see", "want"), 3 noun groups ("person", "year", "hand"), 1 adjective group("good"), 1 adverb group("well). All of the selected words are among the 100 most commonly used words of English.)

[9] In the detailed description of the scale given to the experts: 1 corresponds to "no semantic relation"; 2 corresponds to "semantic relation between some words (less than 50 % of the group); 3 corresponds to "semantic relation between most of the words in the corpus (more than 50 %), but with multiple unrelated words"; 4 corresponds to "semantic relation between most of the words in the corpus, without many unrelated words"

[10] We use only the short PoS tag for this evaluation

preprocessed corpus obtains better groups, but the difference is smaller than the one observed with CBOW.

Methodology	Corpus	Score
W2V-CBOW	4M (raw)	1.6
W2V-CBOW	4M (lemma)	1.4
W2V-CBOW	4M (pos)	1.8
W2V-CBOW	20M (raw)	1.8
W2V-CBOW	20M (lemma)	2.4
W2V-CBOW	20M (pos)	2.5
W2V-CBOW	40M (raw)	2
W2V-CBOW	40M (lemma)	2.1
W2V-CBOW	40M (pos)	2.7
W2V-SG	40M (raw)	1.7
W2V-SG	40M (lemma)	1.8
W2V-SG	40M (pos)	2
CLUTO	4M	2.8
CLUTO	20M	3.2
CLUTO	40M	3.4

Tabla 5: Expert evaluation

Table 6 shows the results for the PoS coherence evaluation. The data shows that the groups obtained from CLUTO are more PoS coherent, compared with the groups obtained by Word2Vec (90-98 % against 69-81 %). For the corpora of size 20M and above, the groups obtained by CLUTO have almost 100 % PoS coherence, meaning that all of the lemmas belong to the same PoS. Both CLUTO and Word2Vec show improved results with the increase of corpus size. The results with Word2Vec indicate that corpus preprocessing largely improves the obtained results (69 %-73 % against 75 %-81 %). In fact, for this experiment the corpus preprocessing have bigger impact than the corpus size: a preprocessed corpus with a size of 4M generates more PoS coherent groups than raw 40M corpus (74-75 % against 73 %). The experiments with Skip-Gram obtain similar results for raw corpus. For Skip-Gram the preprocessed corpus also obtains better overall results, however lemmatized corpus obtains better results than the PoS tagged corpus.

Overall, all three evaluations identify similar patterns in the obtained clusters: (1) the groups obtained by CLUTO perform better than the groups obtained by Word2Vec; (2) Increasing the corpus size improves the quality of the results for both methodologies. This is true for semantic relatedness as well as for PoS coherence. The tendency to obtain more PoS coherent groups justifies the usage of PoS coherence as evaluation criteria; (3) Linguistic preprocessing improves the quality of the groups obtained by Word2Vec (with both algorithms).

Methodology	Corpus	PoS
W2V-CBOW	4M (raw)	69 %
W2V-CBOW	4M (lemma)	74 %
W2V-CBOW	4M (pos)	75 %
W2V-CBOW	20M (raw)	72 %
W2V-CBOW	20M (lemma)	77 %
W2V-CBOW	20M (pos)	80 %
W2V-CBOW	40M (raw)	73 %
W2V-CBOW	40M (lemma)	78 %
W2V-CBOW	40M (pos)	81 %
W2V-SG	40M (raw)	73 %
W2V-SG	40M (lemma)	80 %
W2V-SG	40M (pos)	77 %
CLUTO	4M	90 %
CLUTO	20M	97 %
CLUTO	40M	98 %

Tabla 6: PoS coherence

5 Conclusions and future work

This article compares two methodologies for identifying groups of semantically related words based on Distributional Semantic Models and vector representations. We applied the methodologies to a corpus of English and compared the quality of the obtained groups in terms of semantic relatedness and PoS coherence. We also analyzed the role of different factors, such as corpus size and linguistic preprocessing.

In the comparison of the two methodologies, the results show that CLUTO outperforms Word2Vec with respect to grouping, using corpora of medium size (20M - 40M). However, the quality of the results does depend on the size of the corpus. At 40M CLUTO already obtains very high quality results (98 % PoS coherence and 3.4/4 strength of semantic relationships in the evaluation of the experts) so further increase of the corpus is not likely to show large improvement. On the contrary at 40M Word2Vec still has room for improvement and we expect to narrow the difference between the two methodologies using much larger corpora (1B and above).

In the comparison of the different preprocessing corpora (i.e., raw, lemma, and PoS) in Word2Vec, the results show that lemmatization and PoS tagging largely improve the quality of the groups in both CBOW and Skip-Gram algorithms. This observation is consistent throughout all of the experiments and with respect to all of the evaluation criteria.

The presented comparison opens several lines of future research. First, the evaluation can be extended to bigger corpora, bigger number of vectors, and other languages. Second, the information provided and the suggested criteria for evaluation can be applied to other approaches to DSM and grouping. Finally, the different methodologies and preprocessing options can be evaluated in as part of more complex systems.

Acknowledgments

This work was supported by projects TIN2012-38603-C02-02, SGR-2014-623 and TIN2015-71147-C2-2.

We are grateful to Mariona Taulé, Horacio Rodríguez and the anonymous reviewers for their valuable comments.

References

Arppe, A., G. Gilquin, D. Glynn, M. Hilpert, and A. Zeschel. 2010. Cognitive corpus linguistics: five points of debate on current theory and methodology. *Corpora*, 5(1):1–27.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, 36(4):673–721, December.

Bird, S., E. Klein, and E. Loper, 2009. *Natural Language Processing with Python.*

Boleda, G. and K. Erk. 2015. Distributional semantic features as semantic primitives – or not.

Erk, K. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Harris, Z. 1954. Distributional structure. *Word*, 10(23):146–162.

Karypis, G. 2002. CLUTO a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota.

Lenci, A. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20(1):1–31.

Martí, M. A., M. Taulé, V. Kovatchev, and M. Salamó. Submitted, 2016. Discover: Distributional approach based on syntactic dependencies for discovering constructions. *Natural Language Engineering.*

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Moisl, H. 2015. *Cluster Analisys for Corpus Linguistics.* De Gruyter Mouton.

Turney, P. D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 117-124 recibido 28-03-2016 revisado 20-04-2016 aceptado 03-05-2016

Tratamiento de Redes Sociales en Desambiguación de Nombres de Persona en la Web

Treatment of Social Media in Person Name Disambiguation in the Web

Agustín D. Delgado[1], **Raquel Martínez**[1], **Soto Montalvo**[2], **Víctor Fresno**[1]
1. Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal, 16, 28040 - Madrid
2. Universidad Rey Juan Carlos (URJC), Tulipán, S/N, 28933 - Móstoles
agustin.delgado@lsi.uned.es, raquel@lsi.uned.es, soto.montalvo@urjc.es, vfresno@lsi.uned.es

Resumen: En este trabajo presentamos dos heurísticas para tratar páginas web correspondientes a redes sociales en el problema de desambiguación de nombres de persona en la Web. Este problema consiste en agrupar las páginas web proporcionadas por un motor de búsqueda al consultar un nombre de persona según el individuo al que se refieren. Aunque estas páginas web pueden afectar negativamente en la agrupación de los resultados, la mayoría de sistemas del estado del arte no tienen en cuenta su papel en este problema. Hemos evaluado nuestras heurísticas con dos colecciones que contienen este tipo de páginas web. Para agrupar las páginas web hemos utilizado una extensión de un algoritmo del estado del arte. Ambas heurísticas obtienen mejoras cuando hay un número elevado de páginas sociales y el algoritmo propuesto es más independiente del nivel de ambigüedad de los nombres de persona que otros propuestos por el estado del arte.
Palabras clave: búsqueda de personas en la web, redes sociales, clustering

Abstract: In this work, we present two heuristics to treat web pages from social networks for person name disambiguation in the Web. This problem consists in clustering the results provided by a search engine when the query is a person name according to the individual they refer to. Although these web pages could negatively affect when grouping the results, most of the systems in the state-of-the-art do not take into account their role in this problem. We have evaluated our heuristics with two collections that contain this kind of web pages. We have used an extension of an algorithm of the state of the art to cluster the web pages. Both heuristics get improvements when there is a high number of social web pages, and the proposed algorithm is more independent with respect to the ambiguity degree of person names than other ones in the state of the art.
Keywords: web people search, social media, clustering

1 Introducción

La desambiguación de nombres de personas es un reto dentro del Procesamiento del Lenguaje Natural. Un escenario real donde es de gran ayuda diferenciar entre distintos individuos con el mismo nombre lo encontramos en los motores de búsqueda. Cuando un usuario quiere buscar información sobre una persona en particular, se encuentra con un ranking de links que pueden hablar de diferentes personas que comparten el mismo nombre, de manera que debe seleccionar del ranking aquellos resultados del individuo de su interés. Pese a que entre un 11-17 % de consultas realizadas por usuarios contienen un nombre de persona (Artiles et al., 2010), los motores de búsqueda más conocidos (Google, Yahoo!, Bing) solo proveen herramientas de desambiguación para las celebridades mediante sus grafos de conocimiento (*knowledge graphs*). Por otra parte, recientemente han aparecido varios buscadores de pago especializados en buscar personas (spokeo.com, pipl.com, intelius.com), lo que de-

* Este trabajo ha sido subvencionado por el Ministerio de Ciencia e Innovación [MED-RECORD Project, TIN2013-46616-C2-2-R] y el grupo CVIP de la URJC.

 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

muestra el impacto de este problema en Internet.

Una de las mayores dificultades de este problema reside en que la temática tratada en las páginas web de un mismo individuo puede ser heterogénea. Por ejemplo, en páginas profesionales normalmente encontramos información laboral de una persona, mientras que en blogs o perfiles de redes sociales es habitual encontrar información personal, opiniones, hobbies, etc. Debido a la irrupción de las plataformas de redes sociales en los últimos años, cuando consultamos un nombre de persona en un motor de búsqueda es bastante habitual obtener enlaces a perfiles de este tipo de plataformas. A pesar de ello, la mayoría de los sistemas del estado del arte no tienen en cuenta este factor puesto que han sido evaluados en corpora que contienen un número reducido de webs de este tipo. La aparición de este tipo de páginas web puede afectar negativamente en este problema y, por tanto, deben ser tratadas de manera especial (Berendsen, 2015).

La principal contribución de este trabajo consiste en la propuesta de dos heurísticas para tratar las páginas de redes sociales en la desambiguación de nombres de personas en la Web. Hemos evaluado nuestras heurísticas en dos corpora de desambiguación de nombres de personas que incluyen resultados de redes sociales. En ambas colecciones, el uso de nuestras aproximaciones mejora los resultados obtenidos cuando no se tienen en cuenta este tipo de páginas web.

El resto del artículo se organiza del siguiente modo. En la sección 2, comentamos brevemente el estado del arte en desambiguación de nombres de personas. A continuación, en la sección 3 presentamos el algoritmo de clustering que hemos utilizado para agrupar las páginas web según el individuo al que se refieren. Posteriormente, la sección 4 presenta nuestras propuestas de tratamiento de las redes sociales en este problema. La sección 5 presenta y analiza los resultados obtenidos. Finalmente, en la sección 6 se presentan conclusiones y líneas de trabajo futuro.

2 *Estado del Arte*

Las campañas de evaluación WePS[1] (*Web People Search*) plantearon el problema de desambiguación de personas en la Web, publicando varios corpora anotados. Este marco de evaluación se ha convertido en un referente, puesto que ha permitido realizar estudios comparativos sobre el rendimiento de diferentes sistemas.

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

La desambiguación de nombres de personas en la web se ha tratado como un problema de clustering en el estado del arte, donde el objetivo es estimar el número de individuos diferentes mencionados en el ranking de páginas web, y organizar en grupos dichos resultados según el individuo particular al que se refieren. Los sistemas propuestos dividen el problema en dos pasos: (1) representación de páginas web, donde el objetivo es seleccionar rasgos adecuados para representar las páginas web; (2) aplicar un algoritmo de clustering para agrupar los resultados.

En cuanto a la representación de las páginas web, los sistemas más competitivos han usado el modelo de espacio vectorial. Algunos trabajos (Balog et al., 2009; Grüetze, et al., 2014) han concluido que el uso de modelos probabilísticos logran resultados más pobres. Los rasgos más utilizados han sido bolsas de palabras, Entidades Nombradas (ENs) y sintagmas nominales. Según (Artiles, Amigó y Gonzalo, 2009a) el uso de rasgos lingüísticos como las ENs, no otorgan ventajas sustanciales con respecto a usar rasgos que no requieren pre-procesamientos lingüísticos. Por otra parte, algunos autores (Nuray-Turan, Kalashnikov y Mehrotra, 2012; Delgado et al., 2014a) destacan la precisión de los n-gramas a la hora de agrupar adecuadamente las páginas web. Finalmente, algunos sistemas competitivos (Chen, Yat Mei Lee y Huang, 2012; Nuray-Turan, Kalashnikov y Mehrotra, 2012; Xu et al., 2015) enriquecen la representación tomando tokens de las URLs, snippets, extrayendo información de Wikipedia, realizando consultas adicionales a un buscador, o aplicando extracción de atributos para conseguir datos biográficos de los individuos.

En cuanto a los algoritmos de clustering utilizados, las campañas WePS (Artiles et al., 2010) destacan que sus mejores participantes han usado métodos basados en el algoritmo jerárquico aglomerativo (HAC). Esta conclusión se ha visto corroborada en trabajos posteriores (Liu, Lu y Xu, 2011; Xu et al., 2015), donde se presentan sistemas que obtienen mejores resultados y están basados en versiones de este algoritmo. La mayoría de los anteriores sistemas requieren de datos de entrenamiento para obtener un valor de umbral que corte el dendograma devuelto por HAC. Sin embargo, el comportamiento de HAC es muy sensible a dicho valor y puede conllevar resultados sesgados según la naturaleza del corpus de entrenamiento. Finalmente, otros trabajos

(Delgado et al., 2014a; Xu et al., 2015) presentan sistemas que evitan el uso de datos de entrenamiento. En particular, (Delgado et al., 2014a) presenta el algoritmo de clustering UPND, basado en la compartición de n-gramas y una función de umbral adaptada a los documentos que se comparan, y cuyo uso evita la necesidad de datos de entrenamiento de HAC.

Las colecciones proporcionadas por WePS contienen un número muy pequeño de páginas web correspondientes a redes sociales. Sin embargo, es común que este tipo de páginas web aparezcan cuando se consulta un nombre de persona en un buscador. En (Berendsen, 2015), se estudia el impacto de este tipo de páginas web, concluyendo que su aparición puede llevar a obtener agrupaciones incorrectas y deben ser tratadas de manera diferenciada. Propone un método que distingue las webs sociales del resto, agrupando de forma separada las páginas web no sociales y las sociales. Las páginas no sociales se agrupan mediante HAC aplicando un valor de umbral obtenido mediante datos de entrenamiento, mientras que las páginas sociales se dejan en clusters unitarios. Finalmente, propone un algoritmo de mezcla de ambos grupos basado en penalizar aquellos clusters que contienen páginas sociales. Su propuesta la prueba utilizando un nuevo corpus que contiene un número considerable de páginas web sociales.

3 Algoritmo de Clustering

El algoritmo de clustering que hemos usado para agrupar las páginas web consiste en una extensión del método UPND presentado en (Delgado et al., 2014a), el cual se basa, por un lado, en la compartición de n-gramas largos compuestos por palabras escritas en mayúsculas y, por otro lado, en el uso de funciones que computan automáticamente un umbral cuando se comparan dos páginas web. En la Sección 3.1 detallamos los rasgos adicionales que utiliza nuestra propuesta para mejorar la representación de las páginas web. A continuación, en la Sección 3.2 presentamos una nueva función de umbral y, finalmente, el método propuesto se describe en la Sección 3.3.

3.1 Representación de páginas web

Puesto que tomamos como punto de partida el algoritmo UPND, nuestro método asume las dos hipótesis sobre representación de las páginas web de este algoritmo: (H1) La coaparición de n-gramas permite decidir si dos documentos hablan de un mismo individuo. Además, cuanto mayor sea el valor n, más probable es la afirmación anterior. (H2) Las palabras en mayúsculas aportan información especialmente útil a la hora de desambiguar entre diferentes individuos. Combinando ambas hipótesis, se asume que la coaparicón de n-gramas en mayúsculas es un buen indicador para decidir si dos documentos se refieren al mismo individuo. Hemos añadido una hipótesis adicional: (H3) Dos páginas web de un ranking hablan del mismo individuo si están enlazadas entre sí, esto es, una de ellas contiene como link la URL de la otra.

Una limitación de la combinación de las hipótesis (H1) y (H2) es que quedan páginas web sin representar o infrarepresentadas, por ejemplo, aquellas escritas principalmente en minúsculas. Para evitar este problema, tras aplicar UPND, se ejecutan dos fases en las que los documentos se representan respectivamente mediante 1-gramas de palabras mayúsculas y 1-gramas de todas las palabras.

3.2 Umbrales Adaptativos

A la hora de comparar dos páginas web, el algoritmo UPND emplea una función de umbral que depende únicamente del contenido de las páginas web con el objetivo de evitar el cálculo de umbrales mediante datos de entrenamiento. Las funciones propuestas en (Delgado et al., 2014a; Delgado et al., 2014b) no dependen del valor n de los n-gramas, pese a que se asume que cuanto más largos sean los que comparten dos páginas web, mayor es la probabilidad de que ambas hablen del mismo individuo. Por ello, para cumplir formalmente la hipótesis (H1) proponemos una nueva función de umbral, de manera que decrece el umbral si el valor n de los n-gramas aumenta:

$$\gamma(W_i^n, W_j^n) = \frac{\gamma_{max}(W_i^n, W_j^n) + \gamma_{min}(W_i^n, W_j^n)}{2 \cdot n} \quad (1)$$

donde W_i es una página web, W_i^n denota a su bolsa de n-gramas asociada y

$$\gamma_{max}(W_i^n, W_j^n) = \frac{min(|W_i^n|, |W_j^n|) - |W_i^n \cap W_j^n|}{max(|W_i^n|, |W_j^n|)} \quad (2)$$

$$\gamma_{min}(W_i^n, W_j^n) = \frac{min(|W_i^n|, |W_j^n|) - |W_i^n \cap W_j^n|}{min(|W_i^n|, |W_j^n|)} \quad (3)$$

Dada una función de similitud sim, la *condición de agrupamiento* empleada por el algoritmo UPND para agrupar dos páginas web W_i y W_j es la siguiente: $sim(W_i^n, W_j^n) > \gamma(W_i^n, W_j^n)$. Cuando se cumple la anterior condición de agrupamiento el algoritmo une los clusters a los que pertenecen ambos documentos, por lo que UPND agrupa los documentos de manera transitiva.

3.3 Algoritmo Propuesto

El algoritmo propuesto se muestra en Algoritmo 1. Inicialmente, se agrupan las páginas web que están enlazadas comparando sus links con sus URLs (según H3) mediante el método $groupByLinks$. A continuación, se aplica el algoritmo UPND tomando 3-gramas en mayúsculas. Posteriormente se ejecutan dos fases adicionales que usan 1-gramas para representar las páginas web, evitando así el problema de baja representación de páginas web de UPND. La primera fase extra agrupa los clusters obtenidos previamente usando 1-gramas en mayúsculas. Esto se justifica por dos razones: (i) como los rasgos usados por UPND son muy discriminantes, se asume que puede devolver varios clusters que se refieren al mismo individuo y (ii) se toman rasgos en mayúsculas asumiendo la hipótesis (H2) de UPND. Finalmente, la segunda fase extra agrupa las páginas web que no se han agrupado con anterioridad (*isolated pages*, conjunto I), tomando como rasgos todos los 1-gramas. Para cada *isolated page*, se calcula su similitud con los clusters no-unitarios existentes (conjunto C_{aux}), y se agrupa en el más similar tal que cumpla la condición de agrupamiento de UPND (método $bestCluster$). En caso de no agruparse en ningún cluster, la propia página web *isolated* se trata como un cluster más, de manera que se permite que las páginas *isolated* puedan agruparse entre sí.

Las fases adicionales comparan clusters con clusters y páginas web *isolated* con clusters. La representación de los clusters consiste en una bolsa de palabras, al igual que las páginas web. Para obtener la bolsa de palabras asociada a un cluster, se calcula su centroide y posteriormente se filtran aquellos rasgos no representativos. Se asume que los rasgos representativos de un cluster son aquellos tales que: (i) aparecen en muchos documentos del cluster (tienen un alto valor de frecuencia de documento dentro del cluster (DF)) y (ii) aparecen en pocos clusters (tienen un alto valor de frecuencia inversa por cluster (ICF)). Para obtener los rasgos representativos de un cluster se sigue el siguiente proceso: se calcula el valor DF*ICF de todos los rasgos y se obtiene la mediana de todos esos valores. Finalmente, se filtran del cluster aquellos rasgos cuyos valores DF-ICF no superen la mediana, esto es, los que no son representativos. La elección de la mediana se justifica porque se trata de un estadístico que no es sensible a casos extremos. Dado un cluster C_k, denotamos como CT_k a su centroide obtenido de esta manera.

Algoritmo 1 $ExtendedUPND(\mathcal{W}, sim, \gamma)$

Entrada: Conjunto de páginas web $\mathcal{W}$= $\{W_1, W_2, ..., W_N\}$, medida de similitud sim y función de umbral γ.
Salida: Conjunto de clusters $\mathcal{C} = \{C_1, C_2, ..., C_l\}$

```
// Agrupar páginas web enlazadas
C = groupByLinks(W)
// Algoritmo UPND
para i = 1 to N hacer
    para j = i + 1 to N hacer
        si sim(W_i^3M, W_j^3M) > γ(W_i^3M, W_j^3M)
            C_i = C_i ∪ C_j
            C = C \{C_j}
        fin si
    fin para
fin para
// FASE extra 1: Agrupación de clusters
para k = 1 to |C| hacer
    para l = k + 1 to |C| hacer
        si sim(CT_k^1M, CT_l^1M) > γ(CT_k^1M, CT_l^1M)
            C_k = C_k ∪ C_l
            C = C \{C_l}
        fin si
    fin para
fin para
// FASE extra 2: Agrupación páginas isolated
I = {W ∈ W : ∃C ∈ C : C = {W}}
Caux = {C ∈ C : |C| > 1}
para W_I ∈ I hacer
    C_sim = bestCluster(W_I, Caux)
    si C_sim ≠ ∅
        C_sim = C_sim ∪ {W_I}
        C = C \{{W_I}}
    si no
        C_aux = C_aux ∪ {W_I}
    fin si
fin para
devolver C
```

4 Tratamiento de páginas web sociales

En los últimos años se ha elevado de forma significativa el número de usuarios que utilizan asiduamente redes sociales. Por ejemplo, Facebook[2] sobrepasa los 1000 millones de usuarios activos mensuales mientras que Twitter[3] sobrepasa los 300 millones. De ahí que los buscadores devuelvan habitualmente páginas web perte-

[2]https://newsroom.fb.com/company-info/
[3]https://about.twitter.com/company

necientes a redes sociales cuando la consulta se corresponde con un nombre de persona.

(Berendsen, 2015) fue el primero en estudiar el impacto de las redes sociales en la desambiguación de nombres de personas en la Web y concluye que la aparición de páginas sociales puede provocar que los sistemas del estado del arte realicen agrupaciones incorrectas. Su propuesta consiste en tratar a las páginas sociales de manera diferenciada bajo el supuesto de que cada página web social se refiere a un individuo diferente. Propone aplicar la política *one in one* sobre las webs sociales, lo cual significa mantener cada una de estas páginas web en un cluster unitario. No obstante, esta asunción tiene un par de limitaciones: (i) un individuo puede tener cuenta de usuario en varias redes sociales y (ii) un individuo puede tener varias cuentas de usuario en una misma red social.

En este trabajo proponemos dos heurísticas para tratar las páginas web sociales que corrigen las limitaciones de la política *one in one*. Ambas heurísticas requieren conocer si una página web pertenece a una red social y, en caso afirmativo, a cuál en concreto. Para ello, se ha tomado una lista de redes sociales de Wikipedia[4]. Tomando el dominio de una cierta página web a través de su URL, se puede comparar si se corresponde con algún dominio de la lista de redes sociales y, en caso afirmativo, conocer qué red social es.

4.1 P1: *One in one per social network*

Esta heurística asume que en un ranking las páginas web sociales correspondientes a una misma red social se refieren a individuos diferentes, porque suelen corresponderse con perfiles de usuario y además en un ranking no se repiten páginas. Esta heurística no permite la comparación de páginas web pertenecientes a la misma red social, pero sí permite la comparación de webs sociales de diferente red social, corrigiendo la primera limitación de la política *one in one*. Puesto que el algoritmo UPND agrupa páginas web por transitividad, varias páginas web de una misma red social pueden finalmente acabar en un mismo cluster, solucionando la segunda limitación. Si dos páginas web de la misma red social se agrupan por separado con una tercera página web de otro dominio, entonces las tres pertenecerán al mismo cluster. La aplicación de la heurística en el algoritmo consiste en evitar comparaciones de páginas web sociales de la misma red social.

[4]en.wikipedia.org/wiki/Category:Social_networking_services

4.2 P2: Eliminación de rasgos comunes

Esta heurística asume que muchas agrupaciones incorrectas con páginas sociales se deben a la compartición de vocabulario común de estas plataformas. En un escenario multilingüe, esto se cumple para las páginas web escritas en el mismo idioma. El tratamiento de las redes sociales de esta heurística es el siguiente: se forman grupos de páginas web sociales según la red social a la que pertenezcan e idioma en el que están escritas. En los grupos en los que existan al menos dos páginas web, se calculan qué rasgos aparecen en la mayoría de ellas y se eliminan de esas páginas web, asumiendo que se trata de vocabulario específico de dicha red social. Hemos tomado la política de eliminar aquellos rasgos que aparezcan en al menos el 75 % de las páginas web de un grupo. Esta heurística no impone restricciones a la hora de comparar páginas web sociales entre sí, de manera que se permite tanto la comparación de páginas web de distinta red social, como páginas web de la misma red social. Para poder aplicar esta heurística, se efectúa la eliminación de los rasgos de cada grupo antes de aplicar el algoritmo.

5 Experimentación

En esta sección presentamos las colecciones de páginas web utilizadas. Posteriormente, presentamos y analizamos los resultados obtenidos comparándonos con el otro sistema del estado del arte que hace un tratamiento diferenciado de las páginas web sociales.

5.1 Corpora de evaluación

Los corpora que hemos utilizado se caracterizan por contener un número significativo de páginas web sociales. Por esta razón, no hemos utilizado las colecciones de referencia en la tarea de las campañas de evaluación WePS. Estas colecciones contienen un pequeño número de webs sociales, y en particular, en WePS-2, los organizadores no consideraron para la evaluación de los resultados (Artiles, Gonzalo y Sekine, 2009b).

Las colecciones utilizadas son el corpus ECIR2012[5] y un nuevo corpus denominado MC4WePS[6]. La Tabla 1 muestra el número de nombres de personas y de documentos contenidos en las colecciones, el porcentaje de páginas web sociales y los porcentajes de nombres muy ambiguos y poco ambiguos. Se ha considerado

[5]http://ilps.science.uva.nl/resources/ecir2012rdwps/

[6]http://nlp.uned.es/web-nlp/resources

que un nombre es muy ambiguo si en las páginas web se hace referencia a más de 10 individuos diferentes. En caso contrario, hemos considerado que el nombre es poco ambiguo.

Dato	ECIR2012	MC4WePS
#Nombres	33	100
#Docs	3487	10432
%Social	34.73 %	8.36 %
%MuyAmbiguos	81.82 %	51.00 %
%PocoAmbiguos	18.18 %	49.00 %

Tabla 1: Datos de los corpora ECIR2012 y MC4WePS

Los nombres de persona contenidos en el corpus ECIR2012 son neerlandeses y las páginas web están escritas en dicho idioma. Los resultados de búsqueda incluidos en el corpus fueron devueltos por varios buscadores, concretamente Google, Yahoo! y Bing, de manera que no son rankings reales devueltos tras consultar cada nombre de persona. Esta colección fue construida para estudiar el impacto de las redes sociales en el problema, por lo que se añadieron artificialmente páginas de las redes sociales Facebook, Hyves[7], LinkedIn, Twitter y MySpace. Por esta razón, este corpus contiene un porcentaje muy elevado de webs sociales. La mayoría de los nombres de persona incluidos en esta colección son muy ambiguos.

En el caso de MC4WePS se incluyen nombres de persona de origen diverso, aunque mayoritariamente anglosajón e hispano. Los datos de esta colección fueron recopilados sobre los ranking de links devueltos por el buscador Google al realizar las consultas, de manera que se trata de resultados reales. Además, este corpus se caracteriza por contener páginas web escritas en diferentes idiomas, al contrario que las colecciones de referencia para este problema. Los anotadores identificaron páginas web escritas en 30 idiomas diferentes, prevaleciendo el inglés y el castellano (96.08 % entre ambos idiomas). Por otra parte, en este corpus está equilibrado el número de nombres muy ambiguos y poco ambiguos.

5.2 Resultados

En esta sección presentamos los resultados obtenidos por nuestro algoritmo para las dos colecciones descritas anteriormente. Las métricas utilizadas son las B^3-Cubed (Bagga y Baldwin, 1998): precisión (BEP), recall (BER) y medida-F ($F_{0,5}$). Estas métricas son adecuadas para evaluar sistemas de desambiguación de nombres de persona (Artiles, Gonzalo y Sekine, 2009b). Para la experimentación se han pesado los rasgos utilizando TF-IDF y se ha empleado la similitud coseno para comparar las páginas web.

La Tabla 2 presenta los resultados obtenidos sobre todas las páginas web y las páginas web sociales por el algoritmo propuesto por (Berendsen, 2015), al que hemos denominado BEREN, y por nuestra propuesta teniendo en cuenta varias configuraciones diferentes. En el caso del algoritmo BEREN, se presentan los resultados obtenidos por la mejor y la peor de sus configuraciones, denotados por BERENbest y BERENworse respectivamente. La primera penaliza la presencia de redes sociales a la hora de mezclar clusters no sociales con clusters sociales, mientras que la segunda no aplica dicha penalización. En el caso de nuestro algoritmo, presentamos los resultados obtenidos sin aplicar ningún tratamiento de redes sociales (ExtendedUPND) y aplicando nuestras dos heurísticas de tratamiento de páginas web sociales (ExtendedUPND+P1 y ExtendedUPND+P2). Por otra parte, se han incluido los resultados obtenidos por el algoritmo UPND. La tabla también muestra información relativa a la significancia estadística de los resultados calculada mediante el test de Wilcoxon (Wilcoxon, 1945) con un nivel de confianza del 95 %, tomando los pares de valores $F_{0,5}$ de cada nombre de persona. Para cada columna, los experimentos se marcan con (k) donde $k \in \mathbb{N}$, de modo que dos experimentos con la misma marca obtienen resultados similares, y un experimento marcado con (k) obtiene mejoras signiticativas sobre otro marcado con (l) si $k < l$.

Los resultados muestran que en el caso del corpus ECIR2012, las configuraciones de los algoritmos que tratan de manera especial a las webs sociales obtienen mejores resultados con respecto a las que no lo hacen. Esto corrobora lo concluido por (Berendsen, 2015) respecto al papel de este tipo de páginas web en este problema. En el caso de la colección MC4WePS, los resultados obtenidos son muy similares entre sí, lo cuál puede explicarse por el menor porcentaje de páginas web sociales presentes en dicho corpus con respecto a ECIR2012 (ver Tabla 1), por lo que el impacto de las redes sociales es mucho menor en esta colección.

[7] Se trata de una red social similar a Facebook, popular en los Países Bajos.

Corpus	ECIR2012						MC4WePS					
Webs	Todas las webs			Webs sociales			Todas las webs			Webs sociales		
Ejecución	BEP	BER	$F_{0,5}$	BEP	BER	$F_{0,5}$	BEP	BER	$F_{0,5}$	BEP	BER	$F_{0,5}$
BERENbest	0.90	0.80	0.83 (1)	1.00	0.79	0.87 (1)	0.91	0.43	0.50 (3)	0.99	0.68	0.77 (2)
BERENworse	0.74	0.82	0.76 (2)	0.55	0.85	0.62 (5)	0.91	0.43	0.50 (3)	0.99	0.68	0.77 (2)
ExtendedUPND	0.72	0.75	0.72 (3)	0.45	0.87	0.55 (6)	0.83	0.76	0.77 (1)	0.83	0.80	0.77 (2)
ExtendedUPND+P1	0.92	0.70	0.78 (2)	0.92	0.77	0.82 (2)	0.87	0.74	0.78 (1)	0.92	0.77	0.81 (1)
ExtendedUPND+P2	0.86	0.72	0.77 (2)	0.73	0.83	0.75 (3)	0.86	0.74	0.78 (1)	0.90	0.79	0.80 (1)
UPND	0.80	0.70	0.73 (3)	0.64	0.84	0.70 (4)	0.86	0.67	0.72 (2)	0.84	0.79	0.77 (2)

Tabla 2: Resultados de los algoritmos BEREN, ExtendedUPND y UPND para las colecciones ECIR2012 y MC4WePS sobre todas las páginas web y únicamente las páginas web de redes sociales.

En cuanto a la comparación de los algoritmos de clustering, vemos que en el corpus ECIR2012, la mejor configuración del algoritmo BEREN obtiene mejoras significativas con respecto al resto de ejecuciones, pero las dos configuraciones de este algoritmo obtienen resultados muy pobres en el corpus MC4WePS. Esto se explica porque el rendimiento de HAC depende fuertemente del valor del umbral usado como criterio de agrupamiento (Artiles, Gonzalo y Sekine, 2009b). UPND y ExtendedUPND evitan este problema gracias al uso de la función de umbral. Por otra parte, las diferencias en el grado de ambigüedad de ambos corpora también influyen en los resultados. Por un lado, ECIR2012 se compone de muchos nombres muy ambiguos, mientras que MC4WePS contiene un mayor equilibrio entre nombres poco ambiguos y muy ambiguos (ver Tabla 1), de manera que el umbral que recibe HAC en el sistema BEREN, con valor 0.225, funciona mejor con nombres muy ambiguos (con más clusters), pero mucho peor para nombres poco ambiguos (con menos clusters). En cambio, tanto UPND como ExtendedUPND son menos sensibles al grado de ambigüedad de los nombres de persona contenidos en ambas colecciones. Por otra parte, la penalización aplicada en el algoritmo de mezcla de webs sociales y no sociales de (Berendsen, 2015) no tiene efecto en el corpus MC4WePS, porque incluye un menor número de redes sociales y el umbral que usan en dicha fase, $\tau = 0{,}5$, es demasiado estricto, de modo que en ambas ejecuciones se agrupan el mismo número de webs sociales con webs no sociales. Finalmente, ExtendedUPND, además de corregir los defectos de representación de UPND, logra un mejor equilibrio entre precisión y recall, y en el caso de MC4WePS consigue mejoras significativas con respecto a UPND.

La tabla muestra que aplicando las heurísticas propuestas sobre todas las webs se obtienen resultados similares en los dos corpora, obteniendo mejoras significativas en ECIR2012 respecto UPND y ExtendedUPND. Además, ambas mejoran significativamente los agrupamientos de páginas sociales respecto no aplicarlas. Su efecto consiste en mejorar los resultados de precisión sin alterar drásticamente los valores de cobertura con respecto a no aplicarlas, lo que significa que evitan agrupamientos incorrectos. La heurística P2 obtiene resultados de precisión más bajos en el corpus ECIR2012. Esto se debe a que P2 tiende a agrupar páginas sociales de la misma red social en el mismo cluster pese a la eliminación de rasgos comunes. Este tipo de agrupamientos normalmente son incorrectos en ambos corpora puesto que se corresponden con perfiles de distintos individuos tal y como presupone la heurística P1. En cambio, en el corpus MC4WePS ambas heurísticas se comportan de manera similar. Esto se debe a que la mayoría de grupos de webs sociales por red social e idioma formados por P2 se componen de dos páginas web, de modo que se eliminan todos los rasgos comunes de las páginas de la misma red social, lo que imposibilita su agrupación como sucede con P1. Lo anterior indica que la suposición *one in one per social network* es adecuada, por lo que P1 es preferible a la hora de tratar las redes sociales.

6 Conclusiones y Trabajo Futuro

En este trabajo hemos presentado dos heurísticas para tratar las páginas web pertenecientes a redes sociales en el problema de desambiguación de nombres de personas. Además, hemos utilizado una extensión del algoritmo de clustering UPND que tiene en cuenta si las páginas están enlazadas, y que permite representar un mayor número de páginas web que el algoritmo original. Por un lado, nuestras heurísticas obtienen mejoras en el corpus ECIR2012 compuesto por más páginas

web sociales que MC4WePS. El efecto conseguido consiste en mejorar los resultados de precisión de los agrupamientos, sin que esto implique una caída drástica en los valores de cobertura, de manera que se evitan agrupaciones incorrectas. Por otro lado, ExtendedUPND ofrece mejores resultados que el algoritmo BEREN en el corpus MC4WePS. Este algoritmo es más independiente del grado de ambigüedad de los nombres respecto al método BEREN, cuyo rendimiento depende del umbral obtenido mediante entrenamiento.

Como trabajo futuro, proponemos realizar un tratamiento distintivo sobre los resultados que se corresponden con entradas de Wikipedia. Varios autores (Long y Shi, 2010; Xu et al., 2015) han destacado que la información proporcionada por esta enciclopedia online sirve de gran ayuda a la hora de diferenciar entre distintos individuos.

Bibliografía

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. En *Proceedings of SemEval-2007*, pages 64-69. ACL.

Artiles, J., E. Amigó, and J. Gonzalo. 2009a. The Role of Named Entities in Web People Search. En *Proceedings of EMNLP 2009*.

Artiles, J., J. Gonzalo, and S. Sekine. 2009b. Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. En *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. En *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. En *Proceedings of the COLING/ACL'98 - Volume 1*, pages 79-85.

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. En *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Richard Berendsen 2015. Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation. PhD Thesis. Informatics Institute, University of Amsterdam.

Chen. Y., Yat Mei Lee, S., and Huang, C.R. 2012. A Robust Web Personal Name Information Extraction System. En *Expert Systems with Applications*, Vol. 32, Issue 3, pp. 2690-2699.

Delgado, A. D, R. Martínez, V. Fresno, and S. Montalvo. 2014. A Data Driven Approach for Person Name Disambiguation in Web Search Results. En *Proceedings of COLING 2014*, pages 301-310.

Delgado, A. D, R. Martínez, S. Montalvo, and V. Fresno. 2014. An Unsupervised Algorithm for Person Name Disambiguation in the Web. En *Procesamiento del Lenguaje Natural, 53*, pages 51-58.

Grüetze, T., Kasneci, G., Zuo, Z., and Naumann, F. 2014. Bootstrapping Wikipedia to answer ambiguous person name queries. En *Proceedings of the 30th International Conference on Data Engineering Workshops (ICDE)*, pages 56-61. Chicago, IL, USA.

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

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

Nuray-Turan, R., Kalashnikov, D. V., and Mehrotra S. 2012. Exploiting Web querying for Web People Search. *ACM Transactions on Database Systems (TODS)*, Vol. 37, Issue 1.

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

Xu, J., Lu, Q., Li, M., and Li, W. 2015. Web Person Disambiguation Using Hierarchical Co-Reference Model. En *Proceedings of CICLing 2015, Part I, pages 279-291*.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 125-132 recibido 30-03-2016 revisado 21-04-2016 aceptado 03-02-2016

Using Personality Recognition Techniques to Improve Bayesian Spam Filtering

Uso de Técnicas de Reconocimiento de la Personalidad para Mejorar el Filtrado Bayesiano de Spam

Enaitz Ezpeleta, Urko Zurutuza
Mondragon University
Goiru Kalea, 2
20500 Arrasate-Mondragón, Spain
{eezpeleta, uzurutuza}@mondragon.edu

José María Gómez Hidalgo
Pragsis Technologies
Manuel Tovar, 43-53, Fuencarral
28034 Madrid, Spain
jmgomez@pragsis.com

Resumen: Millones the usuarios se ven afectados por las campañas de envío de correos eléctrinicos no deseados al día. Durante los últimos años diferentes técnicas de detección de spam han sido desarrollados por investigadores, obteniendo especialmente buenos resultados con algoritmos de aprendizaje automático. En este trabajo presentamos una base para un nuevo metodo de filtrado de spam. Durante el estudio hemos validado la hipótesis de que las técnicas de reconocimiento de personalidad pueden ayudar a mejorar el filtrado Bayesiano de spam. Usando estas técnicas de filtrado, añadimos la característica de personalidad a cada correo, y después comparamos los resultados del filtrado Bayesiano de spam con y sin personalidad, analizando los resultados en terminos de exactitud. En un segundo experimento, combinamos las características de personalidad y polaridad de cada mensaje, y comparamos los resultados. Al final, conseguimos mejorar los resultados del filtrado Bayesiano de spam, alcanzando el 99,24% de exactitud, y reduciendo el número de falsos positivos.
Palabras clave: spam, personalidad, polaridad, PLN, seguridad

Abstract: Millions of users per day are affected by unsolicited email campaigns. During the last years several techniques to detect spam have been developed, achieving specially good results using machine learning algorithms. In this work we provide a baseline for a new spam filtering method. Carrying out this research we validate our hypothesis that personality recognition techniques can help in Bayesian spam filtering. We add the personality feature to each email using personality recognition techniques, and then we compare Bayesian spam filters with and without personality in terms of accuracy. In a second experiment we combine personality and polarity features of each message and we compare all the results. At the end, the top ten Bayesian filtering classifiers have been improved, reaching to a 99.24% of accuracy, reducing also the false positive number.
Keywords: spam, personality, polarity, NLP, security

1 Introduction

Millions of users per day are affected by unsolicited email campaigns. Spam filters are capable of detecting and avoiding an increasing number of emails, but according to Kaspersky Lab data, the average of spam in email traffic stood at 55.28% in 2015[1]. This mass mailing of unsolicited emails are used both for the sale of products such as online fraud, and it reports billionaire benefits. Thanks to spam campaigns a market share sufficient to enrich a sector devoted to fraudulent activity in achieved. These facts make those types of activities one of the biggest threats to Internet security.

To deal with this problem different spam detection systems have been designed and developed by researchers during the last years, spending on cyber-security technologies over $83.6 billions in 2015[2] for example.

This paper provides a baseline for a new

[1] https://securelist.com/analysis/kaspersky-security-bulletin/73591/kaspersky-security-bulletin-spam-and-phishing-in-2015/

[2] http://www.bloomberg.com/news/articles/2016-01-19/e-mail-spam-goes-artisanal

ISSN 1135-5948
© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

spam filtering method. The objective is to demonstrate that personality recognition of email messages can help in Bayesian spam filtering. In this paper we hypothesize that being spam an email that generally aims at selling services or products, analyzing its meaning, and specially the personality of the spam, can bring similar personality functions such that classification systems are improved.

We take into account the results published by (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) related to Bayesian spam filtering, and we aims to improve them. First of all, applying personality recognition techniques to a dataset we create a new tagged (personality) dataset. Then, we apply the best ten classifiers of the mentioned study to the new dataset and we analyze the obtained results. In the second experiment we combine the best sentiment classifiers used by (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) with personality and a new combined dataset is created. One more time, we apply the best classifier to the new dataset and we compare all the results in order to give our conclusions.

The remainder of this paper is organized as follows. Section 2 describes the previous work conducted in the area of spam filtering, personality recognition, natural language processing and sentiment analysis. Section 3 describes the process of the aforementioned experiments, regarding emails personality recognition and spam filtering using personality feature. In Section 4, the obtained results are presented, showing the results of the different experiments carried out during the study. Finally, we summarize our findings and give conclusions in Section 5.

2 Related Work

2.1 Spam filtering techniques

Different techniques to detect spam have been developed during the last years (Nazirova, 2011). Among all proposed automatic classifying techniques, machine learning algorithms have achieved more success (Cormack, 2007). In (Tretyakov, 2004) the authors obtained precisions up to 94.4% using those type of techniques.

In this study we focus on a specific section of machine learning algorithms; content-based filters. Those filters are based on analyzing the content of the emails in order to split messages in spam or legitimate emails as it is explained in (Sanz, Hidalgo, and Cortizo, 2008). Content-based spam filters can be separated in several types such as heuristic filtering, learning-based filtering and filtering by compression.

A comparison between various existing spam detection methods is presented in (Savita Teli, 2014): rule-based system, IP blacklist, Heuristic-based filters, Bayesian network-based filters, white list and DNS black holes. As a conclusion they define Bayesian based filters as the most effective, accurate, and reliable spam detection method.

Some of the content-based filtering techniques are also studied and analyzed in (Malarvizhi and Saraswathi, 2013), and again, the Bayesian method is selected as the most effective one (classifying correctly the 96.5% of messages). Furthermore, in (Eberhardt, 2015) authors demonstrated that although more sophisticated methods have been implemented, Bayesian methods of text classification are still useful.

2.2 Personality recognition techniques

As authors defined in (Vinciarelli and Mohammadi, 2014) personality is a psychological construct aimed at explaining the wide variety of human behaviours in terms of a few, stable and measurable individual characteristics. As an effort to formalize it, two main models has been defined (Celli and Poesio, 2014): in the first one, called Myers-Briggs personality model (Briggs Myers and Myers, 1980), four dimensions are used to define the personality: Extroversion/Introversion, Thinking/Felling, Judging/Perceiving and Sensing/iNtuition; Meanwhile, in the Big Five 5 (Costa and McCrae, 1992) traits are used to define the personality: Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism.

Personality recognition became a potential tool for Natural Language Processing as it is possible to extract a lot of information about the personality of the authors from every text (Mairesse et al., 2007). Several research in the last years has been published related to personality recognition in blogs (Oberlander and Nowson, 2006), offline texts (Mairesse et al., 2007) or online social networks (Bai, Zhu, and Cheng, 2012; Rangel et al., 2015).

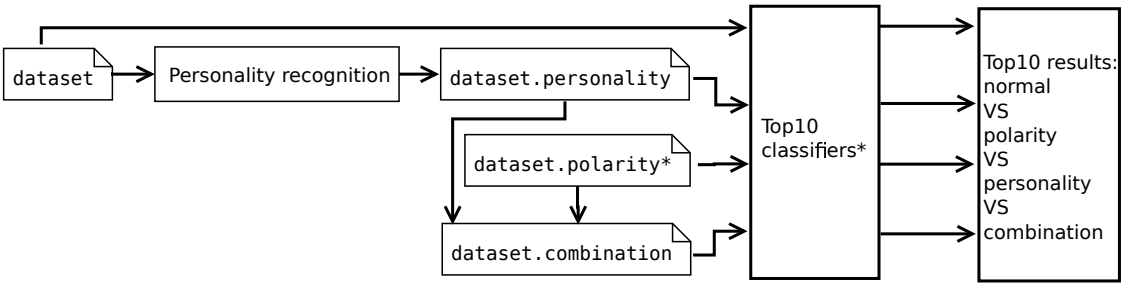


Figure 1: Full process of the study

Email authors personality prediction is possible as it is shown in (Shen, Brdiczka, and Liu, 2013). Authors prove that personality prediction is feasible, and their email feature set can predict personality with reasonable accuracies. This last research is taken into account by the authors as a baseline in spam filtering.

2.3 Sentiment analysis

A brief definition of Natural Language Processing (NLP) is given in (Liddy, 2001), as a theoretical motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis. It aims to achieve human-like language processing for a range of task or applications. Those techniques are becoming more and more useful for spam filtering, as it is demonstrated in (Giyanani and Desai, 2013) using sender information and text content based NLP techniques.

Researchers in (Echeverria Briones et al., 2009) and (Lau et al., 2012) confirmed that it is possible to create an application or a system to detect spam in different formats using text mining techniques and semantic language model respectively.

During the last years Sentiment Analysis (SA) has been used in several research areas, although there has been a continued interest for a while. In (Liu and Zhang, 2012) the most important research opportunities related to SA are described. Based on that, in (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) authors selected document sentiment classification topic as a possible option to improve spam filtering. They tagged a dataset with polarity (positive, neutral or negative) score of each message using sentiment classifiers, and then authors compare spam filtering classifiers with and without the polarity score in terms of accuracy. As the results were positive, authors aim at improving these results adding more semantic features to the text.

3 Design and Implementation

This research has been carried out following the procedure of the figure 1, which is divided in two main experiments.

Taking as a baseline the top ten classifiers identified in (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a), on the one hand, we analyze the influence of the personality in spam filtering comparing the results of the ten classifiers applied to the dataset with and without personality. And on the other hand, we combine personality feature with polarity feature (in the dataset) in order to analyze if it improves Bayesian spam filtering results.

Those experiments are carried out using the 10-fold cross-validation technique and the results are analyzed in terms of false positive rate and accuracy, being the accuracy the percentage of testing set examples correctly classified by the classifier.

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{(Positives + Negatives)}$$

3.1 Dataset

To carry out this study, we use a publicly available dataset called *CSDMC 2010 Spam Corpus*[3]. This dataset is composed by 2,949 legitimate email messages and 1,378 spam messages.

3.2 Bayesian spam filtering

To analyze if personality recognition techniques improve Bayesian spam filtering. First of all we generate Bayesian spam filters as a baseline for the rest of the experiment.

As in (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) the best ten classifiers for spam

[3]http://www.csmining.org/index.php/spam-email-datasets-.html

		Normal		
#	Name	FP	FN	Acc
1	BLR.i.t.c.stwv.go.wtok	13	24	99.15
2	DMNB.c.stwv.go.wtok	21	17	99.12
3	DMNB.i.c.stwv.go.wtok	21	17	99.12
4	DMNB.i.t.c.stwv.go.wtok	21	17	99.12
5	DMNB.stwv.go.wtok	21	17	99.12
6	DMNB.c.stwv.go.stemmer	22	19	99.05
7	DMNB.i.c.stwv.go.stemmer	22	19	99.05
8	DMNB.i.t.c.stwv.go.stemmer	22	19	99.05
9	DMNB.stwv.go.stemmer	22	19	99.05
10	BLR.i.t.c.stwv.go.ngtok.stemmer.igain	14	28	99.03

Table 1: Baseline results

filtering are defined. In table 1, the best results presented in the mentioned study are shown.

During this paper, our main objective is to improve those results using the selected classifiers. To understand the settings of each classifier, table 2 shows the nomenclatures used.

	Meaning
DMNB	DMNBtext
BLR	Bayesian Logistic Regression
.c	idft F, tft F, outwc T
.i.c	idft T, tft F, outwc T
.i.t.c	idft T, tft T, outwc T
.stwv	String to Word Vector
.go	General options
.wtok	Word Tokenizer
.ngtok	NGram Tokenizer 1-3
.stemmer	Stemmer
.igain	Attribute selection using InfoGainAttributeEval

Table 2: Nomenclatures

3.3 Personality recognition

The objective of the next phase is to apply personality recognition technique to each email in order to add this feature to the original dataset and create a new dataset. To do that, we followed the personality recognition process presented in (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016b).

One of the most trusted personality recognition assessment is used in this study: Myers-Briggs personality model. To determine the personality of each emails, it is mandatory to use the four different dimensions of this model: Extroversion/Introversion, Thinking/Feeling, Judging/Perceiving and Sensing/iNtuition. In this case, publicly available machine learning web services for text classification, hosted in *uClassify*[4], are used to calculate each feature. Among all the possibilities offered in this website, we focus on the Myers-Briggs functions developed by Mattias Östmar.

As author explains, each function determines a certain dimension of the personality type according to Myers-Briggs personality model. The analysis is based on the writing style and should not be confused with the Myers-Briggs Type Indicator (MBTI) which determines personality type based on self-assessment questionnaires. Training texts are manually selected based on personality and writing style according to (Jensen and DiTiberio, 1989).

Those are the used functions:

- *Myers-Briggs Attitude:* Analyzes the Extroversion/Introversion dimension.
- *Myers-Briggs Judging Function:* Determines the Thinking/Feeling dimension.
- *Myers-Briggs Lifestyle:* Determines the Judging/Perceiving dimension.
- *Myers-Briggs Perceiving Function:* Determines the Sensing/iNtuition dimension.

Each function returns a float within the range [0.0, 1.0] per each pair of characteristics of the dimension. For example, if we test a certain text and we obtain X value for Sensing, the value for iNtuition is 1-X. Thus, we only record one value per each function: Extroversion, Sensing, Thinking and Judging.

In order to create a new dataset, those four values of each email message are added

[4]https://www.uclassify.com

	Total	Extroversion	Sensing	Thinking	Judging
ham	2949	975	2439	313	1908
spam	1378	591	918	301	915
Percentage(%)					
ham	100	33	83	11	65
spam	100	43	67	22	66

Table 3: Descriptive analysis of the dataset.

to the original dataset. This new dataset is used during the tests to evaluate the influence of the personality in spam filtering. To do that, we apply the top ten classifiers mentioned previously to the original dataset and to the new one, and we compare the results.

3.4 Combination

Once we analyzed the results of the first experiment, in the second part our objective is to explore the possibilities to improve the results published by (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) where authors used polarity feature in Bayesian spam filtering.

We decided to combine both personality and polarity. First, we use the best sentiment classifier defined in the mentioned work and we analyze each email to create a dataset tagged with the polarity of each email. Once we created this dataset, we apply the top ten classifiers in order to obtain the results using the polarity feature. Finally, we create a new dataset adding the personality and the polarity of each email, and we apply the classifiers to compare all the results.

4 Experimental Results

In this Section the results obtained during the previously explained study are shown. To carry out the following experiments the dataset called *CSDMC 2010 Spam Corpus* is used.

4.1 Descriptive analysis

Once the dataset is selected, we perform a descriptive experiment of the dataset. The objective of this step is to analyze the personality features of the authors (spammers and legitimate email writers) applying the previously explained (Section 3.3) personality recognition functions. During this step the personality features are added to the original dataset creating a new tagged dataset, and we extract statistic about the personality. This information is shown in table 3.

Analyzing the data presented in the descriptive table, it is possible to see that there are differences between the emails types. The biggest difference according to Myers-Briggs personality model between spam emails and legitimate emails is given by the Perceiving Function. Taking into account only this dimension, the percentage of *sensing* legitimate emails is 16 point higher than spam emails.

In the next steps different experiments are carried out to see the real influence of personality feature in Bayesian spam filtering.

4.2 Using personality

As we explain in the previous Section, to see if personality improves Bayesian spam filtering, we apply the top ten classifiers to the labelled (personality) dataset, and we compare the results with the results obtained applying the same classifiers to the original dataset.

The results obtained during this experiments are presented in table 4.

	Normal			Personality		
#	FP	FN	Acc	FP	FN	Acc
1	13	24	99.15	14	26	99.08
2	21	17	99.12	22	16	99.12
3	21	17	99.12	22	16	99.12
4	21	17	99.12	22	16	99.12
5	21	17	99.12	22	16	99.12
6	22	19	99.05	22	21	99.01
7	22	19	99.05	22	21	99.01
8	22	19	99.05	22	21	99.01
9	22	19	99.05	22	21	99.01
10	14	28	99.03	13	26	99.10

Table 4: Comparison between normal and personality

Results show that only in one case the previous result is improved (from 99.03% to 99.10%), while in other four cases we obtain the same results (99.12%) and in the other five the results are worst than applying the classifiers to the original dataset.

So, adding the four personality dimensions to the dataset it is not helpful. But if we take into account the information obtained in the descriptive part, we can see that the

	Normal		Polarity		Sensing		Combination	
#	FP	Acc	FP	Acc	FP	Acc	FP	Acc
1	13	99.15	14	99.12	15	99.03	15	99.03
2	21	99.12	22	99.21	21	99.12	**19**	**99.24**
3	21	99.12	22	99.21	21	99.12	**19**	**99.24**
4	21	99.12	22	99.21	21	99.12	**19**	**99.24**
5	21	99.12	22	99.21	21	99.12	**19**	**99.24**
6	22	99.05	22	99.15	22	99.08	23	99.05
7	22	99.05	22	99.15	22	99.08	23	99.05
8	22	99.05	22	99.15	22	99.08	23	99.05
9	22	99.05	22	99.15	22	99.08	23	99.05
10	14	99.03	14	99.03	14	99.08	14	99.10

Table 5: Comparison between all techniques

differentiator dimension is Sensing/iNtuition.

4.2.1 Myers-Briggs Perceiving Function

To see if the mentioned dimension affects in the Bayesian spam filtering, a new dataset is created. We use only the Myers-Briggs Perceiving Function in order to add the *sensing* characteristic of each message to the dataset.

The followed procedure is the same than in the previous experiment: we apply the best ten classifiers to the new dataset and we compare the results with the original ones.

Table 6 summarized the new results.

	Normal			Sensing		
#	FP	FN	Acc	FP	FN	Acc
1	13	24	99.15	15	27	99.03
2	21	17	99.12	21	17	99.12
3	21	17	99.12	21	17	99.12
4	21	17	99.12	21	17	99.12
5	21	17	99.12	21	17	99.12
6	22	19	99.05	22	18	99.08
7	22	19	99.05	22	18	99.08
8	22	19	99.05	22	18	99.08
9	22	19	99.05	22	18	99.08
10	14	28	99.03	14	26	99.08

Table 6: Results using *sensing*

In this case we obtain better results in terms of accuracy than using all the dimensions of the Myers-Briggs personality model. The results are improved in five cases, in four of them the same results are obtained, and only in one case the result is worst.

Those results give a baseline to see the possibilities that personality recognition techniques can improve Bayesian spam filtering. But to confirm that the *sensing* characteristic can be helpful, we carry out one more experiment combining personality feature (sensing) with the polarity of each email.

4.3 Combinational experiment

During this experiment we apply the best ten Bayesian classifiers to the following datasets:

- Original dataset.
- Original dataset with the polarity information of each email. The best sentiment classifier identified in (Ezpeleta, Zurutuza, and Gómez Hidalgo, 2016a) is used to calculate the polarity score of each email.
- Original dataset with the *sensing* feature (as in the previous experiment).
- Original dataset with the polarity and the *sensing* feature of each email (combining the two previous dataset).

We compare the obtained results in terms of accuracy and false positive number, as it is possible to see in table 5.

According to the obtained results, we can say that combining sentiment analysis techniques with personality recognition techniques the best result obtained in Bayesian spam filtering is improved in terms of accuracy. The combination improves (99.24% of accuracy) both the top result of the original dataset (99.15%) and the top result of the polarity analysis (99.21%). Moreover, in those cases where the best result is achieved, the combination of sentiment analysis and personality techniques reduces the false positive number.

5 Conclusions

In this work, we give the initial ground for improving spam filtering techniques.

Results show that with the combination of personality recognition techniques (*sensing*) and sentiment analysis techniques allows it is possible to obtain better results than using those techniques separately. This combination obtains the best results within all different experiments reaching to a 99.24% of accuracy, reducing the false positive number from 21 to 19.

Despite the difference in percentage does not seem to be relevant, from 99.15% to 99.24%, if we take into account the amount of real spam traffic, the improvement is significant.

In addition, we conclude that it is possible to improve spam filtering classifiers adding the *sensing* feature to each email message, as in our experiments 5 results out of the best 10 classifiers are improved and in other 4 cases the same result is obtained. Although using the four dimensions of Myers-Briggs personality model the results are not not significant, using a specific characteristic we demonstrate that those techniques are helpful in spam filtering.

Furthermore, this work presents a new filtering method (combining polarity and personality) that gives to the research community the opportunity of detecting non evident intent in spam emails.

Moreover, taking into account that the personality recognition functions used are independent from the text, the use of manually tagged (personality) emails during the learning process of the function might improve the results.

Finally, taking this work as a reference, several directions can be explored: for example, in order to validate those results a repetition of the experiments using a different dataset; more algorithms and filters settings can be used to obtain more results; analyze different types of spam in order to see if the behaviour of the spammers is the same (SMS spam, blog spam,...).

Acknowledgments. This work has been partially funded by the Basque Department of Education, Language policy and Culture under the project SocialSPAM (PI_2014_1_102).

We thank Mattias Östmar for the valuable tools developed and published. And we thank Jon Kågström (Founder of uClassify[5]) for the opportunity to use their API for research purposes.

[5]https://www.uclassify.com

References

Bai, S., T. Zhu, and L. Cheng. 2012. Big-five personality prediction based on user behaviors at social network sites. *CoRR*, abs/1204.4809.

Briggs Myers, I. and P. B. Myers. 1980. Gifts differing: Understanding personality type.

Celli, F. and M. Poesio. 2014. Pr2: A language independent unsupervised tool for personality recognition from text. *arXiv preprint arXiv:1402.2796*.

Cormack, G. V. 2007. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455.

Costa, P. T. and R. R. McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Eberhardt, J. J. 2015. Bayesian spam detection. *Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal*.

Echeverria Briones, P. F., Z. V. Altamirano Valarezo, A. B. Pinto Astudillo, and J. D. C. Sanchez Guerrero. 2009. Text mining aplicado a la clasificación y distribución automática de correo electrónico y detección de correo spam.

Ezpeleta, E., U. Zurutuza, and J. M. Gómez Hidalgo. 2016a. Does sentiment analysis help in bayesian spam filtering? In *Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Sevilla, Spain, April 18-20, 2016*. Springer.

Ezpeleta, E., U. Zurutuza, and J. M. Gómez Hidalgo. 2016b. Short messages spam filtering using personality recognition. In *Proceedings of the 4th Spanish Conference in Information Retrieval*.

Giyanani, R. and M. Desai. 2013. Spam detection using natural language processing. *International Journal of Computer Science Research & Technilogy*, 1:55–58, August.

Jensen, G. H. and J. K. DiTiberio. 1989. *Personality and the Teaching of Composition*, volume 20. Ablex Pub.

Lau, R. Y. K., S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li. 2012. Text mining and probabilistic language modeling for online review spam detection. *ACM Trans. Manage. Inf. Syst.*, 2(4):25:1–25:30, January.

Liddy, E. 2001. Natural language processing. *Encyclopedia of Library and Information Science, 2nd Ed.*, NY. Marcel Decker, Inc.

Liu, B. and L. Zhang. 2012. A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463.

Mairesse, F., M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500, November.

Malarvizhi, R. and K. Saraswathi. 2013. Content-based spam filtering and detection algorithms-an efficient analysis & comparison 1. *International Journal of Engineering Trends and Technology*, Vol. 4, Issue 9, September.

Nazirova, S. 2011. Survey on spam filtering techniques. *Communications and Network*, 3(3):153–160.

Oberlander, J. and S. Nowson. 2006. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 627–634, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rangel, F., F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September.

Sanz, E. P., J. M. G. Hidalgo, and J. C. Cortizo. 2008. Email spam filtering. *Advances in Computers*, pages 45–114.

Savita Teli, S. B. 2014. Effective spam detection method for email. *IOSR Journal of Computer Science*, pages 68–72.

Shen, J., O. Brdiczka, and J. Liu. 2013. Understanding email writers: Personality prediction from email messages. In *User Modeling, Adaptation, and Personalization*. Springer, pages 318–330.

Tretyakov, K. 2004. Machine learning techniques in spam filtering. In *Data Mining Problem-oriented Seminar, MTAT*, volume 3, pages 60–79.

Vinciarelli, A. and G. Mohammadi. 2014. A survey of personality computing. *Affective Computing, IEEE Transactions on*, 5(3):273–291.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 135-138 recibido 14/03/2016 revisado 13/04/2016 aceptado 03/05/2016

Integración de Paradigmas de Traducción Automática (IMTraP)

Integration of Machine Translation Paradigms (IMTraP)

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Resumen: La Traducción Automática (TA) es un campo altamente interdisciplinar y multidisciplinar porque en él trabajan: ingenieros, informáticos, estadísticos y lingüístas. El objetivo de este proyecto es acercar los diferentes perfiles de la comunidad de la TA para plantear un paradigma integrado de TA que incluya tecnologías lingüísticas y estadísticas. Básicamente, nuestra investigación se centra en el problema de integrar dinámicamente dos de los paradigmas de traducción más populares: el basado en reglas y el estadístico. Una de las principales ideas es usar tecnologías lingüísticas desarrolladas para los sistemas basados en reglas o en el contexto del procesamiento del lenguaje natural. El nuevo paradigma proporcionará soluciones a los retos actuales de la TA como palabras desconocidas, reordenamiento y ambigüedades semánticas. El proyecto se focaliza en tres de las lenguas más hablados en el mundo: Chino, Castellano e Inglés; y todas las combinaciones de traducción entre ellas. Estos pares de lenguas no solo involucran intereses económicos y culturales, sino que además tienen importantes retos de TA como el morfológico, sintáctico y semántico.
Palabras clave: Traducción Automática Híbrida, Morfología, Sintaxis, Semántica, Chino, Castellano

Abstract: Machine Translation (MT) is a highly interdisciplinary and multidisciplinary field approached from the point of view of engineering, computer science, informatics, statistics and linguists. The goal of this research project is to approach the different profiles in the MT community by providing a new integrated MT paradigm which mainly includes linguistic technologies and statistical algorithms. Our research focuses on the problem of dynamically integrating the two most popular MT paradigms: the rule-based and the statistical-based. We will use linguistic technologies developed either for the rule-based MT systems or other natural language processing tasks into statistical MT systems. The new paradigm will provide solutions to current MT challenges such as unknown words, reordering and semantic ambiguities. The project focuses on the three most spoken languages in the world: Chinese, Spanish and English; and all translation combinations among them. These language pairs do not only involve many economic and cultural interests, but they also include some of the most relevant MT challenges such as morphological, syntactic and semantic variations.
Keywords: Hybrid Machine Translation, Morphology, Syntax, Semantics, Chinese, Spanish

1 Introducción

La Traducción Automática (TA) es un campo interdisciplinario y multidisciplinario, que incluye profesionales como: traductores, ingenieros, informáticos, matemáticos y lingüístas. Pero la cooperación y la interacción entre ellos es todavía baja. El objetivo de este proyecto es proponer y validar un paradigma de TA completamente nuevo, capaz combinar de manera eficiente el conocimiento lingüístico con los recursos y algoritmos estadísticos. Se pretende combinar las arquitecturas de la TA basada en reglas y la estadística. El resultado del proyecto será un sistema de TA de tecnología híbrida más allá del estado del arte. El sistema resultante deberá ser, en la medida de lo posible, independiente de la lengua y de código abierto.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

2 Objetivos

Los objetivos del proyecto se resumen de la siguiente manera:

- Desarrollar aproximaciones para integrar la información estructural y lingüística (morfológica, sintáctica y semántica) en TA estadística y formular una nueva arquitectura híbrida.
- Integrar las comunidades de TA (especialmente, lingüístas, informáticos e ingenieros) con el fin de resolver los problemas más comunes de TA incluyendo la morfología, la sintaxis y la semántica.
- Analizar y comparar en detalle la estructura y el funcionamiento de los sistemas basados en reglas y los estadísticos.
- Desarrollar un sistema híbrido de TA que sea entrenable (en la medida de lo posible) en cualquier par de lenguas y de código abierto.

3 Resultados de la primera fase del proyecto

Durante la primera fase (12 meses) del proyecto IMTraP, se han logrado los objetivos principales que se detallan a continuación:

- Revisión de los trabajos previos relacionados y revisión exhaustiva de la TA basada en reglas, que incluye la descripción del estado del arte sobre TA híbrida teniendo en cuenta los diferentes niveles lingüísticos: ortografía, morfología, léxico, semántica y sintaxis (Costa-jussà y Farrús, 2014).
- Definición de los sistemas de TA y corpus experimental, que incluye la recopilación de corpus para el chino-castellano e inglés-castellano y experimentos realizados con TA estadística (Costa-jussà, Henríquez Q, y Banchs, 2012).
- Análisis detallado y comparación de los sistemas basados en reglas y estadística, que incluye: el desarrollo del sistema basado en reglas chino-a-castellano construido en el marco de código abierto Apertium[1]; la descripción y análisis de los sistemas basados en reglas y estadísticos; y propuesta de varias arquitecturas híbridas a estudiar más a fondo en la segunda fase del proyecto (Costa-jussà y Centelles, 2015).

Asimismo, en términos de desarrollo de sistemas de TA, en el marco del proyecto IMTraP:

- Se ha construido con éxito los sistemas de TA estadística de referencia para el chino-castellano e inglés-castellano utilizando los datos recogidos y los parámetros sintonizados (Costa-jussà, Henríquez Q, y Banchs, 2012).
- Se ha desarrollado el primer sistema de código abierto basado en reglas chino–castellano, que ha sido construido utilizando técnicas híbridas mediante la combinación de los conocimientos humanos y las técnicas estadísticas. En particular, el conocimiento humano se ha utilizado para los diccionarios monolingües y bilingües así como para la definición de reglas de transferencia estructural. El conocimiento estadístico ha complementado todos los pasos mencionados. Además, el conocimiento estadístico ha sido la única fuente para las reglas de transferencia léxicas. La mejora del conocimiento estadístico en la TA basada en reglas se ha evaluado y se ha demostrado que proporcionan mejoras notables en la salida de la traducción. En este sentido, se han mostrado eficaces técnicas de construcción de un sistema basado en reglas usando técnicas híbridas. Por otra parte, el sistema basado en reglas supera el sistema estadístico en los experimentos fuera del dominio. El nuevo sistema basado en reglas, así como las metodologías para su construcción se han evaluado de forma automática y con un análisis manual. Por otra parte, la salida de la última versión del sistema basado en reglas ha sido contrastada en esos términos con un sistema estadístico estado del arte y usando un texto fuera del dominio. El sistema basado en reglas supera el sistema estadístico a todos los niveles lingüísticos excepto a nivel sintáctico. Hay una gran mejora en términos de cobertura léxica (Costa-jussà y Centelles, 2015).
- Se ha construido un traductor chino-castellano que está disponible como ser-

[1] https://www.apertium.org/

vicio web[2] y como aplicación en Android (Costa-jussà, Centelles, y Banchs, 2014).

4 Nuevas perspectivas en híbridización

Relacionado con el estado de arte en TA estadística con información lingüística, podemos argumentar que la investigación en el campo de la TA hace que el concepto de hibridación no sea estático (Costa-jussà, 2015a).

- Por un lado, los sistemas híbridos son vistos como una combinación de los sistemas estadísticos con sistemas basados en reglas (sentido estricto).
- Por otra parte, existe un creciente interés en la combinación de los conocimientos lingüísticos en todas sus formas (por ejemplo, morfológico, sintáctico y semántico) en los sistemas estadísticos existentes (sentido amplio de la hibridización). Algunos de los problemas encontrados en TA, específicamente en TA basada en segmentos, han sido superados por la incorporación de técnicas que usan conocimiento morfológico (Toutanova, Suzuki, y Ruopp, 2008), sintáctico (Khalilov y Fonollosa, 2011) y semántico (Banchs y Costa-jussà, 2011). El rendimiento de los sistemas de TA puede ser claramente mejorado mediante el uso de tales conocimientos lingüísticos. Sin embargo, la TA todavía no es capaz de cubrir correctamente todas las variedades problemáticas. Alternativamente, en lugar de ser general, cada extensión a TA tiende a centrarse en un desafío particular para lograr la mejora deseada.

5 Planteamiento de arquitecturas híbridas

Existe un largo camino por recorrer hacia una arquitectura con un mayor nivel de hibridación/integración de paradigmas.

5.1 En sentido estricto

Hemos identificado estrategias, interesantes para la comunidad, de construir una arquitectura híbrida dado un sistema basado en reglas y uno estadístico (Costa-jussà, 2015a):

- A partir de un sistema basado en reglas, existe la necesidad de extraer reglas de transferencia de corpus paralelo. Esto permitiría a la construcción de sistemas basados en reglas por un lingüista monolingüe. Por el momento, los sistemas basados en reglas tienen que ser desarrollados por lingüistas nativos bilingües o por lo menos la gente que son competentes en el idioma de origen y destino. Trabajos en el tema incluyen (Sánchez-Cartagena, Pérez-Ortiz, y Sánchez-Martínez, 2015) y se pueden tomar como punto de partida.
- Con el fin de mejorar los sistemas basados en reglas ser más fluido y natural, sería bueno integrar un modelo de lenguaje en la etapa de generación. El modelo de lenguaje puede ser a base de n-gramas com se propone en (Labaka et al., 2014), o la formación basada en el neuronal basada en la sintaxis. En cada caso, se requere una decodificación diferente para ser integrado en el sistema.
- Comenzando con el núcleo de un sistema basado en la estadística, existe la necesidad de integrar reglas de transferencia y acoplarlas en el modelo de traducción. Se pueden añadir reglas de transferencia jerárquicas a los sistemas estadísticos, asegurándose de que se adopta algún formalismo gramatical compatible. En esta línea de investigación, es importante identificar la mejor manera de hacer que la integración: dé prioridad a las reglas de transferencia sobre los estadísticos o extraiga una puntuación para hacer competir la transferencia de reglas en igualdad de condiciones a los estadísticos. En esta línea también se encuentran trabajos precedentes a los que se podría dar continuidad (Labaka et al., 2014).

5.2 En sentido amplio

Además, si ampliamos el alcance del concepto de hibridización como hemos visto en la sección 4, otros enfoques para mejorar los sistemas estadísticos a nivel de la morfología puede realizar la traducción en dos etapas, haciendo más amplia post-edición automática morfológica (Costa-jussà, 2015b). A nivel de la sintaxis, la gramática otro formalismo puede ser experimentado, así como varias extensiones de los sistemas jerárquicos. Por último, en el campo de la semántica, las proyecciones del vector en el espacio, las reducciones de espacio y redes neuronales abre una nueva vía.

[2] http://www.chispa.me

El proyecto IMTraP, en su segunda fase, se está focalizando en estas últimas direcciones (Gupta et al., 2016; Costa-Jussà y Fonollosa, 2016).

Estas son líneas de investigación alentadoras que pueden dar lugar a un acoplamiento más natural de la lingüística y la estadística. Hay muchas preguntas que quedan por resolver incluyendo la forma correcta de aplicación de la investigación mencionada. Sin embargo, las perspectivas son prometedoras y, sin duda, los grandes avances en TA se pueden derivar de colaboraciones multidisciplinares. Al final, esto es de lo que la hibridación trata.

6 Detalles del proyecto

El proyecto IMTraP ha sido financiado por el *Seventh Framework Program of the European Commission* en el contexto del programa denominado *International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951)*, siendo la *investigadora* la Dr. Marta R. Costa-jussà. La primera fase del proyecto (12 meses) se ha desarrollado en el Institute for Infocomm Research bajo la supervisión del Prof Haizhou Li y el Dr. Rafael E. Banchs. La segunda fase del proyecto (12 meses) se está desarrollando en la Universitat Politècnica de Catalunya bajo la supervisión del Prof. José A. R. Fonollosa. Más detalles sobre el proyecto se pueden encontrar en su correspondiente página web[3].

Bibliografía

Banchs, R. E. y M. R. Costa-jussà. 2011. A semantic feature for statistical machine translation. En *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, páginas 126–134.

Costa-jussà, M. R. 2015a. How Much Hybridization Does Machine Translation Need? *Journal of the American Society for Information Technology*, 6(10):2160–2165.

Costa-jussà, M. R. 2015b. Ongoing study for enahncing chinese-spanish translation with morphology strategies. En *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation*, HyTra.

Costa-jussà, M. R. y J. Centelles. 2015. Description of the chinese-to-spanish rule-based machine translation system developed using a hybrid combination of human annotation and statistical techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1):1:1–1:13.

Costa-jussà, M. R., J. Centelles, y R E. Banchs. 2014. A client mobile application for chinese-spanish statistical machine translation. En *Proc. of the Interspeech 2014 Demo Track*.

Costa-jussà, M. R. y M. Farrús. 2014. Statistical Machine Translation enhancements through linguistic levels: A survey. *ACM Computing Surveys*, 46(3):42.

Costa-Jussà, M. R. y J. A. R. Fonollosa. 2016. Character-based neural machine translation. En *Proceedings of the Annual Conference of the Association of Computational Linguistics (ACL)*.

Costa-jussà, M. R., C. A. Henríquez Q, y R. E. Banchs. 2012. Evaluating indirect strategies for chinese-spanish statistical machine translation. *Journal of Artificial Intelligence Research*, 45(1):761–780.

Gupta, P., M. R. Costa-jussà, P. Rosso, y R. E. Banchs. 2016. A deep source-context feature for lexical selection in statistical machine translation. *Patter Recognition Letters*, 75:24–29.

Khalilov, M. y J.A.R. Fonollosa. 2011. Syntax-based reordering for statistical machine translation. *Computer Speech and Language*, 25(4):761–788.

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

Sánchez-Cartagena, V., J. A. Pérez-Ortiz, y F. Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer MT rules from scarce bilingual corpora. *Computer Speech and Language. Special Issue on Hybrid MT: Integration of Linguistics and Statistics*.

Toutanova, K., H. Suzuki, y A. Ruopp. 2008. Applying morphology generation models to machine translation. En *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*, páginas 514–522.

[3] http://cordis.europa.eu/project/rcn/103170_en.html

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 139-142 recibido 14-03-2016 revisado 13-04-2016 aceptado 28-04-2016

DBpedia del gallego: recursos y aplicaciones en procesamiento del lenguaje

Galician DBpedia: resources and applications in language processing

Miguel Anxo Solla Portela
Universidade de Vigo
Grupo TALG
miguelsolla@uvigo.es

Xavier Gómez Guinovart
Universidade de Vigo
Grupo TALG
xgg@uvigo.es

Resumen: En esta presentación, describimos la metodología utilizada para la creación de la DBpedia del gallego y algunas de sus aplicaciones para el procesamiento lingüístico en los ámbitos del reconocimiento de entidades y de la extracción léxica.
Palabras clave: DBpedia, Wikipedia, WordNet, datos enlazados abiertos, web semántica

Abstract: In this presentation, we review the methodology used in the development of the Galician DBpedia and some of its applications for language processing in the fields of entity recognition and lexical extraction.
Keywords: DBpedia, Wikipedia, WordNet, linked open data, semantic web

1 Introducción

En este artículo[1] se describe la metodología seguida en la creación de la DBpedia del gallego y algunas de sus aplicaciones en el campo del procesamiento del lenguaje. La construcción de este recurso se realizó gracias a la financiación de la Red de Investigación *Tecnoloxías e análise dos datos lingüísticos*, orientada al desarrollo de recursos para el procesamiento lingüístico del gallego, siendo uno de sus objetivos principales la puesta en marcha de nuevas aplicaciones y herramientas con tecnologías de base semántica.

La DBpedia[2] (Lehmann et al., 2015) es un proyecto internacional para crear una versión estructurada de los contenidos de la Wikipedia[3] y publicarla libremente en Internet entrelazada con el conjunto de bases de conocimiento que constituyen la web semántica.

La DBpedia permite realizar consultas complejas a partir del conjunto de datos derivados de la Wikipedia y permite enlazar estos datos con otros conjuntos de datos que hay en la web, siguiendo las especificaciones para los datos enlazados abiertos (Linked Open Data)[4] establecidas por el W3C (World Wide Web Consortium) (Auer et al., 2007).

2 Recursos

La DBpedia del gallego, desarrollada y mantenida por el Grupo TALG (*Tecnoloxías e Aplicacións da Lingua Galega*) de la Universidade de Vigo, contiene 11 millones de tuplas semánticas extraídas a partir de toda la información contenida en la Galipedia[5] y está alojada en el subdominio oficial de dbpedia.org correspondiente a la lengua gallega[6].

La elaboración de la DBpedia del gallego supuso la adaptación de la aplicación de extracción de los datos procedentes de los ficheros *dump* de la Wikipedia, de Wikimedia Commons[7] y de Wikidata[8] para que funcionase satisfactoriamente con los datos procedentes de la Galipedia. Las modificaciones realizadas en el código de la aplicación se pueden consultar en Github[9] y han sido ya implementadas en la aplicación principal de extracción de la DBpedia[10].

[1] Esta investigación se realizó en el marco de la Red de Investigación *Tecnoloxías e análise dos datos lingüísticos* financiada por la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia, ref. CN 2014/007.

[2] `http://dbpedia.org`

[3] `http://wikipedia.org`

[4] `https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[5] `http://gl.wikipedia.org`

[6] `http://gl.dbpedia.org`

[7] `https://commons.wikimedia.org`

[8] `https://www.wikidata.org`

[9] `https://github.com/galician/extraction-framework/`

[10] `https://github.com/dbpedia/extraction-framework/`

Igualmente, con el mismo objetivo de creación del recurso, se elaboraron los ficheros de conversión (*mappings*) necesarios para obtener información estructurada a partir de las *infoboxes* y de las cajas de navegación de la Galipedia[11]. Aunque esta tarea se halla todavía en curso de finalización, la cobertura alcanzada con el trabajo ya realizado resulta bastante amplia, como se puede comprobar en las estadísticas disponibles de los *mappings* de la DBpedia[12]. El conjunto de datos se ha completado, además, con la extracción de los resúmenes de los artículos de la Galipedia ligados a cada recurso.

Los ficheros RDF de la DBpedia del gallego generados a partir de la Galipedia, pueden ser libremente descargados desde el sitio de la DBpedia[13], y sus contenidos pueden consultarse y visualizarse en la web del grupo mediante las aplicaciones Lodview[14] y LodLive[15] (ambas localizadas en gallego como parte del proyecto), utilizando la interfaz adaptada de la propia DBpedia[16] o a través del punto de acceso Virtuoso SPARQL a los datos estructurados[17].

La publicación del punto de acceso SPARQL propició también el modelado en formato de datos enlazados abiertos de Galnet[18] (Solla Portela y Gómez Guinovart, 2015), el WordNet 3.0 del gallego desarrollado por el Grupo TALG que forma parte de la distribución del Multilingual Central Repository (MCR) (González Agirre, Laparra, y Rigau, 2012). La consulta de la versión RDF de Galnet se encuentra disponible a través del servidor SPARQL de la DBpedia del gallego utilizando el grafo `http://sli.uvigo.gal/rdf_galnet`.

El diseño de la estructura de los datos RDF se basó en la versión 3.1 del WordNet de Princeton[19], siguiendo el modelo lemon[20], con ligeras modificaciones respecto al original para poder incorporar los enlaces con las clasificaciones semánticas y ontologías presentes en el MCR y Galnet[21] y mantener su naturaleza plurilingüe a través de un índice interlingüístico (ILI). Además, con el fin de ampliar su cobertura a consultas externas, se alineó cada synset con el correspondiente en la versión 3.1 de Princeton y con la versión 3.0 en formato lemonUby[22]. El resultado de este alineamiento conlleva la compatibilidad del índice interlingüístico de WordNet presente en el MCR con innumerables fuentes de datos enlazados que ya se encuentran disponibles en la web semántica.

3 Aplicaciones

3.1 DBpedia Spotlight

Una vez elaborados los recursos y habilitado el acceso abierto a los datos estructurados, se desarrolló una versión adaptada al gallego de la aplicación DBpedia Spotlight (Daiber et al., 2013) para poder ofrecer una primera herramienta de explotación inmediata de los datos de la DBpedia del gallego en el campo del procesamiento del lenguaje.

DBpedia Spotlight es una utilidad para la anotación de textos con referencias a los conceptos de la DBpedia. La identificación en contexto de las formas relativas a los conceptos se realiza mediante un sistema adaptable que localiza y desambigua de forma automática las menciones a recursos de la DBpedia presentes en el lenguaje natural. En este sentido, la identificación de entidades llevada a cabo por DBpedia Spotlight posee un alcance menos restringido que el reconocimiento de entidades nombradas, habitualmente limitado a ciertas categorías predefinidas como personas, organizaciones y lugares.

La adaptación al gallego de DBpedia Spotlight realizada en el marco de este proyecto identifica y anota en los textos las referencias a conceptos de la DBpedia del gallego, y puede utilizarse libremente desde su interfaz de usuario[23] o como servicio web[24].

[11] `http://mappings.dbpedia.org/index.php/Mapping_gl`

[12] `http://mappings.dbpedia.org/server/statistics/gl/`

[13] `http://downloads.dbpedia.org/2015-10/core-i18n/gl/`

[14] `http://sli.uvigo.gal/dbpedia/lodview/`

[15] `http://sli.uvigo.gal/dbpedia/lodlive/`

[16] `https://github.com/dbpedia/dbpedia-vad-i18n`

[17] `http://gl.dbpedia.org/sparql/`

[18] `http://sli.uvigo.gal/galnet/`

[19] `http://wordnet-rdf.princeton.edu`

[20] `http://lemon-model.net`

[21] Concretamente, los WordNet Domains (Bentivogli et al., 2004), la ontología Adimen-SUMO (Álvez, Lucio, y Rigau, 2012), la Top Ontology (Álvez et al., 2008), los Basic Level Concepts (Izquierdo, Suárez, y Rigau, 2007) y los epinónimos (Solla Portela y Gómez Guinovart, 2015)

[22] `http://lemon-model.net/lexica/uby/wn/`

[23] `http://sli.uvigo.gal/dbpedia/spotlight/`

[24] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service`

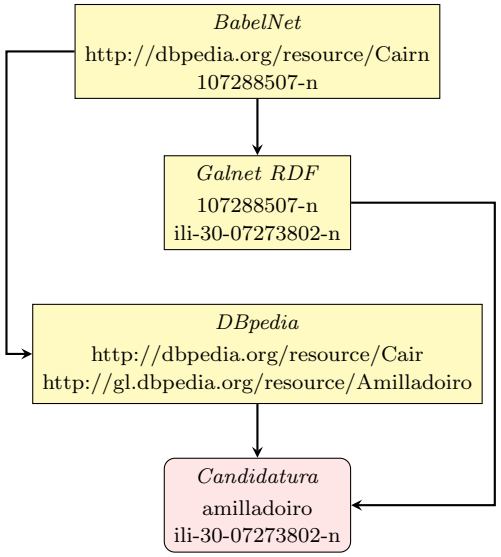


Figura 1: Extracción de variantes (1).

3.2 Extracción léxica

Para poder comprobar las posiblidades de explotación de estos recursos LOD en otras tareas de procesamiento del lenguaje, diseñamos dos experimentos de extracción léxica basados en la DBpedia dirigidos a la ampliación del WordNet del gallego. En el primer experimento de extracción, a parte de la DBpedia del gallego y de Galnet, usamos como fuente LOD remota la versión RDF de BabelNet[25]. El objetivo del experimento consiste en aumentar la cobertura de Galnet mediante variantes gallegas procedentes de la DBpedia limitándose a los synsets de Galnet que aún non tuvieran variantes gallegas.

En primer lugar, se obtuvieron de BabelNet los identificadores de sentido de WordNet 3.1 ligados a recursos de la DBpedia en inglés. El número de alineamientos identificador–recurso obtenidos mediante esta fuente ascendió a 7.796. Segundo, se obtuvieron de Galnet los ILIs de WordNet 3.0 correspondientes a los identificadores de sentido de WordNet 3.1 procedentes de BabelNet. Simultáneamente, se obtuvieron de la DBpedia del gallego los recursos gallegos correspondientes a los recursos de la DBpedia del inglés procedentes de BabelNet[26]. Por último, se identifican los synsets de Galnet correspondientes a los ILIs de WordNet 3.0 obtenidos y se proponen como candidatos a variante los recursos relacionados de la DBpedia del gallego.

Con esta estrategia se consiguieron 910 candidaturas con variantes nominales que apuntaban a synsets que todavía no tenían ninguna variante en gallego. El índice de precisión obtenido en el experimento de extracción, tras su revisión humana, alcanzó el 82,3 %, como se refleja en los resultados de la Tabla 1. Durante la revisión se observó además que, salvo en algunos casos aislados en los que la equivalencia entre idiomas en la DBpedia no es correcta, en la mayor parte de los casos en los que no se puede establecer la validez, el origen del error se encuentra en la inadecuación del alineamiento entre el recurso de la DBpedia y el identificador de WordNet 3.1 en BabelNet. La Figura 1 ilustra este proceso de extracción de variantes de Galnet a partir de los recursos LOD de la DBpedia, BabelNet y Galnet con un ejemplo de candidatura aceptada[27].

Variantes evaluadas	910	
Aceptadas	749	82,3 %
Rechazadas	161	17,7 %

Tabla 1: Evaluación de las candidaturas (1).

En un segundo experimento, exploramos la adquisición de variantes a partir de las equivalencias interlingüísticas de la DBpedia y de las variantes interlingüísticas presentes en los synsets del MCR. Partiendo de los synsets sin variante en gallego, se compararon las variantes existentes en catalán, euskera, portugués, español e inglés con los recursos de la DBpedia para cada una de estas lenguas, a fin de proponer candidaturas de nuevas variantes para el gallego (Figura 2). Con este método se generaron 2.194 candidaturas a partir de recursos con al menos una variante coincidente en alguna de las lenguas de los wordnets del MCR, con un índice de precisión tras la revisión humana del 88,3 % (Tabla 2).

Variantes evaluadas	2.194	
Aceptadas	1.937	88,3 %
Rechazadas	257	11,7 %

Tabla 2: Evaluación de las candidaturas (2).

[25] `http://babelnet.org/rdf/`

[26] Es preciso tener en cuenta que las tuplas de equivalencias interlingüísticas de la DBpedia se generan con el mismo código de extracción de información estructurada que se utiliza para la Wikipedia, pero se toman como fuente los datos procedentes de Wikidata.

[27] `http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-07273802-n`

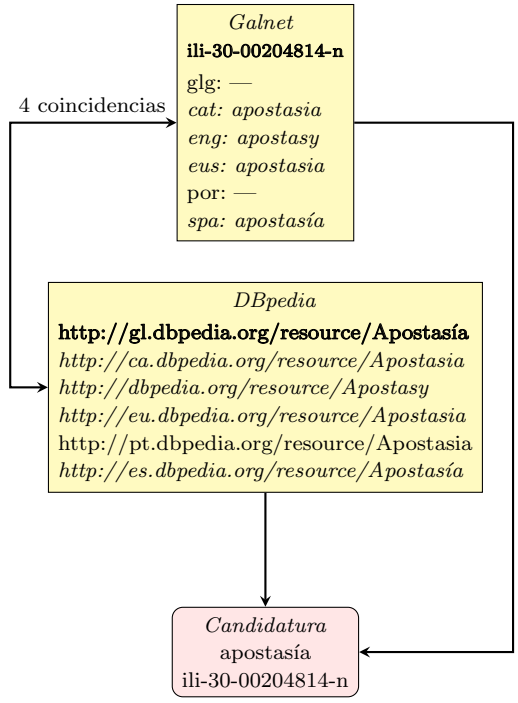


Figura 2: Extracción de variantes (2).

Las variantes aceptadas en estos dos experimentos fueron incorporadas al WordNet del gallego y pueden ser consultadas a través de su interfaz seleccionando como experimento *dbpedia*[28]. Ambas estrategias de extracción léxica pueden ser aplicadas, utilizando los mismos recursos, para sugerir candidaturas de variantes en cualquiera de las lenguas incluidas en los wordnets del MCR.

4 Conclusiones

La publicación de la DBpedia del gallego representa un avance importante para la presencia de información estructurada en lengua gallega en la web semántica. El punto de acceso SPARQL garantiza su aprovechamiento público en aplicaciones derivadas, además de permitir su interacción con los recursos disponibles en otros servidores con tecnologías semánticas. La explotación de la base de conocimientos de la DBpedia del gallego, en combinación con otros recursos en la web semántica, permitirá sin duda dinamizar proyectos, diseñar investigaciones y generar aplicaciones de gran interés en el ámbito del procesamiento del lenguaje.

[28] `http://sli.uvigo.gal/galnet_rev/galnet.php?version=dev&experiment=dbpedia`

Bibliografía

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver, y G. Rigau. 2008. Consistent annotation of EuroWordNet with the Top Concept Ontology. En *Proceedings of the 4th Global WordNet Conference*, Szeged. GWN.

Álvez, J., P. Lucio, y G. Rigau. 2012. Adimen-SUMO: Reengineering an Ontology for First-Order Reasoning. *International Journal on Semantic Web and Information Systems*, 8(4):80–116.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, y Z. Ives. 2007. Dbpedia: A nucleus for a web of open data. En *In 6th Int'l Semantic Web Conference, Busan, Korea*, págs. 11–15. Springer.

Bentivogli, L., P. Forner, B. Magnini, y E. Pianta. 2004. Revising WordNet domains hierarchy: Semantics, coverage, and balancing. En *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, págs. 101–108, Geneva. ACL.

Daiber, J., M. Jakob, C. Hokamp, y P. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. En *Proc. of the 9th International Conference on Semantic Systems*.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *6th Global WordNet Conference*.

Izquierdo, R., A. Suárez, y G. Rigau. 2007. Exploring the Automatic Selection of Basic Level Concepts. En *Proc. of the International Conference on Recent Advances on Natural Language Processing*, págs. 298–302, Shoumen.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195.

Solla Portela, M. A. y X. Gómez Guinovart. 2015. Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas. *Revista Galega de Filoloxía*, 16:169–201.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 143-146 recibido 15-3-2016 revisado 13-04-2016 aceptado 29-04-2016

SomEMBED: Comprensión del lenguaje en los medios de comunicación social-Representando contextos de forma continua

SomEMBED: Social Media language understanding- EMBEDing contexts

Paolo Rosso, Roberto Paredes
PRHLT, Universitat Politècnica de València
Camino de Vera s/n. 46022
Valencia, España
{prosso,rparedes}@prhlt.upv.es

Mariona Taulé, M. Antònia Martí
CLiC, Universitat de Barcelona
Gran Via 585, 08007
Barcelona, España
{mtaule,amarti}@ub.edu

Resumen: SomEMBED es un proyecto coordinado en el que participan el centro de investigación *Pattern Recognition and Human Language Technology* (PRHLT) de la Universitat Politècnica de València (UPV) y el grupo de investigación *Centre de Llenguatge i Computació* (CLiC) de la Universitat de Barcelona. Se trata de un proyecto del programa de I+D (TIN2015-71147) financiado por el Ministerio de Economía y Competitividad. Paolo Rosso coordina el proyecto SomEMBED y lidera el subproyecto SomEMBED-APP y Mariona Taulé lidera el subproyecto SomEMBED-SLang.
Palabras clave: *Embeddings*, representaciones distribuidas, medios de comunicación social, lenguaje no estándar, aplicaciones

Abstract: SomEMBED is a coordinated project involving the research center of *Pattern Recognition and Human Language Technology* (PRHLT) of the Universitat Politècnica de València and the research group of *Centre de Llenguatge i Computació* (CliC) of the Universitat de Barcelona. This is an R&D project (TIN2015-71147) funded by the Spanish Ministry of Economy and Competitiveness. Paolo Rosso coordinates the SomEMBED project and leads the subproject SomEMBED-APP and Mariona Taulé leads the SomEMBED-SLang subproject.
Keywords: Embeddings, distributed representations, social media, non-standard language, applications

1 Introducción

El proyecto SomEMBED tiene como objetivo general avanzar en el área de la Lingüística Computacional (LC) y el Procesamiento del Lenguaje Natural (PLN) con el fin de afrontar y dar solución a los retos que plantea el uso de la lengua en los medios de comunicación social en la web.

Desde la LC nuestro objetivo es el desarrollo de técnicas y métodos para la modelización de la lengua no estándar a partir de corpus representativos de los medios de comunicación social. Desde el PLN, nuestro objetivo es el desarrollo de nuevas técnicas y métodos a partir del estado actual de los conocimientos científico-técnicos para la resolución de tareas específicas en el marco de aplicaciones concretas.

En este proyecto convergen tres líneas de investigación estrechamente relacionadas: 1) la exploración y producción de diferentes metodologías para la extracción automática de patrones sintáctico-semánticos con el fin de representar semánticamente el contenido de los documentos, teniendo como eje central los métodos basados en representaciones continuas de texto (*embeddings*) que permiten modelar el contexto de un modo eficaz y eficiente; 2) el desarrollo de aplicaciones para la resolución de tareas concretas de PLN que permitan mejorar la comprensión automática del texto (p.e.: la detección del lenguaje figurado), e identificar aspectos claves del perfil de autores - edad, sexo, personalidad, lengua nativa, variedad lingüística (Franco-Salvador et al., 2015)- con especial interés en distinguir a los usuarios de los países de lengua hispana (España, México, Perú, etc.), aspectos que además permiten

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

utilizar su información en tareas como la minería de productos y servicios, en especial para la detección de opiniones falsas, y 3) la creación de recursos lingüísticos, especialmente corpus anotados, orientados al análisis de la lengua no estándar que servirán de base para la metodología de extracción de patrones y para las aplicaciones mencionadas que se van a desarrollar. Estas tres líneas de investigación se concretan en los objetivos que se detallan en el siguiente apartado.

2 Objetivos

1) La exploración y producción de nuevos métodos para la detección de patrones sintáctico-semánticos:

a. Experimentación con técnicas de generación de representaciones continuas de palabras para la obtención de *clusters* de palabras relacionadas[1].
b. Aplicación de la metodología DISCver (desarrollada en el proyecto anterior[2]) a una lengua distinta del español. En concreto, proponemos aplicarla al corpus del inglés *ukWaC* (Baroni et al., 2009) con el objetivo de ratificar la transportabilidad del método.
c. Experimentar con métodos alternativos para la obtención de patrones sintáctico-semánticos, basados en representaciones continuas de conjuntos de palabras (Le y Mikolov 2014).
d. Búsqueda de patrones que cumplan determinadas restricciones. En concreto, detección de patrones relacionados con los operadores de incertidumbre (*backward entailment operators*), fundamentales para la correcta interpretación de la polaridad (Danescu et al., 2009).

2. Creación de la infraestructura básica de recursos lingüísticos orientados al análisis de la lengua no estándar.[3]

a. Creación y anotación lingüística de corpus de lengua no estándar para diferentes variantes del español: *HispaSocialMedia* y *HispaLearners*. Por razones expositivas, denominaremos *HispaSocialMedia* al conjunto de corpus extraídos de distintos medios de comunicación social (microblogs, blogs, *reviews*, fórums) e *HispaLearners* para referirnos al conjunto de corpus formados por producciones orales (transcritas) o escritas producidas por aprendices de español como lengua extranjera.
e. Desarrollo de una base de conocimiento de patrones sintáctico-semánticos organizados de forma jerárquica según diferentes niveles de abstracción lingüística. Los patrones serán el resultado de la aplicación de los diferentes métodos descritos en el primer objetivo a los corpus *HispaSocialMedia* e *HispaLearners*.

3. Desarrollo de aplicaciones para los medios de comunicación social que cubran las tareas siguientes:

a. Identificación de lenguaje figurado (metáfora, analogía, humor, ironía y sarcasmo) frente al lenguaje literal (se utilizará el corpus *HispaSocialMedia*).
b. Detección de comunidades de usuarios en los medios de comunicación social con el objetivo de analizar las similitudes a nivel de patrón sintáctico-semántico entre los grupos de usuarios que utilizan y comparten diferentes tipos de lenguaje, por ejemplo el figurado.
c. Identificación de la variedad lingüística. En concreto, la detección del español peninsular frente al español de Latinoamérica: México, Argentina, etc. (se usará el corpus *HispaSocialMedia*).
d. Identificación de la lengua nativa de usuarios que escriben en español (se usará el corpus *HispaLearners*).
e. Identificación de los rasgos de los autores de textos en medios de comunicación social: sexo, edad, personalidad, tendencia política, entre otras (se usará el corpus corpus *HispaSocialMedia*).

3 Hipótesis

Existe la hipótesis de partida comúnmente aceptada de que la variante de lengua informal usada en los medios de comunicación social difiere sensiblemente de la variante estándar. La mayoría de herramientas de PLN disponibles actualmente están pensadas y adaptadas para la

[1] Véanse los *clusters* resultantes en: http://clic.ub.edu/corpus/es/Diana-Araknion-KB

[2] DIANA-Construcciones (TIN2012-38603).

[3] Se entiende por lengua no estándar todas aquellas variantes lingüísticas que se desvían de la forma oral o escrita estándar y que suelen transgredir la norma gramatical.

lengua estándar. Existen, a nuestro entender, dos maneras posibles de abordar el problema del tratamiento de la lengua no estándar: la primera consiste en la extensión o adaptación de las herramientas existentes, mientras que la segunda aborda el problema desde otra metodología radicalmente distinta. Esta metodología consiste en la extracción de patrones sintáctico-semánticos que modelizan el contenido de los corpus. La primera opción presenta un problema grave ya que la variación en los corpus de lengua informal es impredecible, de manera que hace inviable la adaptación de las herramientas existentes.

Nuestra hipótesis de trabajo es que: a) la modelización de los corpus en base a patrones sintáctico-semánticos permite una adecuada representación de su contenido; b) esta modelización abre la puerta al desarrollo de aplicaciones de PLN más eficientes en medios de comunicación social, y c) se obtiene una información lingüística relevante que sirve de base a estudios teóricos sobre el lenguaje y apunta a un modelo de gramática de la actuación, es decir, una gramática de la lengua en uso. La ventaja es que la mayoría de métodos que se desarrollan en el marco de esta aproximación son independientes de la lengua y de la variante en que se utiliza. Esto es debido a que para cada colección de corpus, nuestra aproximación obtiene una representación en términos de patrones y son estos patrones los que proporcionan las características lingüísticas de la misma. Esta representación constituye el input para el desarrollo de aplicaciones y tareas de PLN.

4 Metodología

La extracción automática de los patrones se realizará a partir del tratamiento masivo de corpus, aplicando principalmente técnicas de *deep learning*. Siguiendo la línea de investigación abierta por Mikolov et al., (2013b), se procederá a generar nuevos patrones, utilizando los conjuntos de herramientas de word2vec[4] y Glove[5]. Se trata de un método basado en las representaciones continuas de palabras (*embeddings*). Estas representaciones consisten en vectores n-dimensionales que han sido generados mediante algoritmos log-lineales de modo que vectores de palabras similares en contexto guardan una similitud próxima. A los modelos originales de generación de vectores continuos de palabras, compilados en el conjunto de herramientas de word2vec, le han seguido otras alternativas igualmente eficientes (Pennington et al., 2014), además de otros modelos que permiten generar representaciones continuas de conjuntos de palabras para su aplicación en frases y documentos (Le y Mikolov, 2014).

Otra línea de investigación se centrará en la obtención automática de operadores de incertidumbre para el español basándonos en la propuesta de Danescu et al., (2010) que, a partir de operadores comúnmente aceptados (por ejemplo: no, jamás, dudar, imaginar, etc.) aplica algoritmos recursivos para extraer nuevos operadores mediante la identificación de contextos prototípicos que los suelen acompañar.

Los diferentes métodos y técnicas que se lleven a cabo se aplicarán para la resolución de tareas específicas en el marco de aplicaciones concretas (véase apartado 3 de la sección 2).

En cuanto a los corpus, se recopilarán mediante el uso de algoritmos de web *crawling* para la compilación del corpus *HispaSocialMedia* y el diseño de tareas (p.e. redacción de textos) para la creación del corpus *HispaLearners*. Una vez obtenidos los corpus se procederá a su procesamiento morfosintáctico automático. Se anotarán manualmente aspectos específicos (p.e.: anotación de los operadores de incertidumbre, especialmente la negación) en una selección de los corpus compilados.

5 Resultados esperables

Destacamos el desarrollo de diferentes técnicas y métodos susceptibles de ser incorporados en aplicaciones de entornos de medios de comunicación social con diferentes finalidades:

- La identificación de la variedad lingüística y la lengua nativa del autor, que permitirá geolocalizar y discriminar la relevancia de noticias y eventos en medios de comunicación social para su posterior análisis y explotación. Dada la activa participación de la población en las redes sociales y la complejidad de distinguir la procedencia geográfica del autor (España, México, Perú, Chile, Argentina, etc.), dicha aplicación tiene mucha importancia en los medios de comunicación (análisis de opinión, difusión e impacto) y la industria (análisis de

[4] https://code.google.com/p/word2vec/

[5] http://nlp.stanford.edu/projects/glove/

opinión, estudio de mercado, explotación de productos, etc.).

- La identificación de rasgos del autor (sexo, edad, personalidad, tendencia política, entre otros) que proporcionará información detallada sobre el perfil del autor del texto ante el que nos encontramos y potencialmente contribuirá notablemente en la lingüística forense de cara a la identificación y estudio de infractores, delincuentes y criminales. Puede ser también de utilidad para los intereses de los medios de comunicación social y la industria mencionados anteriormente.
- La detección de opiniones fraudulentas y otros casos de engaño creados por parte de personas específicamente contratadas para este fin. Nuestra contribución consistirá en el desarrollo de técnicas para la extracción de las construcciones o patrones lingüísticos más recurrentes en la expresión de opiniones fraudulentas y otros casos de engaño sobre personas, organizaciones, productos y servicios, lo cual tendrá un potencial uso para la industria.
- La detección de mecanismos de lenguaje figurado en textos subjetivos (ironía, metáfora, humor, etc.) de los medios de comunicación social y en comunidades de usuarios. Las herramientas que se desarrollarán permitirán superar la barrera que supone el uso de lenguaje figurado, que altera radicalmente el significado del mensaje, y contribuirá a la mejora de tareas mencionadas como el análisis de opinión, además de proporcionar información útil de cómo la información y el lenguaje se comparte dentro de comunidades concretas de usuarios.
- Los nuevos recursos y herramientas que se prevé desarrollar han de incidir en la mejora de las técnicas y métodos de PLN que se usarán en aplicaciones sobre los medios de comunicación social. En concreto, los patrones sintáctico-semánticos generados mediante representaciones continuas servirán para modelizar el contenido de corpus mejorando los resultados obtenidos hasta el momento.
- Por otro lado, los corpus que se prevé construir -*HispaSocialMedia* y *HispaLearners*- sobre la lengua no estándar para las diferentes tareas sobre las que se trabajará en este proyecto, fomentarán y contribuirán a la investigación y avance por parte de otras entidades.

El uso de la lengua no estándar dificulta el uso de las herramientas tradicionales de análisis de la lengua y requiere del estudio detallado de este tipo de variante lingüística en estos medios, y de la investigación y desarrollo de nuevos algoritmos y metodologías para el procesamiento del lenguaje que permitan poder compilar, comprender y explotar el conocimiento que se encuentra dentro de los medios de comunicación social. Los resultados del proyecto coordinado *SomEMBED* podrán incidir favorablemente en el desarrollo de aplicaciones en dicho ámbito (redes sociales, blogs, foros, canales de opinión) y en nuestro conocimiento sobre la lengua no estándar que allí se utiliza.

Bibliografía

Baroni M., S. Bernardini, A. Ferraresi y E. Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43: 209-226, Spinger.

Danescu-Niculescu-Mizil, C., L. Lee, y R. Ducott. 2009. Without a'doubt'?: unsupervised discovery of downward-entailing operators. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, páginas 137-145. Association for Computational Linguistics.

Franco-Salvador, M., F. Rangel, P. Rosso, M. Taulé y A. Martí. 2015. Language Variety Identification using Distributed Representations of Words and Documents. *Proceedings of the 6th International Conerence. of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction* (CLEF 2015), Springer-Verlag, LNCS (9283): 28-40.

Le, Q. V., y T. Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv*:1405.4053.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Pennington, J., R. Socher y C.D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing* (EMNLP 2014), 12: 1532-1543.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 147-150 recibido 15-03-2016 revisado 13-04-2016 aceptado 3-05-2016

ASLP-MULAN: Audio speech and language processing for multimedia analytics

ASLP-MULAN: Procesado de audio, habla y lenguaje para análisis de información multimedia

Javier Ferreiros, José Manuel Pardo
GTH-Universidad Politécnica Madrid
Ciudad Universitaria s/n. Madrid
jfl@die.upm.es

Lluís-F Hurtado, Encarna Segarra
ELiRF-Universitat Politècnica València
Camino de Vera s/n 46022 Valencia
lhurtado@dsic.upv.es

Alfonso Ortega, Eduardo Lleida
Universidad de Zaragoza
C/ María de Luna 1. 50018 Zaragoza
ortega@unizar.es

María Inés Torres, Raquel Justo
Universidad del País Vasco UPV/EHU
Campus de Leioa. 48940 Leioa. Vizcaya
manes@we.lc.ehu.es

Abstract: Our intention is generating the right mixture of audio, speech and language technologies with *big data* ones. Some audio, speech and language automatic technologies are available or gaining enough degree of maturity as to be able to help to this objective: automatic speech transcription, query by spoken example, spoken information retrieval, natural language processing, unstructured multimedia contents transcription and description, multimedia files summarization, spoken emotion detection and sentiment analysis, speech and text understanding, etc. They seem to be worthwhile to be joined and put at work on automatically captured data streams coming from several sources of information like YouTube, Facebook, Twitter, online newspapers, web search engines, etc. to automatically generate reports that include both scientific based scores and subjective but relevant summarized statements on the tendency analysis and the perceived satisfaction of a product, a company or another entity by the general population.
Keywords: Audio processing, speech recognition, language processing, sentiment analysis, emotional speech synthesis, multimedia social Web, information retrieval.

Resumen: Nuestra intención es generar la mezcla ideal de tecnologías del audio, el habla y el lenguaje con las de *big data*. Algunas tecnologías automáticas del procesado de audio, habla y lenguaje están adquiriendo suficiente grado de madurez para ser capaces de ayudar a este objetivo: transcripción automática del habla, métodos de búsqueda por habla, recuperación de documentos hablados, procesado del lenguaje natural, transcripción y descripción de contenidos multimedia no estructurados, resumen de ficheros multimedia, detección de emoción en el habla y análisis de sentimientos, comprensión de texto y habla, etc. Parece que merece la pena unirlos y ponerlos a trabajar sobre secuencias de datos obtenidos automáticamente procedentes de diversas fuentes de información como YouTube, Facebook, Twitter, periódicos digitales, buscadores de internet, etc. para generar informes que incluyan tanto puntuaciones basadas en análisis cuantitativo como expresiones resumidas subjetivas pero significativas sobre el análisis de tendencias y la satisfacción percibida sobre un producto, una empresa u otra entidad.
Palabras clave: Procesamiento de audio, reconocimiento del habla, procesamiento del lenguaje, análisis de sentimientos, síntesis de emociones, Web social multimedia, recuperación de información.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

1 About the project

This Project is founded by the "Ministerio de Economía y Competitividad" TIN2014-54288-C4 and there are four reseach groups involved: ELiRF (Universitat Politècnica de València), ViVoLab (Universidad de Zaragoza), SPIN (Universidad del Pais Vasco), GTH (Universidad Politécnica de Madrid).

2 Introduction

Society moves motivated by a lot of influences from fashions to established tendencies. Moreover, nowadays this movement is also highly modulated by the instant exchange of information fostered by social media far beyond TVs and radios. Internet social media sharing opportunities have reached a high percentage of the population and it is crucial, not only for the companies but for all economic and administration drivers in general, to know about the opinions, reputation feeling, political polarities and tendencies auto induced inside the social media. Having this information is relevant to drive new marketing policies and also have high relevance for security and defense in other contexts.

This relevance has been already detected by some companies that offer market surveys and reputation studies acquired in the social media by products, companies and other entities as political parties and administrations. The study is mostly based on costly polls and superficial hand analysis (as opposed to an automatic one) on a sampling on some limited sources using simple criteria in the analysis and we believe that they need a technological impulse to improve their capacities.

Big data science community has begun to apply their specific abilities to these data content analysis. In parallel, some audio, speech and language automatic technologies are now available or gaining enough degree of maturity as to be able to help to this objective. Some of these technologies are: automatic speech transcription, query by spoken example, spoken information retrieval, natural language processing, unstructured multimedia contents transcription and description, multimedia files summarization, spoken emotion detection and sentiment analysis, speech and text understanding, etc. They seem to be worthwhile to be joined and put at work on automatically captured data streams coming from several sources of information like YouTube, Facebook, Twitter, online newspapers, web search engines, etc. Out of this analysis, we will automatically generate reports that include both scientific based scores and subjective but relevant summarized statements on the tendency analysis and the perceived satisfaction of a product, a company or another entity by the general population.

Our intention is working in this direction and generating the right mixture of audio, speech and language technologies with big data ones as to be able to offer it to both the analytics companies interested in this information, improving their capacity to offer their services with increased quality, accuracy and usability of their reports. Also directly to companies or administrations willing to gain this information on their own via deploying our new solutions in their marketing or intelligence departments.

Relevant information about the feelings of the general public about products, companies and institutions can be distilled from multimedia information which is spread on the Internet on several content sharing applications like YouTube, Facebook, twitter, etc. Additionally, a lot of people is interested in finding the most appropriate documents related to a personal need of informative data connected with their interests or on products to be able to compare them to choose the best one under their own criteria.

From several potential ways to make use of social digital multimedia data we choose a couple of applications for the motivation and demonstration of this proposal: the retrieval of the most appropriate multimedia documents under a certain interest and the analysis of the opinion on a brand, person or event making use of multimedia files available in social digital media.

The problem is that most of this information is presented in an unstructured format: it may be directly in text or interleaved in the audio of a video or in the image of the video itself.

A lot of analytics effort is based on text or video, but audio, speech and language is not fully considered yet. We will use several audio, speech and language technologies to extract as much information as possible from these sources. First of all we face a problem of selecting the messages applicable to the specific search we are interested in. Information retrieval techniques including query by example will help us score and select the most appropriate. Then, we have to use techniques able to extract the

messages from the unstructured sources. After extracting the message, natural language processing has to be employed to obtain the semantics behind. Then, other kind of analysis have to be applied to analyze the emotions and polarities represented by these messages.

Afterwards, statistical analysis must be made in order to know the relevance of each analyzed tendency, polarity or sentiment. Finally, summarization techniques help us compose the report in the most usable format.

3 Project objectives

3.1 Strategic objectives

The main goal of this proposal is to explore the maturity of diverse technologies, progress or develop them when necessary as well as use them to deal with multimedia document retrieval and analysis, focusing on the information provided by audio and speech. We thus contribute to multimedia information retrieval and other use cases, for example, the building of automatic market and reputation analysis from social media sources. In this context our two main strategic objectives are:

- **Developing audio, speech and language technologies devoted to**
    1. Speech and audio information retrieval.
    2. Multimedia information analytics.
    3. Automatic output generation.
- **Transferring the acquired knowledge to the society trough dissemination and tec**hnology transfer actions.

3.2 Scientific-Technological objectives

Following the project structure, our scientific-technological goals are:

- **To develop technologies for audio and speech processing intended to**
    1. Transcribe audio files, sometimes coming from videos, into text.
    2. Detect and Classify multimedia events from the acoustics
    3. Language and speaker identification and diarization
    4. Label emotions directly from the acoustics.
- **To develop technologies for audio and speech processing aimed to**
    1. Key term detection and concept discovery.
    2. Multilingual language understanding.
- **To develop technologies for multimedia analytics devoted to**
    1. Integrate developed audio, speech and language processing techniques for Multimedia information retrieval purposes
    2. detect, analyze and classify emotional states and language
    3. Analyze reputation, polarity and tendencies from multimedia social web.
- **To develop technologies for output generation and results presentation dealing with**
    1. Automatic report and summary generation
    2. Natural language generation
    3. Emotional speech generation

3.3 Transferring knowledge objectives

To deal with the second strategic objective we propose three main objectives related to the knowledge transfer to the society:

- **To develop and evaluate application demonstrators related with two use cases:**
    1. Multimedia information retreival
    2. Polarity and tendencies report
- **To develop multimedia annotated resources and software tools** freely available.
- **To train experts in the developed technologies** that may be employed by companies interested in our results.

4 Acknowledgments

This work is founded by the “Ministerio de Economía y Competitividad” TIN2014-54288-C4.

References

García F., L. Hurtado, E. Segarra, E. Sanchis, and G. Riccardi, “Combining multiple translation systems for Spoken Language Understanding portability,” in Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012), 2012, pp. 282–289.

Hurtado L.F., J. Planells, E. Segarra, E. Sanchis, D. Griol (2010): “A stochastic finite-state

transducer approach to spoken dialog management". Proc. of Interspeech, pp. 3002-3005

Justo R., T. Corcoran, S. M. Lukin, M. Walker: "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web". Knowledege-Based Systems. 2014

Justo R., M. I. Torres "Integration of complex language models in ASR and LU systems". Pattern Anal. Appl. 18(3): 493-505, 2015

Martinez F.F., J. Ferreiros, R. Cordoba, J.M. Montero, R. San-Segundo and J.M. Pardo " A bayesian networks approach for dialog modeling: The fusion bn". Proceedings of ICASSP 2009, pp. 4789-4792

Miguel A., J. Villalba, A. Ortega, E. Lleida, C. Vaquero "Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition". INTERSPEECH 2014

Pla, F., L.F. Hurtado. "Political tendency identifcation in twitter using sentiment analysis techniques". Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics

Planells J., L.F. Hurtado, E. Sanchis and E. Segarra (2012): "An online generated transducer to increase dialog manager coverage". Proc. of Interspeech, pp. 1-4

Vaquero C., A. Ortega, A. Miguel, J. Villalba, E. Lleida "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification". Interspeech, Makuhari, Japan. 2010

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 151-154 recibido 24-03-2016 revisado 06-04-2016 aceptado 25-04-2016

CLARIN Centro-K-español

Spanish CLARIN K-Centre

Núria Bel
Universidad Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
nuria.bel@upf.edu

Elena González-Blanco García
Universidad Nacional de
Educación a Distancia
Paseo Senda del Rey 7
28040 Madrid
egonzalezblanco@flog.uned.es

Mikel Iruskieta
Universidad del País Vasco
(UPV/EHU)
Sarriena auzoa z/g
48940 Leioa
mikel.iruskieta@ehu.eus

Resumen: Presentamos CLARIN Centro-K-español que forma parte de la infraestructura europea CLARIN, *Common Language Resources and Technology Infrastructure*, y cuyo objetivo es ofrecer los conocimientos y experiencia de los tres grupos que inicialmente lo componen en la utilización de tecnología para la investigación en humanidades y ciencias sociales.
Palabras clave: infraestructura lingüística, análisis de textos, humanidades, ciencias sociales.

Abstract: We introduce Spanish CLARIN Centre-K, a node of the European infrastructure CLARIN, Common Language Resources and Technology, whose objective is to share knowledge and experience of the three funding constituent groups for research in humanities and social sciences.
Keywords: Language infrastructure, text analytics, humanities, social sciences.

1 Introducción

CLARIN[1], la e-infraestructura europea de investigación para humanidades y ciencias sociales se ha ido desplegando desde hace 10 años como una red de centros que comparten misión, tecnología y recursos para ponerse a disposición de los investigadores que trabajan en el procesamiento y explotación de textos (escritos u orales) en áreas de humanidades y ciencias sociales. El objetivo es garantizar el acceso, integración y explotación de la gran cantidad de datos lingüísticos (o relacionados con las lenguas) y la tecnología relacionada. Actualmente CLARIN da servicios a investigadores en psicología, lingüística, filología, ciencias políticas, o sociología entre otros.

CLARIN fue uno de los proyectos seleccionados por el Comité ESFRI (*European Strategy Forum on Research Infraestructures*) y que figuran en la primera "Hoja de ruta"[2] de las infraestructuras que habían de ser construidas, por su importancia para la investigación, a diez años vista. Ya se han cumplido los diez años y su despliegue como infraestructura europea demuestra lo acertado de aquella decisión.

La Comisión de la Unión Europea dispuso la cofinanciación de una fase preparatoria para estos proyectos dentro del área Infraestructuras del VII Programa Marco. En España, el Ministerio de Educación y Ciencia, Dirección General de Política Tecnológica, Subdirección General de Promoción e Infraestructuras Tecnológicas y Grandes Instalaciones (CAC-2007-23) primero, y después el Ministerio de Ciencia e Innovación, Dirección General de Planificación y Coordinación (ICTS-2008-11) asumieron la cofinanciación de dicha fase preparatoria para la participación española. Por otra parte, el Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya ha participado en la financiación del proyecto CLARIN-CAT, apoyando la presencia de datos y herramientas específicamente en y para la lengua catalana. El Centro de Competencias CLARIN-IULA-UPF se cofinanció mediante el programa FEDER Catalunya 2007-2013.

[1] www.clarin.eu
[2] http://cordis.europa.eu/esfri/roadmap.htm

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

CLARIN está constituido formalmente como un consorcio europeo[3] de países y cuenta en la actualidad con diecisiete asociados y un país observador, Los asociados pueden incorporar centros de investigación específicos como miembros de la red. La red de centros CLARIN es actualmente de 137, entre los que se incluyen instituciones como universidades, centros de investigación, academias nacionales de ciencia y academias nacionales de la lengua.

Además, el consorcio CLARIN invita a incorporarse, como centro de conocimiento o *Centre-K*, a instituciones, de estados asociados o no –que es el caso de España–, que dispongan de servicios estables y que quieran ofrecer sus conocimientos y experiencia en el uso de infraestructuras lingüísticas en la investigación en una lengua o tema particular.

En 2015, CLARIN concedió el reconocimiento de Centre-K al centro distribuido especializado para España y tres de sus lenguas oficiales (castellano, catalán y euskera), el CLARIN Centro-K-español. Constituido como una asociación de centros ya existentes, en el CLARIN Centro-K-español se reúnen las especialidades del Centro de Competencias CLARIN IULA de la Universidad Pompeu Fabra (UPF), el Laboratorio de Innovación en Humanidades Digitales de la Universidad Nacional de Educación a Distancia (UNED) y del Grupo IXA de la Universidad del País Vasco/Euskal Herriko Unibertsitea (UPV/EHU).

2 *CLARIN, e-Infraestructura de Servicios*

CLARIN está concebida como una e-Infraestructura de tecnología lingüística que ofrece, entre otras cosas, un observatorio virtual de tecnología lingüística común para los países participantes, varios repositorios que almacenan, recogen y preservan proyectos y datos, clasificados con metadatos que conllevan índices de calidad (Thompson), y sistemas de identificación permanente, además de *login* unificado para las diferentes instituciones que participan. Todo esto, acompañado de un potente grupo de investigadores con una intensa actividad académica manifestada a través de congresos, seminarios, talleres y proyectos específicos en los diferentes grupos de interés, que han generado a su vez la creación e inclusión en otros proyectos europeos, como Europeana-DSI[4], LT-Observatory[5], EUDAT2020[6] o Parthenos[7].

3 *Descripción de los servicios del CLARIN Centro-K-español*

Los participantes en el Centro-K-español[8] tienen en común el objetivo de fomentar y asistir en el uso de la tecnología en la investigación en Humanidades y Ciencias Sociales. La experiencia y competencias de los tres grupos son complementarios, abarcando así buena parte del espectro de conocimientos necesarios para asesorar a los investigadores de estas áreas. IXA y IULA-UPF-CCC están especializados en *Text Analytics* y Tecnologías y Recursos del Lenguaje. LINHD está especializado en los beneficios del enriquecimiento de textos, Web Semántica, bases de datos y tecnologías de visualización para las Humanidades Digitales. Los tres ofrecen servicios a investigadores que trabajan con textos en español e inglés y, además, IXA aporta competencias en el manejo de textos en euskera y IULA-UPF-CCC de textos en catalán: disponen de recursos y herramientas para el análisis y la anotación automática de los textos en estas lenguas.

Así, el CLARIN Centro-K-español consta de los servicios que aportan sus grupos participantes y que en el Centro-K se asocian para constituirse en un punto de acceso único y actuar así de agente de distribución de solicitudes de servicios. Como Centro-K, suman las fortalezas de los tres centros participantes con el fin de ofrecer servicios a la comunidad de forma unificada y organizada.

Los servicios que el CLARIN Centro-K-español ofrece son:

– Consultoría virtual para asesorar y responder dudas sobre cuestiones prácticas relacionadas con estándares, uso de herramientas de procesamiento y acceso a recursos lingüísticos. El centro ofrece contacto por correo electrónico con garantía de respuesta en 24 horas. Se trata de un servicio de orientación personalizada y que también pone al servicio de los investigadores

[3] *European Research Infrastructure Consortium*.

[4] http://pro.europeana.eu/

[5] http://www.lt-observatory.eu/

[6] http://eudat.eu/

[7] http://www.parthenos-project.eu/

[8] http://www.clarin-es.org

documentación y casos de uso conocidos en los que encontrar buenas prácticas.
- Soporte para auto-aprendizaje con recursos especializados: catálogos especializados donde encontrar información sobre prácticas actuales y acceso directo a herramientas, en aplicaciones web, que facilitan el uso de las tecnologías existentes, videotutoriales, MOOCs, etc.
- Organización de programas de enseñanza y formación para investigadores, estudiantes, proyectos o grupos de interés.
- Servicios tecnológicos y de gestión y planificación de proyectos personalizados en función de las necesidades de los solicitantes.

4 Descripción de los grupos del Centro-K-español

El Centro de Competencias CLARIN IULA-UPF[9] está especializado en Análisis y Minería de Textos y en Tecnologías y Recursos del Lenguaje. Dirigido por el grupo Tecnología de los Recursos Lingüísticos, ha liderado la participación de España en CLARIN como proveedor de servicios web de procesamiento del lenguaje natural. Ha desarrollado aplicaciones web como ContaWords[10] con la que el usuario puede obtener información cuantitativa de textos (aportados por el usuario o de la web): frecuencia de palabras, por categorías morfosintácticas, reconocimiento y clasificación de entidades nombradas. Los resultados se obtienen en un formato fácilmente manipulable por el usuario. También ofrece acceso web a analizadores de dependencias sintácticas para el español[11], y a otras herramientas básicas de análisis populares como FreeLing (Padró y Stanilovsky, 2012) y MaltParser (Nivre et al., 2004). El centro de competencias, también cuenta con la experiencia del HDLab@UPF, un nuevo entorno de investigación y aprendizaje desarrollado por el Departamento de Humanidades de la UPF.

El Laboratorio de Innovación en Humanidades Digitales[12] (LINHD) fue fundado en 2014 gracias a la financiación inicial de la UNED. Se trata de un centro interdisciplinar de investigación en el que participan todas las facultades de humanidades y ciencias sociales, además de la ETSI de Ingeniería Informática de la UNED, la sección digital de la Biblioteca, el CEMAV e Intecca (medios audiovisuales de la UNED). El centro agrupa hoy día más de una veintena de proyectos, es un centro de referencia en formación en humanidades digitales en español, pues ofrece tres títulos propios, una escuela de verano con más de un centenar de alumnos y diferentes actividades colaborativas con otros centros e instituciones a nivel nacional e internacional (entre los que destaca el Secrit-Conicet, en Argentina).

El LINDH está especializado en el enriquecimiento de textos, la web semántica, las bases de datos y la edición digital. Se ocupa de velar por los estándares propios de las humanidades digitales, como el TEI-XML para el marcado de textos, además de ofrecer un acercamiento al mundo hispanohablante de los proyectos, recursos y actividades de otros países –principalmente anglófonos–. En este sentido, destacan las actividades colaborativas y de traducción del entorno virtual de edición Textgrid de Dariah-DE[13], del vocabulario semántico de TADIRAH[14], del índice de herramientas Dirt Directory[15] y de la gestión de los dos últimos eventos de blogging DayofDH 2015 y 2016. En investigación es, además un centro puntero, pues ha recibido recientemente un proyecto ERC y está trabajando en la creación de un entorno virtual de investigación para humanistas que se pondrá en breve a disposición del centro-K.

El Grupo IXA es un grupo de investigación multidisciplinario de la UPV/EHU que incluye miembros de cinco departamentos: Lenguajes y Sistemas Informáticos, Arquitectura y Tecnología de Computadores, Ciencia de la Computación e Inteligencia Artificial, Lengua Vasca y Comunicación (Filología Vasca), y Didáctica de la Lengua y la Literatura. El grupo IXA trabaja con el procesamiento del lenguaje natural desde el año 1988. Está especializado en el Análisis y Minería de Textos y en las Tecnologías y Recursos del Lenguaje. La lengua de estudio principal es el euskera, aunque también se han desarrollado productos para otras lenguas, como el inglés y el

[9] http://clarin-es-lab.org
[10] http://contawords.iula.upf.edu
[11] http://lod.iula.upf.edu/resources/278
[12] http://linhd.uned.es
[13] https://de.dariah.eu/
[14] http://tadirah.dariah.eu/vocab/index.php
[15] http://dirtdirectory.org/tadirah

castellano. Sus productos más reconocidos son el corrector ortográfico Xuxen, el traductor automático Matxin integrado en la plataforma Opentrad, Basque WordNet, el corpus de Ciencia y Tecnología (ZT), y el Corpus de Referencia para el Procesamiento del Euskera (EPEC). Asimismo, ha desarrollado más de 20 herramientas[16], algunas de estas herramientas han sido reutilizadas y rediseñadas para el Centro-K. Estas nuevas herramientas son útiles para un uso masivo e intuitivo, como por ejemplo, el programa ANALITZAK[17] basado en los IXA-PIPES (Agerri et al., 2014), con el que se obtienen frecuencias de morfemas, palabras por categorías o entidades de textos y webs tanto en euskera, como en inglés y en castellano.

Para la consulta de información lingüística de diferentes niveles se ofrece, además, acceso web a bases de datos y herramientas de uso sencillo: i) EDBL[18], base de datos lexical de euskera, en la que se detalla la información morfosintáctica de palabras y morfemas, utilizada en el corrector Xuxen. ii) Konbitzul[19], para consultar combinaciones (y su información morfosintáctica) de nombres y verbos útiles para la traducción de euskera a castellano o viceversa, útil para que el traductor automático Matxin traduzca adecuadamente dichas unidades fraseológicas. iii) e-ROLda[20], donde se puede realizar consultas para obtener información sintáctico-semántico de verbos (e incluso de nombres) en corpus anotado. iv) EusEduSeg[21], el segmentador discursivo automático para el euskera y diferentes bases de datos[22] con información discursiva basados en el trabajo desarrollado por Iruskieta et al. (2013), donde se pueden realizar consultas sobre relaciones de coherencia en lenguas tan dispares como el euskera, inglés, castellano y portugués (y muy pronto también el chino), útil para tareas de análisis de sentimiento, traducción y análisis de textos científicos, políticos y de crítica literaria.

[16] https://ixa.si.ehu.es/Ixa/Produktuak
[17] http://ixa2.si.ehu.es/clarink
[18] http://ixa2.si.ehu.es/edbl/
[19] http://ixa2.si.ehu.es/konbitzul/
[20] http://ixa2.si.ehu.es/e-rolda/en/
[21] http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl
[22] http://ixa2.si.ehu.es/rst/

5 Conclusión y trabajo futuro

Las herramientas brevemente descritas en este trabajo han sido desarrolladas para que la investigación en humanidades y ciencias sociales pueda tener una base lingüística fiable y de fácil uso, como puede ser el ejemplo de Villegas et al. (2012). Evidentemente las herramientas aquí descritas no satisfacen las necesidades existentes hoy en día, pero el proyecto CLARIN Centro-K-español ofrece la posibilidad de explorar la utilización de herramientas de interés general contando con servicios de consultoría o soporte para el aprendizaje, así como para diseñar y desarrollar mejores herramientas en colaboración con los grupos de investigación, empresas y agentes educativos que estén interesados en este proyecto y que pueden unirse al centro.

Referencias

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Proceedings of the 9th Language Resources and Evaluation Conference, LREC2014.

Iruskieta, M., M.J. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, O. Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In the 4th Workshop RST and Discourse Studies, pp. 40-49, October, Fortaleza, Brasil.

Nivre, J., J. Hall, and J. Nilsson. 2004. Memory-Based Dependency Parsing. In Ng, H. T. and Riloff, E. (eds.) *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, pp. 49-56

Padró, L., and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the 8th Language Resources and Evaluation Conference, LREC2012.

Villegas M., N. Bel, C. Gonzalo, A. Moreno, and N. Simelio. 2012. Using Language Resources in Humanities research. In Proceedings of the 8th Language Resources and Evaluation Conference, LREC2012.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 155-158 recibido 23-03-2016 revisado 06-04-2016 aceptado 27-04-2016

Deteami research-transference project: natural language processing technologies to the aid of pharmacy and pharmacosurveillance

Proyecto de transferencia tecnológica Deteami: tecnologías de procesamiento del lenguaje natural para la ayuda en farmacia y en farmacovigilancia

A. Casillas[(1)]∗, **A. Díaz de Ilarraza**[(1)], **K. Gojenola**[(1)],
L. Mendarte[(2)], **M. Oronoz**[(1)], **J. Peral**[(3)], **A. Pérez**[(1)]
[(1)]IXA research group (UPV-EHU);
∗arantza.casillas@ehu.eus
[(2)]Basurto University Hospital;
[(3)]Galdakao-Usansolo Hospital

Abstract: The goal of the Deteami project is to develop tools that make clinicians aware of adverse drug reactions stated in electronic health records of the clinical digital history.The records produced in hospitals are a valuable though nearly unexplored source of information among others due to the fact that are tough to get due to privacy and confidentiality restrictions. To leverage the clinicians work of reading and analyzing the health records looking for information about the health of the patients, in this project we explore the records automatically, identify among others disorder and drug entities, and infer medical information, in this case, adverse drug reactions. In this project a research-framework was settled with the Galdakao-Usansolo and Basurto Hospitals from Osakidetza (the Basque Health System). Osakidetza provided both the texts and the final user feedback, as well as, specialists that annotate the corpora, an in this way, we obtained a gold-standard.
Keywords: Technological transference, clinical text mining, entity recognition

Resumen: El objetivo del proyecto Deteami es el desarrollo de herramientas para ayudar al personal clínico a identificar reacciones adversas a medicamentos en informes médicos electrónicos de la historia clínica digital. Los informes que se generan en los hospitales son una valiosa fuente de información aún no debidamente explotada debido principalmente a restricciones de privacidad y confidencialidad. Con el objetivo de aliviar el trabajo del personal clínico que se dedica a leer y analizar los informes médicos buscando información sobre la salud de los pacientes, en este proyecto analizamos automáticamente los informes, identificamos entre otras entidades que describen enfermedades y medicamentos, y finalmente, inferimos información médica; en este caso, reacciones adversas a medicamentos. En este proyecto hemos establecido un marco de colaboración con los hospitales de Galdakao-Usansolo y Basurto pertenecientes a Osakidetza (Servicio Vasco de Salud). Osakidetza participa mediante la provisión de los textos y retroalimentándo el trabajo técnico con su experiencia, así como expertos que anotan el corpus para la obtención del gold-standard.
Palabras clave: Transferencia tecnológica, minería de textos clínicos, reconocimiento de entidades

1 Introduction

Typically, clinical documentation is produced in natural language on a free text basis, basically without or with little structure. From the research point of view, clinical text pro-

∗ This work was partially supported by the Spanish Ministry of Science and Innovation (EXTRECM: TIN2013-46616-C2-1-R, TADEEP: TIN2015-70214-P) and the Basque Government (DETEAMI: M inistry of Health 2014111003, IXA Research Group of type A (2010-2015), ELKAROLA: KK-2015/00098).

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

cessing dates from the early eighties (Friedman et al., 1983). There are easy understandable examples of the interest of natural language processing in this domain such as in (Taira, Soderland, and Jakobovits, 2001) where they proposed an approach to extract valuable information in the framework of radiology. The information extracted can be used as a decision support system.

The research goes ahead with evidences of the potential use of these techniques, while it is steadily being implemented in the hospitals. The goal of this project is twofold: on the one hand, to develop and transfer NLP technologies to the clinical domain and, on the other hand, to extract automatically adverse drug reactions from electronic health records. According to the World Health Organization, an Adverse Drug Reaction (ADR) "is a response to a drug which is noxious and unintended, and which occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease...".

The detection of ADRs is a key issue for the pharmacy and pharmacovigilance services that attempt to prevent medicine-related adverse effects in humans, in order to promote patient safety, but also the rational use of medicines. (Henriksson et al., 2015) state that ADRs cause the 3-5% of hospital admissions world-wide.

While there is some research in the clinical language processing, mainly in the language used in journals, there are still few applications integrated in the health services and that deal with the language used in patient's records. One of the reasons, is the difficulty to access information about patients due to privacy and confidentiality restrictions. For the Deteami project, although all the records were collected without any private data about the patients (without any name, age, address...) three ethical committees were passed and a confidentiality agreement was signed between the UPV-EHU University and Osakidetza.

From the point of view of computational linguistics, adverse drug reactions can be represented as a pair of entities (drug, disease) in which the drug was the causative agent of the disease. Hence, the task can be formulated as finding cause-effect events from a drug to a disease.

To work on this task, a set of real data was collected and manually annotated by two experts. This corpus serve as i) a gold-standard for the development of FreeLing-Med, a clinical entity recognition tagger and ii) training and evaluation sets for an ADR event retrieval system. The ADR retrieval tool shall serve to the aid of the Pharmacy and Surveillance services in their task of detecting adverse drug reactions. Besides, it will be an attempt to carry out technological transference. To do so, the prototype shall be validated by virtue of experts from the two involved hospitals. The ultimate goal of such a prototype would be to contribute in the early diagnosis of diseases, with its impact on the wellbeing of the society.

An added value to this project is that it shall be developed in Spanish, while there are few tools available for the clinical domain.

2 *Materials and Methods*

Our first purpose has been to acquire a corpus representative of the ADR event detection task and get it annotated by experts (clinicians, pharmacists, etc.) following the process in (Oronoz et al., 2015). In parallel, we are developing FreeLing-Med (Oronoz et al., 2013). Together with this, we mean to use NegEx adapted to this domain in order to detect negation and hence, help discarding negated events early. While we focus on the semantic tags provided by FreeLing-Med that correspond to medical entities, the remaining information would feed the subsequent stages, mainly the event retrieval system. Figure 1 summarizes the processes involved in the Deteami project, each of which is explained in the following sub-sections. As a result, the system extracts structured information from free text in the clinical domain in Spanish.

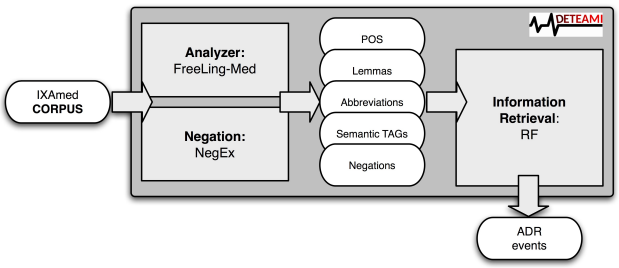


Figure 1: Deteami project overview.

2.1 Corpus acquisition

While there are several efforts in the literature that aim at the clinical domain, due to the lack of data, the majority focus on clinical journals and abstracts from PubMed. There are a few examples in the literature aiming at clinical text notes, the most of them for English, but there is also a remarkable one for Swedish (Dalianis et al., 2012). For the Spanish language this is the first corpus composed of real electronic records, we are aware of. In the Deteami project, the two hospitals involved, the Galdakao-Usansolo and Basurto Hospitals, contribute with medical records. Each of the hospitals make use of different platforms to store the data and most of the records are fully unstructured. At the IXA group, we collect and classify the records, we properly encode them and analyze them with FreeLing-Med. After that, we filter the terms indicating disorders and drugs, and we obtain the information that is shown to the human annotators in the "Brat Rapid Annotation Tool" format. In this way, the expert annotators following the guidelines can create the gold-standard.

2.2 Analyzer

A key issue in the detection of (drug, disease) events is to appropriately recognize clinical entities (particularly: drugs, signs, symptoms, substances, etc.).

There are a few examples of analyzers well adapted for the clinical domain, an example of them is GENIA (Tsuruoka et al., 2005), an analyzer for English. There is a version of MetaMap Transfer (MMTx) for Spanish (Carrero et al., 2008) that translates into English the text in Spanish by means of Google Translate and next applies the English version of MMTx in order to extract the medical entities that appear in the English SNOMED-CT. The arising question is the impact of error propagation.

In this project, instead, we adapted a general purpose linguistic analyzer FreeLing (L. Padró, 2011) to the medical domain. We enhanced the dictionaries with samples from the manually annotated corpus, terminology from standard medical ontologies (SNOMED CT, CIE-9-CM, ATC classification...), abbreviations within the medical domain and also tackle ambiguity. This tool analyzes the records and obtains valuable features such as tokens, lemmas, POS tags and semantic tags (disorder, body part...). As we said before, some of these features are shown to the annotators, but also serve for the event detection classifier.

2.3 Negation and speculation

Negation detection is crucial to discard early several potential ADR pairs. The negation in the medical domain has been tackled with two approaches:

1. Rule-based: NegEx (Chapman et al., 2001) is an outstanding toolkit in this area. It consists of a set of regular expressions built automatically from trigger phrases (pre-, post- and pseudo-negation) that can negate a clinical finding on which the negation is focused. It provides competitive results with 94.5% precision in clinical abstracts. The tool was adapted for Spanish as well (Costumero et al., 2014).

2. Machine Learning: there are alternatives that mean to infer negation patterns, such as (Averbuch et al., 2004). While this technique is language-independent, it was tested for English with an F1-score of 99.7%.

We are studying the approach that fits better the kind of texts we have, and we started using NegEx.

2.4 Information retrieval

The approaches used to extract information from clinical texts can be divided into two main groups: those that make use of rules and those based on machine learning. Rules, in comparison with inferred classifiers, tend to have better precision while lower recall. Our idea is to combine both methods in an hybrid system. First, we aim to use Kybots or abstract schemas defined in the Kyoto project (Vossen et al., 2008), to define adverse drug event patterns. Second, for event retrieval, we are making use of random forests, a supervised ensemble classification strategy.

3 Concluding remarks

In this project two hospitals attached to Osakidetza and a research group with experience in NLP are cooperating in an attempt to develop and transfer technological solutions based on NLP to the clinical domain,

and hence, fill the technological gap in clinical text mining in Spanish. The project is carried out for real clinical texts that lack of structure and so as to support language technologies in Spanish. This is the first year out of three of the Deteami project. Some of the mentioned tools have been already developed but are being improved. That is the case of the analyzer FreeLing-Med, currently in use but under development in order to improve i) the identification of non standard or local medical language and ii) the disambiguation of semantic tags and abbreviations. The detection of negation and speculation is in its early stage. Some experiments in the retrieval of adverse drug events have been carried out, but need to solve the problem of having an unbalanced set of disorder-drug pairs (there are very few pairs indicating ADRs and many indicating prescriptions). All the analyzers will improve with the manual annotation of new medical records. The two expert clinicians that are already annotating the corpus will continue with this task and reviewing the results automatically obtained for one more year. In this way, both the health system and the NLP researching group will benefit for this fruitful collaboration.

References

Averbuch, M., T.H Karson, B. Ben-Ami, O. Maimon, and L. Rokach. 2004. Context-sensitive medical information retrieval. *MedInfo*, page 282.

Carrero, F.J., J. Carlos Cortizo, J.M. Gómez, and M De Buenaga. 2008. In the development of a spanish metamap. In *Proceedings of the 17th CIKM conference*, pages 1465–1466. ACM.

Chapman, W.W., W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Costumero, R., F. Lopez, C. Gonzalo-Martín, M. Millan, and E. Menasalvas. 2014. An Approach to Detect Negation on Medical Documents in Spanish. In *Brain Informatics and Health*. Springer, pages 366–375.

Dalianis, H., M. Hassel, A. Henriksson, and M. Skeppstedt. 2012. Stockholm EPR corpus: A clinical database used to improve health care. In *Swedish Language Technology Conference*, pages 17–18. Citeseer.

Friedman, C., N. Sager, E.C. Chi, E. Marsh, C. Christenson, and M. S. Lyman. 1983. Computer Structuring of Free-Text Patient Data. In *Symposium on Computer Applications in Medical Care*, pages 688–691. American Medical Informatics Association.

Henriksson, A., M. Kvist, H. Dalianis, and M. Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics*, 57:333–349.

L. Padró. 2011. Analizadores Multilingües en freeling. *Linguamatica*, 3(2):13–20, December.

Oronoz, M., A. Casillas, K. Gojenola, and A. Pérez. 2013. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. *Lecture Notes in Computer Science*, 8259:536–547.

Oronoz, M., K. Gojenola, A. Pérez, A. Díaz de Ilarraza, and A. Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56(0):318 – 332.

Taira, R., S. Soderland, and R. Jakobovits. 2001. Automatic structuring of radiology free-text reports 1. *Radiographics*, 21(1):237–245.

Tsuruoka, Y., Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392.

Vossen, P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, and J. VanGent. 2008. KYOTO: a System for Mining, Structuring and Distributing Knowledge across Languages and Cultures. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 23–37, may.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 159-162 recibido 29-03-2016 revisado 06-04-2016 aceptado 29-04-2016

TALENT+ Tecnologías avanzadas para la Gestión del Talento

TALENT+ Advanced Technologies for Talent Management

Julio Villena Román, José Carlos González Cristóbal, José Antonio Gallego Vázquez
s|ngular
E-28034 Madrid, Spain
{jvillena, jgonzalez, jagallego}@sngular.team

Resumen: TALENT+ es un proyecto de I+D+i cuyo objetivo es el desarrollo de un sistema inteligente avanzado de soporte a la toma de decisiones para la gestión del talento, orientado a cualquier organización que busque identificar, atraer y retener al mejor talento disponible, entender cuáles son los factores clave para motivar a sus empleados y maximizar su rendimiento, y quiera estar alerta sobre cualquier posible cambio que altere el clima de convivencia en la organización. La solución se compone de dos módulos: el sistema basado en conocimiento, que codifica la experiencia, forma de trabajo y mejores prácticas de los gestores, y el sistema inteligente que explota este conocimiento aplicando técnicas de análisis de datos.
Palabras clave: gestión del talento, sistema de soporte a la decisión, análisis de datos, RR.HH.

Abstract: The objective of TALENT+ project is the development of an advanced intelligent decision support system for talent management, aimed at any organization that seeks to identify, attract and retain the best talent available, understanding the keys to motivate employees and maximize their performance, and want to be aware of any changes that alter the coexistence in the organization. The solution consists of two modules: the knowledge based system, which encodes the experience, methods and best practices of managers, and the intelligent system that exploits this knowledge using data analysis techniques.
Keywords: Talent management, decision support system, data analytics, HR.

1 Introducción

Uno de los retos más importantes y ambiciosos surgidos en la actual sociedad es la llamada *gestión del talento*, una nueva forma de abordar las funciones de RR.HH. que pretende seleccionar la persona adecuada para el trabajo adecuado en el momento adecuado. En este contexto, se considera *talento* a cualquier persona que tiene la capacidad de producir una diferencia significativa en el rendimiento actual y futuro de la organización (Lynne, 2005).

La identificación del talento se ha convertido en uno de los retos principales de la gestión de RR.HH. (Tower Perrin, 2005). El proceso consiste en reconocer las áreas de talento clave en la organización, identificar las personas que constituyen su talento, atraer e incorporar personas con talento externas a la organización y realizar actividades de desarrollo del pool de talento para cubrir las necesidades de la organización, reteniendo el talento y aumentando el compromiso de estas personas clave con la organización para que asuman funciones más importantes (Cubbingham, 2007).

En este contexto surge TALENT+, proyecto propio del plan de I+D+i de la división Data & Analytics de s|ngular, planificado a dos años y que pretende generar para la compañía una propuesta de servicios avanzados, diferentes y exclusivos, en torno a la gestión del talento, en el que entren en juego tecnologías y conceptos innovadores que aún no han sido de aplicación en la gestión de los RR.HH.

La gestión del talento implica decisiones de negocio que suelen afectar a diferentes niveles en la organización. El proceso actual de toma de decisiones se basa en la experiencia humana de los gestores, su conocimiento, preferencias y juicios personales, de carácter incierto y difíciles de optimizar. Estos factores pueden causar inconsistencias, imprecisiones, desigualdades y decisiones imprevistas.

Sin embargo, la tecnología actual permite aplicar un enfoque más sistemático. Cada día el empleado de una empresa emite múltiples

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

señales que miden su rendimiento, grado de satisfacción respecto a su trabajo y compromiso con la organización. La tecnología ha simplificado enormemente la recolección de parámetros tanto objetivos (control de presencia, tiempo en reuniones, reservas de salas, calendarios compartidos…) como de carácter subjetivo (*feedback* del empleado sobre la empresa con diferentes métricas propias de la organización o estandarizadas como el Employee Net Promoter Score, mensajes del trabajador en redes sociales, etc.).

Así, el objetivo del proyecto es el diseño y desarrollo de un sistema inteligente de soporte a la toma de decisiones, elemento clave de automatización integrado perfectamente en el conjunto de herramientas tecnológicas empleadas en la organización, para ayudar a los responsables de la gestión del talento durante las diferentes fases de la toma de decisiones, mediante la integración de herramientas de modelado y conocimiento humano, sobre todo en aquellas áreas donde existe incertidumbre o información incompleta y donde las decisiones que involucran riesgo deben realizarse utilizando juicios humanos y preferencias.

El proyecto está orientado a cualquier organización que busque identificar, atraer y retener al mejor talento disponible, entender cuáles son los factores clave para motivar a sus empleados y maximizar su rendimiento, y quiera estar alerta sobre todo cambio que altere el clima de convivencia en la organización.

El sistema planteado en TALENT+ implica unos retos científico-tecnológicos de primer nivel, que, en general, se pueden describir como i) estudiar la mejor manera de aplicar las técnicas de análisis de datos para detectar y predecir el talento en una organización, y ii) definir la arquitectura tecnológica del sistema inteligente de gestión del talento para soporte a la toma de decisiones en RR.HH.

2 Estado del arte

El estado actual de la tecnología en RR.HH. se centra básicamente en dominios de gestión específicos, tales como selección de personal, la formación y la evaluación del rendimiento laboral. La mayoría de las aplicaciones son sistemas de gestión, sin inteligencia automática subyacente, o bien son sistemas expertos, con reglas manuales, sin capacidad de adaptación o aprendizaje. Los sistemas basados en conocimiento representan una gran oportunidad para mejorar la práctica de la gestión de RR.HH. (Martinsons, 1995), ya que con ellos las tareas son automatizables de manera más estable, más fáciles de replicar, menos costosas y se documentan de forma automática. Como desventaja, tienen problemas con el conocimiento informal, difícil de verbalizar.

Por citar algunos, se han empleado sistemas de este tipo para la selección de personal (Hooper, 1998), en sistemas guiados para formación online (Chen, 2007), o, combinados con el paradigma de agentes, para planificación de recursos y horarios (Glenzer, 2003).

Por otra parte, las técnicas de análisis de datos, englobadas bajo la disciplina hoy denominada Data Science, están viviendo un nuevo resurgir que puede realizar grandes aportaciones en este campo, por la posibilidad de convertir todos datos de RR.HH. en información y conocimiento útil (Han, 2006). Sin embargo, aunque el análisis de datos se aplica en numerosos campos como finanzas, marketing, fabricación, etc., su aplicación en RR.HH. todavía es muy poco frecuente.

La figura siguiente presenta el diagrama arquitectónico de un sistema de RR.HH. para predicción del talento (Hamidah, 2009). En este trabajo, el sistema realizaba predicciones mostrando el talento potencial estimado para un determinado empleado, detallando las razones justificando dicha predicción, y sugiriendo las tareas más adecuadas para dichas capacidades.

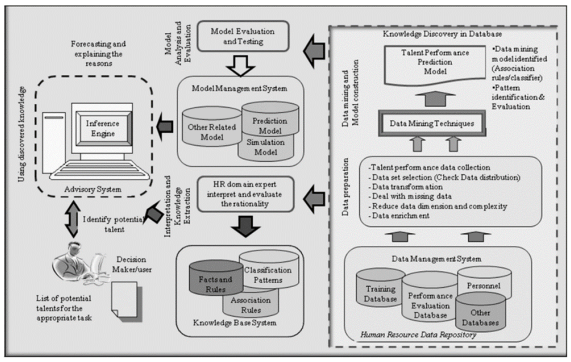


Figura 1: Arquitectura del sistema de RR.HH. para predicción del talento (Hamidah, 2009)

TALENT+ va un paso más allá, ya que se plantea un IDSS de naturaleza híbrida, que por un lado incorporará conocimiento de dominio específico (a través de una base de casos, una base de reglas, un subsistema de adquisición de conocimiento o modelos de dominio concretos, en definitiva, la cultura de la organización

desde el punto de vista de los RR.HH) pero se diseñará para incorporar también mecanismos inteligentes, como razonamiento y aprendizaje, para la toma de decisiones (Viademonte, 2006).

El proceso básico sería la recopilación de conocimiento de la organización, para posteriormente aprender de los conocimientos adquiridos, mediante algoritmos análisis de datos, razonamiento sobre esos conocimientos y, finalmente, ante demandas de información, emitir una respuesta (predicción, clasificación, recomendación) justificando los resultados (con métricas objetivas).

3 Líneas de trabajo

El proyecto persigue una nueva definición completamente funcional, operativa y utilizable del perfilado del capital humano, más allá del modelo simple actual de triplas (persona, conocimiento/habilidad, valor) que se representa en los actuales Curriculum Vitae. La novedad es representar el talento representando cada persona como un grafo de conocimientos y habilidades, no aislados y transversales, sino relacionados y con medidas de distancia entre ellos que modelen hasta cuánto la persona es clave en la organización y su grado de implicación con su puesto de trabajo.

Esta representación del talento, adecuada para su procesamiento tecnológico, será la base para abordar las tres líneas de acción definidas en el proyecto, en tareas relacionadas con la identificación del talento de un organización (los empleados clave), la detección de grupos intrínsecos de empleados con similares características respecto a la gestión del talento en la organización, la predicción del rendimiento de un empleado en una determinada tarea/puesto de trabajo según su talento, la selección del candidato óptimo en un proceso de selección, la determinación del empleado más adecuado al perfil laboral/tarea indicados, el asesoramiento automatizado sobre la carrera profesional de los empleados de una organización y el análisis de la fuga de talento.

3.1 Identificación del talento

El objetivo es aplicar técnicas de análisis de datos para identificar qué variables son las que más impactan en el rendimiento y motivación de los trabajadores en el ámbito de una organización para actuar sobre ellas mejorando el entorno laboral y favoreciendo la identificación, retención y desarrollo del talento. El fin último es identificar aquellos trabajadores clave en la organización, por sus conocimientos o capacidades en su puesto de trabajo, o por su contribución al clima laboral.

3.2 Atracción de talento externo

El objetivo es ayudar a la organización en el proceso de selección (*recruiting*) del personal más apropiado. Con el denominado *data driven recruitment* se pretende identificar, mediante la aplicación de técnicas de análisis de datos, qué características, tanto personales como profesionales son garantía de éxito (o fracaso) específicamente en la organización (nivel de formación, experiencia, análisis psicométrico) y en base a las mismas realizar una ordenación de candidatos de mayor a menor relevancia (o grado de adecuación) en un proceso de selección de un puesto específico.

3.3 Desarrollo del pool de talento

El objetivo es ayudar a los gestores de RR.HH. en el proceso de desarrollo del pool de talento de la organización, para ofrecer a los empleados un plan de desarrollo de carrera profesional que ofrezca un entorno de trabajo agradable y motivador que potencie sus conocimientos y habilidades y evite la fuga de talento.

La primera línea de trabajo es la recomendación de acciones: clasificar a los empleados aplicando técnicas de análisis de datos en función de las variables que les motivan positiva y negativamente, encontrando clusters de empleados similares entre sí y diferentes a los de otros grupos. El escenario sería ayudar al *mentoring* de los empleados dado recomendaciones de acciones efectivas sobre su carrera laboral, en comparación con empleados y circunstancias similares.

La segunda línea es la monitorización activa: aplicar análisis de información no estructurada para la creación de canales de escucha automática de la "voz del empleado", con una serie de índices que avisan en tiempo real del estado de ánimo. Por último, la tercera línea es la predicción de fuga de talento: predecir con tiempo suficiente el posible abandono de un empleado y las causas que le llevan a ello, para tomar medidas para evitarlo.

4 Resultados preliminares

El proyecto ha comenzado recientemente, por lo que los resultados obtenidos todavía son muy

preliminares. En una primera fase, se ha desarrollado un modelo genérico de análisis de la Voz del Empleado (VoE) sobre la plataforma de analítica de contenidos MeaningCloud[1], que consiste en un conjunto de reglas semánticas para la extracción de *insights* relevantes para la interpretación de mensajes recogidos en encuestas de satisfacción en el trabajo.

Este modelo se ha utilizado internamente, por una parte, en estudios de competencias profesionales y satisfacción laboral con el análisis automatizado de encuestas a 300 empleados, y, por otra, para el procesamiento de entrevistas de salida para analizar motivos de abandono de la compañía. Los resultados son muy positivos en cuanto a precisión, cobertura y relevancia de las conclusiones obtenidas, proporcionando información muy valiosa.

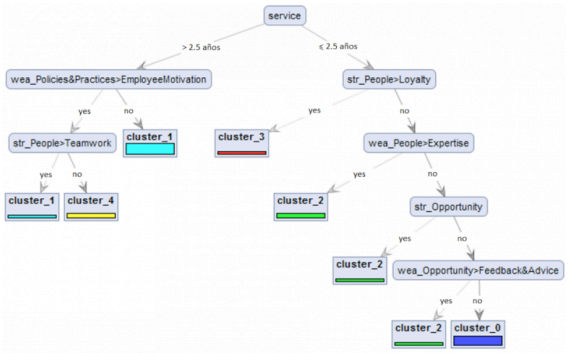


Figura 2: Análisis de entrevistas de salida

5 Conclusiones

TALENT+ es un proyecto ambicioso, una apuesta estratégica de futuro para s|ngular, que busca generar una propuesta de servicios avanzados, diferentes y exclusivos, en torno a la gestión del talento, incorporando tecnología al campo de los RR.HH. donde todavía no se han hecho grandes aportaciones. Estos nuevos servicios y tecnologías crearán un ecosistema de aplicaciones mucho más inteligentes, proporcionando una experiencia en la que el elemento primordial es el trabajador.

Es necesario un esfuerzo investigador importante en tecnologías que se encuentran en una fase de madurez temprana. Por ello, actualmente nos encontramos en una fase de búsqueda activa de financiación externa que impulse el esfuerzo investigador y de desarrollo incorporando otros grupos de investigación.

[1] MeaningCloud (www.meaningcloud.com) es una marca registrada y el nombre de la filial en EE.UU. de Singular Meaning SL, una compañía s|ngular (sngular.team).

Referencias

Chen, K.K., et al. 2007. Constructing a Web-based Employee Training Expert System with Data Mining Approach. 4th IEEE International Conference on Enterprise Computing, 2007.

Cubbingham, I. 2007. Talent Management: Making it real. Development and Learning in Organizations: An International Journal, Vol. 21 Iss: 2, pp. 4 - 6, 2007.

Glenzer, C. 2003. A conceptual model of an interorganizational intelligent meeting-scheduler (IIMS). Strategic Information Systems, 2003. 12(1): p. 47-70. , 2003.

Hamidah, J., H. Abdul Razak, and A.O. Zulaiha. 2009. Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application. International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering Vol:3, No:2, 2009.

Han, J. and M. Kamber. 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann Publisher, 2006.

Hooper, R.S., et al. 1998. Use of an Expert System in a personnel selection process. Expert Systems and Applications, 1998. 14(4): p. 425-432., 1998.

Lynne, M. 2005. Talent Management Value Imperatives: Strategies for Execution. The Conference Board, 2005.

Martinsons, M.G. 1995. Knowledge-based systems leverage human resource management expertise. International Journal of Manpower, 1995. 16(2): p. 17-34. , 1995.

Tower Perrin. 2005. TP Track Research Report Talent Management: State of the Art. 2005.

Viademonte, S. and F. Burstein. 2006. From Knowledge Discovery to Computational Intelligent: Framework for Intelligent Decision Support Systems. Springer London, 2006.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 163-166 recibido 29-03-2016 revisado 06-04-2016 aceptado 30-04-2016

eGovernAbility: Marco para el desarrollo de servicios personalizables accesibles en la Administración electrónica

eGovernAbility: Framework for the development of customizable accesible services in the Electronic Administration

Paloma Martínez, Lourdes Moreno
Departamento de Informática
Universidad Carlos III de Madrid
{pmf, lmoreno}@inf.uc3m.es

Julio Abascal, Javier Muguerza
Departamento de Arquitectura y Tecnología de Computadores
Universidad del País Vasco
{julio.abascal, j.muguerza}@ehu.eus

Resumen: El objetivo general del proyecto eGovernAbility es obtener una arquitectura de software basado en modelos para el desarrollo metodológico de servicios electrónicos (eServicios) inclusivos personalizados que permitan a cualquier usuario interactuar de una manera satisfactoria, sin importar el dispositivo utilizado. Esto requiere la integración en los modelos de técnicas apropiadas de perfilado de usuario para adaptar los eServicios a las características de los usuarios, a la tecnología disponible y a la funcionalidad del servicio. Por ello, este proyecto tiene un carácter multidisciplinar y se abordará mediante la colaboración entre investigadores en tecnologías de la información y expertos de eAdministración, combinando diversas disciplinas científicas: Modelado de servicios para eGovernment, minería de datos, interacción adaptada al usuario (incluyendo accesibilidad universal y acceso multidispositivo), procesamiento de lenguaje natural (PLN) y arquitecturas de software basado en modelos.
Palabras clave: eGobernanza, eServicios, minería de datos, aprendizaje automático, perfilado de usuario, accesibilidad, adaptabilidad, diseño de interfaces.

Abstract: The aim of eGovernAbility project is to produce a model-based software architecture for methodologically developing personalized inclusive public on-line services (eServices) that allow any user to interact with them in a satisfactory way, no matter the device used. This requires integrating appropriate user profiling and adaptation techniques into the model to tailor eServices to users' characteristics, available technology, and service's functionality. Hence, this project has a multidisciplinary nature and will be addressed through collaboration between researchers and experts in information technology and professionals of the eGovernment, combining diverse scientific backgrounds: Modelling eGovernment services, Data mining, human-computer interaction for user tailored interfaces (including universal accessibility and multidevice access), Natural Language Processing (PLN) and Model based software architectures.
Keywords: eGovernment, eServices, web mining, user profiling, accessibility, adaptability, model-based user interface design.

1 Introducción

Las administraciones públicas avanzan rápidamente hacia la provisión a través de la Web de servicios para el ciudadano básicos y extendidos (Informe eGovernment de la UE de 2014). Además de la disminución de los costes, este esfuerzo apoya el derecho de acceso a los servicios públicos para todas las personas (incluyendo personas con discapacidad y de edad avanzada). La UE ha lanzado iniciativas para "*Satisfacer nuevas necesidades de la sociedad mediante el uso de las nuevas tecnologías en el sector público*" con el fin de "*promover servicios públicos eficientes y abiertos, centrados en los ciudadanos*". Se han hecho esfuerzos para mejorar la usabilidad y accesibilidad de los sitios web de las administraciones, pero diversos estudios han revelado que no es suficiente. El objetivo

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

general del proyecto es definir una arquitectura de software basado en modelos para el desarrollo metodológico de eServicios inclusivos personalizados que permitan a cualquier usuario interactuar de una manera satisfactoria, sin importar el dispositivo utilizado. El proyecto eGovernAbility[1] tiene una duración prevista de 3 años (2015-2017) y se estructura en dos subproyectos:

- Subproyecto 1: liderado por la UPV/EHU, que además es coordinadora del proyecto, explotará datos reales de interacción de usuario con eServicios prestados por la Diputación de Gipuzkoa (DFG) para extraer patrones de uso, uso anormal de los servicios y barreras de accesibilidad. Se construirán modelos de eServicios a partir de los cuales se proporcionará acceso a la Web mediante adaptaciones de presentación, contenido y navegación.
- Subproyecto 2: liderado por el grupo LABDA[2] de la UC3M, creará una arquitectura de software para el desarrollo basado en modelos de aplicaciones de eAdministración con soporte para la accesibilidad y el uso de múltiples dispositivos. Se aplicarán técnicas de simplificación de textos para la adaptación de contenidos en web.

Los resultados del proyecto incluirán modelos de aplicaciones de eAdministración, usuarios de eServicios y adaptaciones web para la accesibilidad universal y multidispositivo que se integrarán en la arquitectura basada en modelos para la creación de herramientas de desarrollo de aplicaciones de eAdministración accesibles y de calidad.

2 Antecedentes

La administración electrónica (eAdministración) se ha convertido en una tendencia adoptada en muchos países. Dado que la eAdministración es un área importante de las TIC, los gobiernos intentan incorporarla en sus sistemas de información y procesos de gobierno. En septiembre de 2011 España estaba entre los diez países más avanzados en esta área y en quinto lugar a nivel europeo en términos de disponibilidad y sofisticación de servicios públicos on-line, ver SIPA (2011). La Ley de acceso electrónico de los ciudadanos a los Servicios Públicos de junio de 2007 tiene como objetivo que los ciudadanos pudieran acceder a todos los servicios públicos y gestionar la documentación administrativa utilizando internet desde cualquier sitio en cualquier momento.

Desde el punto de vista regulador en materia de accesibilidad, se debe seguir en España y en la mayoría de los países el estándar en accesibilidad de referencia *Web Content Accessibility Guidelines (WCAG) 2.0* (W3C, 2008) (ISO, 2012). Este estándar indica cómo desarrollar los sitios web tal que puedan ser accedidos por cualquier persona, incluyendo las personas con discapacidad. Las administraciones públicas en concreto deben cumplir con este estándar desde el punto de vista legislativo.

Sin embargo, el cumplimiento de este estándar de accesibilidad no garantiza una experiencia de usuario satisfactoria en la Web En este sentido, la personalización de las interfaces de acuerdo a las necesidades de los usuarios parece ser un método efectivo para superar las barreras de accesibilidad y asegurar una experiencia satisfactoria en la Web. La eAdministración puede beneficiarse de este método con el fin de desarrollar eServicios que sean accesibles, fáciles de usar, efectivos y diseñados para responder a las necesidades de todos los ciudadanos.

Por último, la sistematización de la captura y representación de los requisitos asegurando la portabilidad desde diferentes interfaces de usuario abstractos a diferentes dispositivos (multidispositivo) y diversas plataformas podría beneficiar el desarrollo de eServicios. Estos retos serían posibles si se proporcionara un entorno metodológico a las partes interesadas que incluya una arquitectura para el desarrollo e integración de servicios, que es precisamente el objetivo principal del proyecto eGovernAbility.

3 Estructura del proyecto

El proyecto se organiza en 5 paquetes de trabajo.

WorkPackage01: Identificación de los sistemas de la eAdministración a modelar. Según el estado de la cuestión se llevará a cabo un análisis preliminar de los datos de uso y trabajo con el usuario del proyecto, la Diputación foral de Guipuzcoa, que permitirá identificar los eServicios a modelar y adaptar.

[1] Proyecto TIN2014-52665-C2-1-R, https://egovernability.wordpress.com/

[2] labda.inf.uc3m.es

WorkPackage02: Extracción de patrones de comportamiento de usuario. Se utilizarán técnicas de minería de datos para descubrir patrones de uso en la interacción de los usuarios con los eServicios. Se extraerán distintos tipos de patrones de uso como los caminos frecuentes, URLs de entrada/salida, accesos rápidos versus accesos lentos, etc. tanto para los datos de uso común como para los casos especiales con interacciones problemáticas.

WorkPackage03: Generación de perfiles dinámicos de usuario para personalización multidispositivo de eServicios públicos accesibles. Se generará un modelo de perfiles de usuario apropiado para el desarrollo de eServicios públicos accesibles personalizados y adaptativos. Se analizarán diversas soluciones de personalización que incluyan adaptaciones de contenido, de navegación y de presentación de la información, sin olvidar la personalización de interfaces móviles. Es precisamente en la adaptación de contenidos donde se están aplicando técnicas PLN para simplificación según distintos tipos de discapacidad.

WorkPackage04: Detección de anomalías para soporte a usuarios y administradores. Los patrones extraídos del uso de la Web y de los eServicios permitirán identificar los problemas en la interacción. Ello permitirá ayudar tanto a los usuarios como a los administradores de los sistemas. Por otro lado, el conocimiento extraído se utilizará para enriquecer los perfiles de usuario y mejorar la experiencia de usuario.

WorkPackage05: Marco formal metodológico basado en la experiencia adquirida. Se definirá un marco metodológico para ayudar a los profesionales en el desarrollo de interfaces de usuario de eServicios accesibles. Se analizarán varios enfoques MDD (model driven development) y MBD (model-based development) para definir interfaces de usuario abstractos y concretos teniendo en cuenta los requisitos según los estándares de accesibilidad y las necesidades de personalización de los eServicios.

4 Marco formal metodológico

Con el objetivo de diseñar sistemas inclusivos de calidad se tiene como espacio de solución el uso de enfoques metodológicos que proporcionen servicios públicos robustos teniendo en cuenta las necesidades de los usuarios, en especial de las personas con discapacidad y personas mayores.

Como elemento integrador en el proyecto se incluye un espacio de trabajo que comprende un soporte metodológico con una arquitectura en el desarrollo de interfaces de usuarios de eServicios inclusivos y personalizados.

4.1 Enfoque metodológico

Ante la diversidad de factores a considerar en el desarrollo de un sistema de acceso público como son la heterogeneidad de plataformas, dispositivos, necesidades especiales de las personas, modalidades de interacción, etc. se está utilizando como herramienta metodológica enfoques MDD y MBD integrados en arquitecturas dirigidas por modelos como la que se muestra en la figura 1.

Esta arquitectura está estructurada de acuerdo a los diferentes niveles de abstracción definidos en este caso por el marco de referencia Cameleon (Calvary et al., 2003). Con esta aproximación se proporcionan mecanismos de diseño capaces de modelar los requisitos de accesibilidad definidos en el proyecto con un único diseño, el cual genere interfaces finales que den respuesta a toda la heterogeneidad de tecnología, plataforma, dispositivos, modalidades de interacción, etc.

Este enfoque apoya la creación de servicios accesibles desde las aplicaciones de la eAdministración de manera sostenible ya que se consigue una sistematización de los requisitos de accesibilidad y personalización en todo el ciclo de vida del desarrollo.

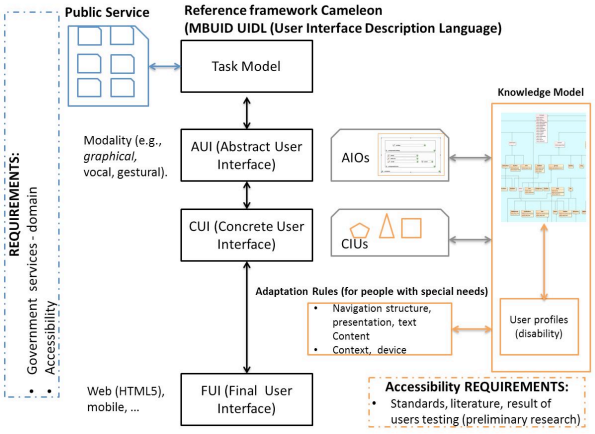


Figura 1: Arquitectura dirigida por modelos del proyecto eGobernAbility

4.2 Accesibilidad siguiendo WCAG 2.0

Con el objetivo de diseñar servicios accesibles se hace un tratamiento de la accesibilidad

conforme al estándar de referencia (WCAG) 2.0. Este estándar se tiene en cuenta de manera esencial en este proyecto, los requisitos según las WCAG 2.0 se integran desde el diseño en los distintos niveles de abstracción de la arquitectura.

Por otro lado, el cumplir con este estándar no garantiza necesariamente una experiencia de usuario satisfactoria, por ello que en este proyecto se están teniendo en cuenta técnicas de personalización de la interfaz que mejore la experiencia de usuario al acceder a servicios en la eAdministración.

4.3 Personalización de Interfaces de usuario

La personalización de los servicios de la eAdministration tiene sus propias particularidades ya que los servicios presentan mayor dificultad que un sitio web informativo. Aspectos tales como necesidades de los usuarios, el nivel de experiencia, familiaridad con el dominio o nivel de dificultad de la tarea se deben tener en cuenta. Aunque existen ejemplos de sitios web accesibles como Discapnet con noticias en lectura fácil (www.noticiasfacil.es), el *e-Journal of Inclusion Europe* con resúmenes de noticias en lectrua fácil (www.e-include.info), y la Wikipedia Simple en Inglés (simple.wikipedia.org), todavía hay mucho margen para la mejora principalmente en servicios de la eAdministración.

Con este objetivo, se han definido reglas de adaptación y personalización de las interfaces (Valencia et al., 2013) (Moreno et al., 2015). En estas adaptaciones se aborda la personalización de contenido, estructura y presentación para distintas discapacidades como la visual, auditiva, física y cognitiva. Con ellas se favorece la inclusión de personas ciegas, con baja visión, sordas, con hipoacusia, personas mayores, etc. en la eAdministración.

Las adaptaciones relacionadas con las barreras de accesibilidad sensoriales se pueden abordar aplicando técnicas dadas en las WCAG 2.0, sin embargo para abordar reglas de adaptación que eviten barreras de accesibilidad en las personas con discapacidad cognitiva es necesario utilizar técnicas de otras disciplinas como es el Procesamiento de Lenguaje Natural (PLN).

Las técnicas de PLN dan soporte a la transformación del contenido textual en textos más sencillos que faciliten la comprensión y lectura a las personas, (Saggion et al., 2011). Las técnicas utilizadas en el proyecto siguen un enfoque de simplificación léxica de textos (Moreno et al., 2015). Con esta personalización de las interfaces de los servicios públicos se quieren evitar las barreras cognitivas que puede tener un ciudadano, por ejemplo una persona de edad avanzada al realizar una gestión en la eAdministración en la que no comprende los pasos a seguir y su contenido por la complejidad del texto en términos de compresión y facilidad de lectura.

Bibliografía

Calvary, G. ; Coutaz, J. ; Thevenin, et al. 2003.A Unifying Reference Framework for Multi-Target User Interfaces. In Interacting with Computer, p.289–308, 2003.

ISO/IEC 40500:2012, Information technology -- W3C Web Content Accessibility Guidelines (WCAG) 2.0, http://www.iso.org

Moreno, L., Martínez, P., Segura-Bedmar, I., Revert, R.. Exploring language technologies to provide support to WCAG 2.0 and E2R guidelines. Interacción '15. ACM, New York, NY, USA, , Article 57, (2015)

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

SIPA (Spanish Institute of Public Administration). eGovernment in Spain. Report to the European Commission. www.epractice.eu/files/eGovernmentSpain.pdf (2011).

Valencia, X., Arrue, M., Pérez, JE., Abascal, J.. User individuality management in websites based on WAI-ARIA annotations and ontologies. W4A, (2013).

W3C, WAI, Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation 11 December 2008. http://www.w3.org/TR/WCAG20/

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 167-170 recibido 30-03-2016 revisado 06-04-2016 aceptado 01-05-2016

Extracción de contextos definitorios en el área de biomedicina

Extraction of Definitional Contexts from Biomedical Corpora

César Aguilar[a], Olga Acosta[b], Gerardo Sierra[c]
Sergio Juárez[d], Tomás Infante[b]
[a]Católica de Chile - Campus San Joaquín, Santiago de Chile caguilara@uc.cl
[b]Cognitiva –IBM - Apoquindo No. 5400, Las Condes, Santiago de Chile {oacosta, tinfante}@cognitiva.la
[c]Grupo de Ingeniería Lingüística - Instituto de Ingeniería, UNAM - Torre de Ingeniería, CU, Ciudad de México gsierram@ii.unam.mx
[d]Facultad de Estadística e Informática, Universidad Veracruzana - Xalapa, Veracruz, México sejuarez@uv.mx

Resumen: En este proyecto se formula una metodología para extraer contextos definitorios desde corpus de biomedicina en español, con el fin de generar los siguientes productos: (i) un listado de candidatos a términos, (ii) un listado de candidatos a definiciones, y (iii) una taxonomía de términos biomédicos basada en relaciones de hiponimia/hiperonimia. Nuestro método permite crear un sistema capaz de extraer tales contextos, el cual puede verse como un módulo que cubriría las primeras etapas a seguir para construir una ontología basada en información textual.
Palabras clave: Contexto definitorio, término, definición, extracción de información, taxonomía.

Abstract: In this project we formulate a methodology for extracting definitional contexts from corpus of biomedicine in Spanish, in order to generate the following products: (i) a list of candidate terms, (ii) a list of candidates for definitions, and (iii) a taxonomy of biomedical terms relationships based on hyponym/hyperonym. Our methodology allows the creation of a system capable of extracting such contexts, which can be seen as a module that would cover the first steps to follow to build an ontology based on textual information.
Keywords: Definitional Context, Term, Definition, Information extraction, Taxonomy.

1 Introducción

Debido al incremento de información, la biomedicina tiene un gran interés en el desarrollo de herramientas que le ayuden a identificar, extraer y clasificar tal información, con miras a obtener conocimientos relevantes. Por ello, la Biblioteca Nacional de Medicina (NLM) de Estados Unidos desarrolló la base de datos *MedLine*, la cual cuenta con un total aproximado de 21 millones de referencias a artículos provenientes de 4.500 revistas. Para acceder a tal repositorio, se ha implementado el motor de consulta *PubMed*[1], el cual brinda acceso gratuito a sus datos. La implementación de esta clase de recursos ha hecho que la biomedicina mantenga una estrecha relación con el área de procesamiento del lenguaje natural (PLN).

A pesar de los avances logrados por parte del PLN en la explotación de *MedLine*, (mayoritariamente en inglés, aunque también hay aportes en español), en Chile tales logros han tenido un bajo impacto hasta hoy.

De acuerdo con la *Estrategia Nacional de Salud 2010-2020*[2], elaborada por el Ministerio de Salud, existen varias limitaciones dentro de los sistemas de gestión de información médica, entre las cuales cabe destacar: (i) ausencia de normativa para la generación y recepción de información; (ii) escasa infraestructura tecnológica; (iii) ausencia de sistemas en línea, y (iv) errores en el traspaso manual de la información.

2 Objetivos

El objetivo de este proyecto es diseñar un sistema de extracción de contextos definitorios (CDs) en el área de biomedicina en español.

[1] Al respecto, véase el siguiente sitio WEB: www.ncbi.nlm.nih.gov/pubmed.

[2] Al respecto, véase el siguiente sitio WEB: http://web.minsal.cl/portal/url/item/c4034eddbc96ca6de0400101640159b8.pdf.

Para diseñar este sistema, se toma en cuenta la metodología para extraer CDs desde corpus planteada por Sierra *et al*. (2008), y Sierra (2009), la cual considera los siguientes aspectos:

- El análisis lingüístico respecto a la estructuración de un CD.
- La implementación de un sistema de búsqueda, el cual genere como *output*: (a) un conjunto de candidatos a términos, (b) un conjunto de candidatos a CDs, clasificados según el tipo de definición asociada, y (c) una posible taxonomía de términos jerarquizados conforme a relaciones léxicas identificables entre ellos.
- El uso de métodos probabilísticos para evaluar la eficacia del sistema.

3 Descripción del proyecto

Definimos un CD como fragmentos textuales que contengan términos y definiciones ligadas por predicaciones verbales (Sierra *et al*., 2008; Sierra, 2009). Un ejemplo es el siguiente:

> Generalmente [marcador discursivo], [la **célula** Término/Marcador tipográfico] [puede definirse como Frase predicativa] [una porción de protoplasma individualizado, dotado de núcleo y de una membrana plasmática, que nace, crece, se reproduce y muere Definición]

En este ejemplo, se pueden reconocer los componentes básicos de un CD: un término, una definición y una frase predicativa que opera como conector entre las unidades anteriores. A estas unidades se añaden marcadores discursivos (el adverbio *generalmente*) o marcadores tipográficos (signos ortográficos o tipos de fuente).

Una forma de representar lo anterior es el esquema de la figura 1, en donde se muestra la configuración de un CD en torno a sus componentes básicos (término, frase predicativa y definición), así como opcionales (marcadores discursivos y tipográficos). La combinación de estos elementos es lo que proyecta la armazón lingüística de un CD dentro de un texto, lo que facilita la detección automática de términos y definiciones circunscritas a dicha armazón.

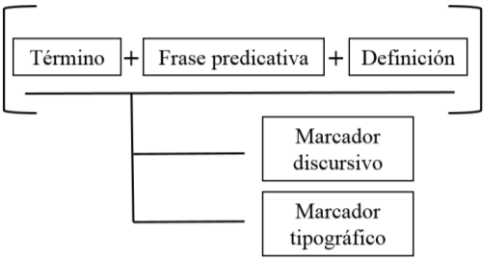


Figura 1: Estructura de un CD

Un aspecto relevante a observar aquí es que las definiciones asumen rasgos específicos para expresar el concepto asociado a un término, dependiendo del tipo de verbo que aparece en la frase predicativa. Estos rasgos tienen que ver con la formulación de los dos componentes básicos: género próximo y diferencia específica. Tomando en cuenta esto, Sierra *et al*. (2008) proponen 4 tipos de definiciones, derivables del modelo analítico:

- **Definición analítica o aristotélica:** se da cuando el género próximo y la diferencia específica aparece de manera explícita dentro de una definición.
- **Definición sinonímica:** se da cuando en una definición se hace explícito el género próximo, estableciendo una equivalencia conceptual con el término que es definido.
- **Definición funcional:** se da cuando se hace explícita la diferencia específica, ofreciendo una definición de un término a partir de su uso o aplicación en una situación dada.
- **Definición extensional:** se da cuando se hace explícita la diferencia específica, presentando una definición que enumera los componentes que conforman un objeto representado por el término a definir. Esta enumeración de componentes sigue un orden basado en relaciones que van de un todo hacia las partes, o de las partes hacia el todo.

Estos 4 tipos brindan la posibilidad de establecer una taxonomía que se puede reforzar a partir de la corroboración de relaciones léxicas existentes en CDs, a saber: hiponimia/hiperonimia, meronimia y sinonimia. Conforme a la propuesta de Buitelaar, Cimiano y Magnani (2005), un CD contiene tanto unidades conceptuales (términos y definiciones)

como relaciones léxicas y semánticas, de acuerdo con este esquema:

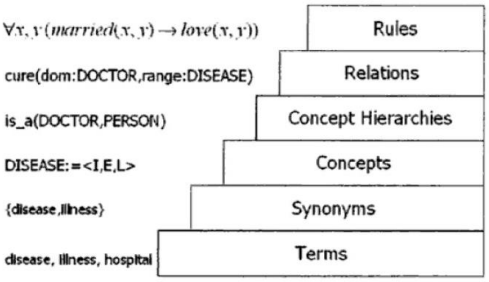


Figura 2: Esquema de desarrollo de una ontología según Buitelaar, Cimiano y Magnini

Este esquema describe un proceso escalonado para construir una ontología con información textual. En el primer escalón se concibe un proceso de extracción de términos (en nuestro caso, éstos se reconocen en CDs previamente delimitados). Después se hace una búsqueda de sinónimos ligados a los términos extraídos, ya sea a partir de la consulta de diccionarios, o ya sea empleando *WordNet* (Fellbaum, 1998).

En el siguiente escalón se pasa a formalizar la información conceptual asociada al término definido (en nuestro caso, tal información está contenida en CDs), conforme a una jerarquía léxica basada en relaciones de hiperonimia/hiponimia, meronimia y sinonimia. Luego, en un plano de representación conceptual, se pueden construir marcos semánticos como los que propone *FrameNet* (Baker, Fillmore y Cronin, 2003). Finalmente, el último escalón representa la formulación de axiomas, tomando en cuenta la información léxica y semántica obtenida.

4 Avances

Lo que mostramos aquí es el desarrollo de un método de extracción de términos, así como un método para detectar relaciones de hiponimia/hiperonimia y meronimia.

4.1 Implementación de un método de extracción de términos

Hemos implementado un método de extracción de términos, el cual aplica un contraste entre corpus, usando 4 medidas para asignar relevancia a palabras que ocurren tanto en el corpus de dominio como en un corpus de lengua general. Tales medidas son: razón *log-likelihood*; (ii) diferencia de rangos aplicada por Kit y Liu (2008); (iii) razón de frecuencia relativa, y la aproximación a la distribución binomial mediante el uso de la distribución normal estándar, planteada por Drouin (2003).

Los resultados obtenidos muestran un desempeño mejor de las medidas diferencia de rangos y razón de frecuencias relativas. Igualmente, nuestro método es útil para asignar relevancia a palabras de un dominio, ya que el vocabulario estrechamente relacionado con un dominio tendrá mayor probabilidad de ocurrencia en éste, que en un corpus general. Para más detalles véase a Acosta, Aguilar e Infante (2015).

4.2 Identificación de relaciones de hiponimia/hiperonimia y meronimia

En el caso de relaciones de hiponimia/hiperonimia, se ha desarrollado un método híbrido que utiliza un *chunker* capaz de reconocer sintagmas nominales cuyo modificador es un adjetivo relacional (p.e.: *cáncer* ***pancreático***, *infección* ***sanguínea***, *médula* ***ósea***, etc.). Para obtener esta clase de candidatos, se filtran todos aquellos sintagmas que no contengan esta clase de adjetivos, considerando una heurística reportada en Acosta, Sierra y Aguilar (2015).

Estos experimentos de extracción han sido útiles para desarrollar la arquitectura de un sistema prototipo, el cual se muestra en la figura 3. De acuerdo con tal esquema, el *input* es un corpus especializado, el cual se etiqueta con anotado morfosintáctico (o POST), para luego hacer una identificación de segmentos que contengan candidatos a CDs. Una vez localizados estos candidatos, se pasa a identificar aquellos sintagmas nominales que posean o bien un adjetivo relacional, o bien un sintagma prepositivo introducido por *de*.

Tras localizar estos candidatos, se contrastan sus frecuencias de aparición tanto en el corpus de dominio como en un corpus general, con miras a determinar su grado de unicidad (ing.: *unithood*), junto con su valor terminológico (ing.: *termhood*). Al final, el output esperado se compondrá de: (i) un conjunto de candidatos a CDs, un conjunto de candidatos a términos, y (iii), un conjunto de posibles hipónimos e hiperónimos (en el caso que se busque esta relación), o en su defecto candidatos a merónimos.

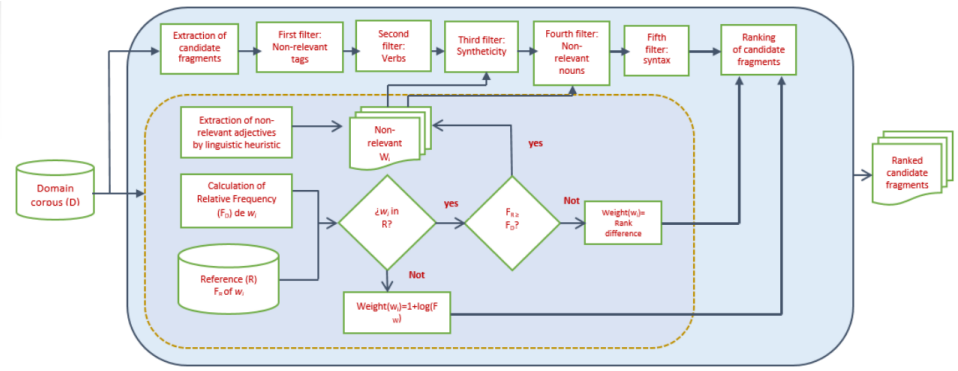


Figura 3: Arquitectura para un prototipo de extractor de CDs

5 *Trabajo a desarrollar*

Los puntos por cubrir son:

- Concretar un sistema definitivo para la extracción de CDs, considerando una clasificación conforme a los tipos de definiciones descritos.
- Continuar con los procesos de evaluación, incluyendo una revisión manual para verificar la calidad de los resultados obtenidos.
- Generar una taxonomía jerarquizada, basada en las relaciones detectadas, la cual integraría un método para generar sinónimos a través de un proceso de permutación, p. e.: dar como sinónimo de *infección ocular* (nombre + adjetivo relacional) *infección del ojo* (nombre + sintagma prepositivo precedido por *de*).

6 *Colaboraciones*

Este proyecto cuenta con la colaboración de cuatro instancias:

(i) El Grupo de Procesamiento del Lenguaje Natural, de la Facultad de Letras de la Pontificia Universidad Católica de Chile.
(ii) El equipo de lingüística computacional de *Cognitiva Latinoamérica* (http://cognitiva.la).
(iii) El Grupo de Ingeniería Lingüística, del Instituto de Ingeniería de la UNAM, México.
(iv) La Facultad de Estadística e Informática de la Universidad Veracruzana, México.

Agradecimientos

Este proyecto ha sido patrocinado por la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), del Gobierno de Chile. Número de proyecto: 11130565.

Bibliografía

Acosta, O., Aguilar, C., e Infante, T. 2015. Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus, *Linguamática*, 7(2): 19-34.

Acosta, O., Sierra, G., y Aguilar, C. 2015. Extracting definitional contexts in Spanish through the identification of hyponymy-hyperonymy relations. En Žižka, J., y Dařena, F. (eds.) *Modern Computational Models of Semantic Discovery in Natural Language*. IGI Global, Hershey, Penn, USA: 48-70.

Buitelaar. P., Cimiano, P. y Magnini, B. 2005. *Ontology learning from text*. IOS Press, Amsterdam.

Baker, C., Fillmore, Ch., y Cronin, B. 2003. The structure of the FrameNet database. *International Journal of Lexicography* 16(3): 281-296.

Drouin, P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1): 99–115.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Kit, Ch., y Liu, Y. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2): 204–229.

Sierra, G. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, 1(2): 13-37.

Sierra, G., Alarcón, R., Aguilar, C., y Bach, C. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* 14(1):74-98.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 173-176 recibido 27-03-2016 revisado 11-04-2016 aceptado 25-04-2016

Sistema de predicción de peticiones de trabajos y servicios en sectores profesionales

Prediction system for job and service requests in professional sectors

Christian Moreno Bermúdez
Universidad de Jaén
Campus Las Lagunillas, s/n, 23071 Jaén
christianmorenobermudez@gmail.com

Arturo Montejo Ráez
Universidad de Jaén
Campus Las Lagunillas, s/n, 23071 Jaén
amontejo@ujaen.es

Resumen: El presente trabajo presenta un sistema que predice peticiones de trabajos y servicios en formato de texto en categorías o sectores profesionales. Se realiza una comparativa de distintos algoritmos de Categorización Automática de Textos para evaluarlos y construir el sistema. El sistema forma parte de una aplicación web que intermedia entre particulares que demandan presupuestos sobre trabajos y profesionales que buscan clientes y ofertan servicios.
Palabras clave: Categorización Automática de Textos, Minería de Textos, Minería de Datos, profesionales, presupuestos.

Abstract: System that predicts job requests and services in text format into categories or sectors. A comparison of different algorithms for Automatic Text Categorization is performed in order to build the final system. The system is part of a web application that mediates between individuals who demand estimates about jobs and professionals who seek clients and offer services.
Keywords: Automatic Text Categorization, Text Mining, Data Mining, professionals, budgets.

1 Introducción

En la actualidad, existe la necesidad de encontrar a profesionales que satisfagan las peticiones de personas en cuanto a trabajos y servicios se refiere, por ejemplo, para realizar una reforma en el hogar. Cuando una persona necesita de los servicios de un profesional debe consultar directorios de empresas, buscadores, preguntar a otras personas, etc. Si desea comparar presupuestos de varios profesionales, el proceso anterior se repite para cada profesional. Por otra parte, profesionales y empresas necesitan ofertar sus servicios a clientes para el desarrollo de su actividad.

El sistema de predicción, que se explicará a continuación, forma parte de una aplicación web desarrollada como prueba de concepto. Esta aplicación permite al particular registrar su necesidad mediante una descripción en formato de texto. Un ejemplo sería: *"Quisiera presupuesto para instalar rejas en 4 ventanas, las dimensiones son de 1,5x1,5m"*. Una vez enviada la petición, el sistema de predicción la clasifica en el sector profesional correspondiente, que en el ejemplo anterior coincidiría con "Carpintería Metálica". Posteriormente, la aplicación se encarga de avisar a los profesionales correspondientes dados de alta en dicho sector para que manden sus presupuestos. Este sistema constituye, pues, una demostración práctica del uso de la categorización automática de textos.

Este proyecto está parcialmente financiado por el Gobierno de España a través del proyecto REDES (TIN2015-65136-C2-1-R).

Para realizar el sistema se llevan a cabo las siguientes etapas: recopilación y preparación de datos que se explica en la sección 3, selección de algoritmos a comparar en la sección 4 y evaluación de resultados en la sección 5.

2 Antecedentes

La Categorización y clasificación Automática de Textos (Sebastiani, 2002) es una de las tecnologías del lenguaje humano que mayor aplicación ha demostrado (Jackson y Moulinier, 2007). Entendemos la categorización automática de textos como la asignación de un texto a una o varias categorías (mono-etiqueta o multi-etiqueta). La

gran mayoría de los algoritmos orientados a la clasificación, entre los que destacan las *Support Vector Machines* (SVM) (Joachims, 1998), suelen ser de tipo binario, esto es, el algoritmo clasifica el texto a una de dos posible categorías. Sobre estos clasificadores binarios se construyen los multiclase (una de varias categorías) y los multietiqueta (una o más de varias categorías posibles). En cualquier caso, actualmente se ha encontrado niveles de precisión altos que posibilitan el uso de estos algoritmos con garantías en aplicaciones reales, como el filtrado de información en la web (Gentili et al., 2001), procesamiento de textos legales (Cardie et al., 2008) o la clasificación de artículos para la Física de Altas Energías (Montejo-Ráez et al., 2005).

3 Recopilación y preparación de datos

Es necesario un conjunto de peticiones de trabajos clasificados para entrenar el sistema de predicción. En Internet encontramos diferentes sitios web que ayudan a resolver la problemática planteada en la introducción y muestran un amplio conjunto de ejemplos de peticiones reales ya categorizadas. Nosotros usaremos la herramienta *wget*[1] de UNIX para obtener un total de 43.601 documentos HTML, donde un documento contiene una petición de trabajo clasificada en una de 24 categorías posibles.

Se eliminan aquellas categorías que son una agregación de varias categorías, por ejemplo, "Reformas" que incluye ejemplos donde necesitan fontaneros, carpinteros, albañiles, etc. simultáneamente. A pesar de ser ejemplos válidos, pertenecen a un problema de clasificación multi-etiqueta en el que no se centra este proyecto.

Tras el sesgo anterior, el conjunto de dato queda con 31.056 ejemplos y 16 categorías. Se genera una muestra aleatoria de 10.000 ejemplos para realizar las pruebas.

A continuación, se aplican los procesos que construyen el índice de un buscador sobre ficheros de texto:

1. Eliminación de caracteres indeseables.
2. Eliminación de palabras que carecen de significado o *stopwords*.
3. Extracción de raíces, con la ayuda de un *stemmer*[2] para el español.
4. Aplicación del Modelo Espacio Vectorial, para generar la matriz de pesos que dará soporte al clasificador. Se obtiene una matriz con 7.652 atributos: *calder, repar, papel, termostat*, etc.

Para mejorar el tiempo de entrenamiento, se reduce el tamaño de la matriz eliminando aquellos términos que aparecen menos de 10 veces en todo el conjunto de datos. El conjunto de atributos desciende a 1.283.

Por último, se plasman los datos en formato *.arff*. Este es el formato usado por la herramienta Weka[3], que contiene los diferentes algoritmos de minería de datos que usaremos para la comparativa.

Antes de lanzar los experimentos, conviene reducir aún más la dimensión del conjunto de datos aplicando un método de selección de atributos, como el filtro basado en consistencia (Liu y Setiono, 1996), que mejora en gran medida el tiempo de entrenamiento e incluso la tasa de acierto entre 1 y 3 puntos porcentuales.

Finalmente, el conjunto resultante contiene:

- Nº de instancias: 9.984
- Nº de atributos: 280
- Nº de categorías: 16

4 Selección de algoritmos

Para llevar a cabo el experimento, se ha creído conveniente elegir cinco familias de algoritmos diferentes en las que se incluyen los siguientes algoritmos y parámetros:

- Algoritmos basados en probabilidad:
  - **BayesNe**t (O. Pourret et al., 2008), redes bayesianas con los parámetros por defecto definidos en Weka.
  - **Naïve Bayes** (Rish, 2001), con los parámetros por defecto.
- Máquinas de Vectores de Soporte:
  - **LibSVM** (Chih-Chung y Chih-Jen, 2013), combinando:
    - kernel: linea, polinómico, de base radial y sigmoidal.
    - C (factor de complejidad): 0,125, 1 y 2.
  - **SMO** (Platt, 1998), combinando:
    - kernel: polinómico, gaussiano.
    - C (factor de complejidad): 0,125, 1 y 2.

[1] https://www.gnu.org/software/wget/

[2] http://snowball.tartarus.org/

[3] http://www.cs.waikato.ac.nz/ml/weka/

    - Exponente (solo para kernel polinómico): 0,125, 1 y 2.
    - Gamma (solo para kernel gaussiano): 0,1, 0,01 y 0,001.
- Algoritmos basados en proximidad:
    - **KNN** (Altman, 1992), con los siguientes número de vecinos (k): 1, 3, 5 y 7.
- Algoritmos basados en reglas:
    - **RIPPER** (William, 1995) o JRip, con los parámetros por defecto.
- Algoritmos basados en árboles:
    - **C4.5** (Quinlan, 1993) o J48, con los parámetros por defecto.
    - **RandomForest (**Breiman, 2001), con número máximo de árboles: 20, 50, 80 y 110.

Esta combinación de algoritmos y parámetros da como resultado un total de 42 clasificadores entrenados en Weka.

Aquellos que mejor se comporten, constituirán las bases para construir 3 meta-clasificadores diferentes, con los que se espera una mejora en los resultados:

- **StackingC** (Seewald, 2002): Se seleccionarán los cuatro mejores algoritmos anteriores para construir un clasificador por regresión.
- **Bagging** (Breiman, 1996): Bagging escogiendo porciones del dataset del 33,33% con el mejor algoritmo base.
- **AdaBoostM1** (Freund y Schapire, 1996): Boosting con el mejor algoritmo base.

5 Evaluación de resultados

A continuación, se expone la Tabla 1 que incluye la comparativa de la tasa de acierto en predicción de los cuatro mejores algoritmos base:

ALGORIT	PARAM	ACIERTO
BayesNet	Por def.	81,32 %
LibSVM	C = 2, Kernel = Lineal	83,95 %
SMO	C = 2 , exp = 1, Kernel = Polin.	83,89 %
Random Forest	Tam = 80	81,53 %

Tabla 1: Comparativa de los mejores algoritmos base

Dado los resultados de la tabla anterior, se escoge el algoritmo LibSVM con factor de complejidad igual a 2 y kernel lineal para construir un meta-clasificador por Bagging y por Boosting. Para construir el meta-clasificador por Stacking se elige a los cuatro algoritmos de la tabla.

En este caso, se tiene en cuenta, además de la tasa de acierto, las métricas de Error Medio Absoluto y coeficiente de Kappa (Cohen, 1960). El Error Medio Absoluto verifica con qué grado de exactitud el sistema acierta, es decir, cómo de lejos están las predicciones de los datos reales. El coeficiente Kappa es una medida que muestra como de bueno es un clasificador estudiando la concordancia de los resultados obtenidos por varios clasificadores del mismo tipo. Valores cercano a 1 afirman una concordancia buena, mientras que valores cercanos a 0 muestran una concordancia debida exclusivamente al azar.

Tras el lanzamiento de las pruebas, se obtienen los siguientes resultados (véase Tabla 2):

ALGORIT	ACIERTO	E.M.A	KAPPA
StackingC	84,93 %	0,0307	0,8326
Bagging	82,90 %	0,0235	0,8094
AdaBoostM1	83,21 %	0,0470	0,8135

Tabla 2: Comparativa de meta-clasificadores

De los resultados que se muestran en la Tabla 2, se desprende que el algoritmo más prometedor para implementar el sistema de predicción es **StackingC**, ya que presenta la mayor tasa de acierto, una media del error absoluto en un intervalo razonable y el mejor estadístico de Kappa.

6 Implementación en la aplicación web

Para implementar el algoritmo en la aplicación web, se ha utilizado la API de Weka en lenguaje JAVA. En un proyecto a parte, se puede cargar el dataset de instancias y atributos con el que entrenar un objeto de la clase weka.classifiers.meta.StackingC. Posteriormente, el objeto se serializa y queda listo para recuperarlo y usarlo desde la aplicación web.

La secuencia de pasos a realizar en la aplicación es bastante simple. El usuario introduce la descripción de aquello que necesita en un formulario, junto a sus datos de

contacto. Seguidamente, el sistema de predicción clasifica la petición en el sector profesional que corresponde. Acto seguido, se notifica a los profesionales de dicho sector y próximos al usuario de que tienen una solicitud de presupuesto interesante. Los profesionales pueden contactar con el cliente y mandar sus presupuestos para que el usuario compare fácilmente y decida. Dado que el sistema no es infalible, el usuario puede modificar la categorización realizada manualmente.

7 Conclusiones y trabajo futuro

Hemos construido un sistema funcional para el enrutado de peticiones de trabajo a profesionales gracias a la aplicación de algoritmos de categorización de textos. Hemos evaluado su precisión con resultados que llevan a concluir la factibilidad de estas tecnologías en este dominio de cara a su explotación real. Cabe destacar que el sistema se trata de un prototipo en un contexto limitado. Es necesario un conjunto de ejemplos más rico tanto en descripciones como en categorías empresariales.

Por otra parte, sería muy interesante desarrollar un clasificador multi-etiqueta, ya que en muchos casos encontraremos descripciones que hacen referencia a varios sectores profesionales. Por ejemplo, en una descripción como *"quisiera mudarme a un piso nuevo y se necesitan trasladar todos los muebles. Además es necesario pintar el piso"* se debería avisar a tanto a profesionales de la mudanza como a pintores.

Bibliografía

Altman, N. S. 1992. *An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46 (3)*: 175–185.

Breiman, Leo. (1996). *Bagging predictors. Machine Learning.* 24(2):123-140.

Breiman, Leo. (2001). *Random Forests. Machine Learning* 45 (1): 5–32.

Cardie, C., Farina, C. R., Rawding, M., & Aijaz, A. (2008). *An eRulemaking Corpus: Identifying Substantive Issues in Public Comments.*

Chih-Chung Chang and Chih-Jen Lin (2013). *LIBSVM: A Library for Support Vector Machines.* National Taiwan University.

Cohen J. *A coefficient of agreement for nominal scales*. Educ Psychol Meas 1960; 20: 37-46.

Freund Yoav and Schapire Robert E. (1996): *Experiments with a new boosting algorithm.* 148-156.

Gentili, G. L., Marinilli, M., Micarelli, A., & Sciarrone, F. (2001). *Text categorization in an intelligent agent for filtering information on the Web.* 15(03), 527-549.

H. Liu, R. Setiono (1996): A *probabilistic approach to feature selection - A filter solution*. 319-327.

Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization* .

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137-142). Springer Berlin Heidelberg.

Montejo-Ráez, A., Urena-Lopez, L., & Steinberger, R. (2005). *Text categorization using bibliographic records: beyond document content.* 119-126.

O. Pourret, P. Naim and B. Marcot. 2008. Bayesian Networks: *A Practical Guide to Applications*. Chichester, UK: Wiley.

Platt, John. 1998, *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.*

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.*

Rish, Irina. 2001. *An empirical study of the naive Bayes classifier*.

Sebastiani, F. (2002). *Machine learning in automated text categorization. ACM computing surveys (CSUR), 34*(1), 1-47.

Seewald, A. K. (2002). *How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness*. 554-561.

William W. Cohen. 1995. *Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning*, 115-123.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 177-180 recibido 28-03-2016 revisado 11-04-2016 aceptado 20-04-2016

EasyLecto: Un sistema de simplificación léxica de efectos adversos presentes en prospectos de fármacos en español

EasyLecto: A lexical simplification system for adverse drug effects in Spanish patient information leaflets

Luis Núñez-Gómez, Isabel Segura-Bedmar, Paloma Martínez
Universidad Carlos III de Madrid
Av. Universidad, 30, 28913
lununezg@pa.uc3m.es,{*isegura*, *pmf*} @inf.uc3m.es

Resumen: Presentamos EasyLecto, un sistema de simplificación léxica de efectos adversos presentes en prospectos de fármacos en español. El método de simplificación léxica que utiliza EasyLecto se basa en la frecuencia de las palabras para determinar los sinónimos más simples. Este sistema propone los mejores sinónimos y definiciones, obtenidos a partir de los recursos MedlinePlus y MedDRA. El sistema puede ayudar a los lectores con bajo nivel de alfabetización, con problemas cognitivos o discapacidad a la hora de entender los efectos adversos presentes en prospectos de fármacos.
Palabras clave: Simplificación léxica, Procesamiento de Lenguaje Natural

Abstract: We introduce EasyLecto, a lexical simplification system of adverse effects in patient information leaflet in Spanish. The method of lexical simplification using EasyLecto is based on the frequency of words to determine the simplest synonyms. This system proposes the best synonyms and meanings, obtained from the MedlinePlus and MedDRA resources, representing their benefit for readers with low literacy, with cognitive problems or handicapped in understanding the adverse drug effects in patient information leaflet in Spanish.
Keywords: Lexical simplification, Natural Language Processing

1 Introdución

La simplificación automática de textos tiene como objetivo transformar un texto complejo en uno sencillo de leer y de entender. Es una tarea en Procesamiento de Lenguaje Natural (PLN) que en los últimos años ha crecido considerablemente (Abrahamsson et al., 2014) y que puede beneficiar a distintos grupos de usuarios como las personas mayores, personas con bajo nivel de alfabetización, personas con discapacidad lectora o incluso personas que están aprendiendo un idioma. Para simplificar un texto es aconsejable aplicar las pautas de lectura fácil[1], que pueden resumirse en: uso de lenguaje directo y sencillo, empleo de voz activa, expresar una idea por oración, evitar el uso de jerga y tecnicismos así como abreviaturas, estructurar el texto de forma clara y coherente, y usar palabras que representen un único concepto. En la simplificación automática de textos las oraciones complejas se dividen en oraciones más simples y el vocabulario complejo se reemplaza por un vocabulario más fácil de entender. La aplicación de técnicas de simplificación de textos puede ser beneficiosa para disminuir la complejidad de los textos en dominios tan diversos como la administración electrónica, la enseñanza o la salud, entre muchos otros. En este artículo nos centramos en el dominio de la salud y en particular en la simplificación de prospectos de medicamentos. Los prospectos de fármacos no son fáciles de comprender ya que contienen un gran número de términos técnicos del ámbito médico y porque sus oraciones suelen contener un gran número de estructuras gramaticales complejas (Davis et al., 2006). Por lo general, las personas desconocen la terminología médica y cuando tratan de entender los efectos adversos de un fármaco no logran hacerlo y terminan quitando importancia. Por esta razón es importante reducir la complejidad de los prospectos y de esta manera contribuir a evitar un uso inadecuado y peligroso de los medicamentos.

[1] www.lecturafacil.es/es

ISSN 1135-5948
© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

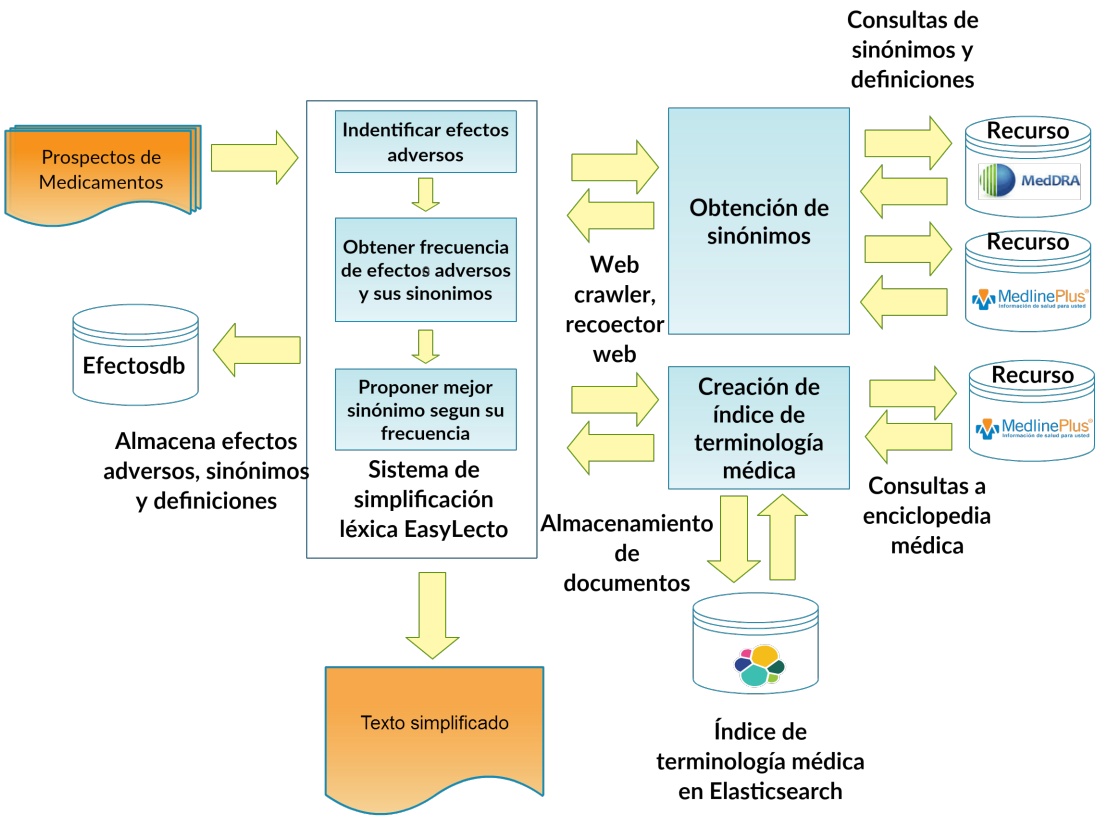


Figura 1: Esquema del proceso de simplificación léxica de EasyLecto.

Aunque existen herramientas de PLN basadas en técnicas de simplificación léxica que permiten reducir la complejidad de documentos, muy pocas están orientadas al español y al dominio de la salud (Bott, Saggion, y Mille, 2012; Grigonytea et al., 2014).

En las siguientes secciones se presenta la herramienta EasyLecto, un sistema de simplificación léxica de los efectos adversos presentes en los prospectos de fármacos en español. La herramienta combina recursos terminológicos para obtener los sinónimos de los efectos adversos y el cálculo de sus frecuencias en una colección de textos recopilada de la enciclopedia online MedLinePlus[2]. Nuestra principal hipótesis es que el sinónimo más frecuente debería ser el más simple. En el siguiente apartado, se describe con más detalle el método de simplificación propuesto.

2 *Proceso de simplificación Léxica de EasyLecto*

El sistema EasyLecto propone los mejores sinónimos para cada efecto adverso presente en un prospecto de medicamentos. El método de simplificación léxica que utiliza EasyLecto es la frecuencia de términos para determinar el sinónimo más simple; siendo el más frecuente. Figura 1 muestra la arquitectura del sistema EasyLecto.

El índice de terminología médica se crea utilizando la herramienta Elasticsearch [3], un gestor de base de datos distribuido y orientado a documentos, el cual permite almacenar gran cantidad de información, facilita las consultas mediante JSON [4] y está basado en tecnología Apache Lucene [5]. El proceso de creación del índice empieza con un algoritmo web-crawler que utiliza jsoup[6], una biblioteca en java que permite extraer y manipular datos de la web. El algoritmo recupera los artículos de la enciclopedia de MedlinePlus de forma alfabética; la enciclopedia incluye una gran cantidad de artículos acerca de temas de salud como enfermedades, exámenes médicos, síntomas, lesiones y procedimientos quirúrgicos. Una vez descargado, cada artículo de MedLinePlus se representa como un objeto

[2]https://www.nlm.nih.gov/medlineplus/spanish/

[3]https://www.elastic.co

[4]http://www.json.org

[5]https://lucene.apache.org/core/

[6]http://jsoup.org/

JSON, donde además de almacenar el contenido de su página web, también se guarda un conjunto de sinónimos del concepto descrito en el artículo (dichos sinónimos están identificados con los metadatos "Otros nombres" y "Nombres alternativos" en la página web del artículo) y la primera oración del texto del artículo, que será considerada como la definición del concepto descrito en el artículo. De esta forma, además de obtener un índice de los términos y sus frecuencias en MedLinePlus, también se ha construido de forma automática un posible diccionario de términos médicos y sus sinónimos. En total se almacenaron un total de 6900 documentos en el índice de terminología médica (ver Figura 1).

EasyLecto utiliza un sistema de reconocimiento de entidades, basado en diccionarios, para identificar las menciones de efectos adversos descritos en un prospecto. El lector puede encontrar información más detallada sobre dicho sistema en (Segura-Bedmar et al., 2015).

Para cada uno de los efectos adversos detectados, se obtiene un conjunto de sinónimos y definiciones a partir de los recursos MedDRA y MedlinePlus. En el caso de MedLinePlus, en concreto se utiliza la información almacenada en la base de datos de ElasticSearch, descrita en el apartado anterior. MedDRA [7], un tesauro multilingüe médico con información sobre productos médicos y con información sobre los efectos de los medicamentos. Desde el punto de vista de EasyLecto, la principal ventaja de MedDRA es que los efectos están agrupados en conjuntos de sinónimos.

La técnica de simplificación léxica que utiliza EasyLecto, como ya se ha mencionado, es la frecuencia de los sinónimos en el índice de terminología médica creado a partir de la colección MedLinePlus. Nuestra hipótesis es que el sinónimo más frecuente será el más sencillo para sustituir cada efecto adverso.

Después de recuperar los sinónimos de los recursos, en el subproceso "obtener frecuencia de efectos adversos y sus sinónimos", se consulta el índice de terminología médica; para recuperar la frecuencia de los efectos adversos y de sus sinónimos candidatos. El algoritmo realiza consultas mediante JSON a todos los documentos de nuestro índice. En el subproceso "proponer mejor sinónimo según su frecuencia", y una vez obtenidas las frecuencias de los efectos adversos y sus sinónimos candidatos, se propone el sinónimo más simple. En caso que el efecto adverso es más frecuente que todos los sinónimos candidatos, el efecto adverso no será reemplazado por otro sinónimo. Para mejorar los tiempos de ejecución, una vez procesado un documento, EasyLecto almacena sus efectos adversos y sus mejores sinónimos en la base de datos "Efectosdb", como se puede ver en la Figura 1. De esta forma, cada vez que se simplifique, EasyLecto en primer lugar buscará los efectos y sus mejores sinónimos en la base de datos, y sólo en el caso de no existir, buscará sus sinónimos y sus frecuencias en los recursos MedDRA y en el índice de terminología médica.

3 Funcionamiento del sistema EasyLecto

El sistema de simplificación léxica EasyLecto, permite identificar y simplificar los efectos adversos en un prospecto. A modo de ejemplo, al abrir el prospecto "Ceftriaxona", si el usuario necesita saber cuál es el mejor sinónimo para el efecto adverso "tromboflebitis", basta que el usuario haga clic sobre dicho efecto adverso y de manera inmediata se abrirá una ventana emergente. La Figura 2 muestra la salida que proporciona la herramienta EasyLecto para el término tromboflebitis. El mejor sinónimo que propone EasyLecto en el recurso MedlinePlus para el efecto adverso "tromboflebitis" es *"trombosis venosa profunda"*, sin embargo, MedDRA propone como mejor sinónimo el mismo término. La definición que ofrece MedlinePlus para este sinónimo es *"la trombosis venosa profunda o tvp, es un coágulo sanguíneo que se forma en una vena profunda en el cuerpo"*, como se puede ver en la Figura 2. Como MedDRA sólo propone sinónimos, nuestro sistema EasyLecto toma como posible definición el término MedDRA de mayor longitud.

4 Conclusiones y trabajo futuro

En este trabajo se presentó EasyLecto, el primer sistema de simplificación léxica de efectos adversos en prospectos de fármacos en español. Una demo del sistema está disponible en `http://163.117.129.251:8080/EasyLecto`. En la actualidad, se está creando un corpus gold-standard anotado con efectos y sus sinónimos, que nos permitirá dar

[7] http://www.meddra.org/

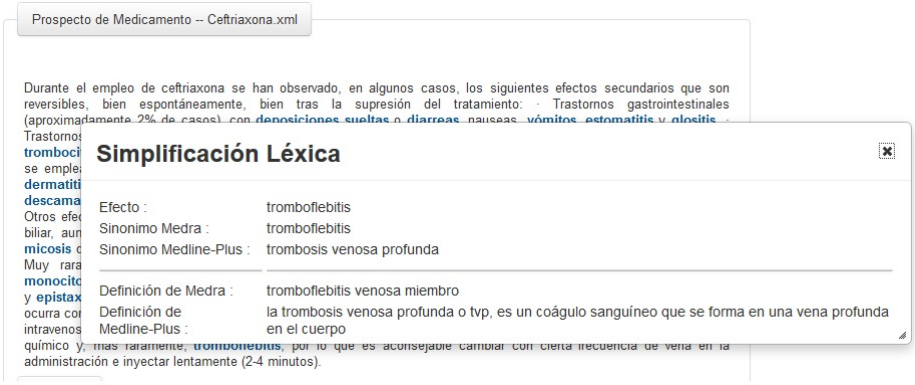


Figura 2: Simplificación léxica de los efectos adversos.

una evaluación cuantitativa de los resultados de la herramienta EasyLecto. En un futuro se pretende integrar otros recursos como BabelNet[8] y abordar la simplificación de otros conceptos médicos (pruebas médicas, tratamientos, enfermedades, etc). También se explorarán otras técnicas para la selección del mejor sinónimo, como por ejemplo el uso de modelos de vectores de palabras.

Agradecimientos

Este trabajo ha sido financiado por el proyecto eGovernAbility-Access (TIN2014-52665-C2-2-R).

Bibliografía

Abrahamsson, E., T. Forni, M. Skeppstedt, y M. Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, páginas 57–65.

Bott, S., H. Saggion, y S. Mille. 2012. Text simplification tools for spanish. En N. C. C. Chair) K. Choukri T. Declerck M. U. Doğan B. Maegaard J. Mariani A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may.

Davis, T. C., M. S. Wolf, P. F. Bass, J. A. Thompson, H. H. Tilson, M. Neuberger, y R. M. Parker. 2006. Literacy and misunderstanding prescription drug labels. *Annals of Internal Medicine*, 145(12):887–894.

Grigonytea, G., M. Kvistbc, S. Velupillaib, y M. Wiréna. 2014. Improving readability of swedish electronic health records through lexical simplification: First results. *EACL 2014*, páginas 74–83.

Segura-Bedmar, I., P. Martínez, R. Revert, y J. Moreno-Schneider. 2015. Exploring spanish health social media for detecting drug effects. *BMC medical informatics and decision making*, 15 Suppl 2(Suppl 2):S6.

[8]http://babelnet.org/

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 181-184 recibido 29-03-2016 revisado 11-04-2016 aceptado 22-04-2016

Through the Eyes of VERTa

A Través de los Ojos de VERTa

Elisabet Comelles
Universitat de Barcelona (UB)
Gran Via de les Corts Catalanes, 585, 08007,
Barcelona (Spain)
elicomelles@ub.edu

Jordi Atserias
IXA Group
University of the Basque Country
(UPV/EHU)
Spain
jordi_atserias001@ehu.eus

Abstract: This paper describes a practical demo of VERTa for Spanish. VERTa is an MT evaluation metric that combines linguistic features at different levels. VERTa has been developed for English and Spanish but can be easily adapted to other languages. VERTa can be used to evaluate adequacy, fluency and ranking of sentences. In this paper, VERTa's modules are described briefly, as well as its graphical interface which provides information on VERTa's performance and possible MT errors.
Keywords: Machine translation evaluation, Automatic metric, Demo, Spanish, Linguistic knowledge, Error analysis

Resumen: Este artículo describe la demostración práctica de VERTa para el castellano. VERTa es una métrica de evaluación de traducción automática que combina información lingüística a diferentes niveles. VERTa ha sido desarrollada para el inglés y el castellano pero se puede adaptar fácilmente a otras lenguas. La métrica puede evaluar la adecuación, la fluidez y ranking de frases. En este artículo se describen brevemente los módulos de VERTa y su interficie gráfica, la cual proporciona información sobre el rendimiento de la métrica y posibles errores de traducción.
Palabras clave: Evaluación de la traducción automática, Métrica automática, Demo, Castellano, Conocimiento lingüístico, Errores de traducción

1 Introduction

Automatic Machine Translation (MT) Evaluation has become a key field in Natural Language Processing due to the amount of texts that are translated over the world and the need for a quick, reliable and inexpensive way to evaluate the quality of the output text. Therefore, a large number of metrics have been developed, which range from very simple metrics, such as BLEU to more complex ones, which involve combining a wide variety of linguistic features using machine-learning techniques (Gautam and Bhattacharyya, 2014; Joty et al., 2014; Yu et al., 2015) or in a more simple and straightforward way (Giménez and Márquez, 2010; González et al., 2014). Nevertheless, little research has been carried out in order to explore the suitability of the linguistic features used and how they should be combined, from a linguistic point of view. In order to address this issue, VERTa, a linguistically-motivated metric (Comelles and Atserias, 2015), has been developed. This metric uses a wide variety of linguistic features at different levels and aims at moving away from a biased evaluation, providing a more holistic approach to MT evaluation.

This paper reports a demo of VERTa and aims at exploring the results provided by the metric and its potential use as an error analysis tool. Therefore, we provide a brief description of the different modules in the Spanish version of VERTa and how the results and the information in these modules can be visualized to better understand the metric's performance and to help developers carry out error analysis. VERTa is available at http://grial.ub.edu:8080/VERTaDemo/.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

2 VERTa

VERTa claims to be a linguistically-motivated metric because before its development a thorough analysis was carried out in order to identify which linguistic phenomena an MT evaluation metric should take into account when evaluating MT output by means of reference translations. With the results of this analysis (Comelles, 2015) we decided on the linguistic features that would be more appropriate and on how they should be combined depending on whether Adequacy or Fluency was evaluated. Therefore, VERTa consists of six modules which can work independently or in combination: *Lexical Similarity Module (L)*, *Morphological Similarity Module (M)*, *N-gram Similarity Module (N)*, *Dependency Similarity Module (D)*, *Semantic Similarity Module (S)* and *Language Model (LM) Module*[1].

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc.).

Next, all modules forming VERTa are described.

2.1 Lexical Similarity Module

The Lexical Similarity Module captures similarities between lexical items in the hypothesis and reference sentences. This module does not only use superficial information such as the wordform, but it also takes into account lemmatization, lexical semantics (i.e. synonymy, hypernymy and hyponymy), and partial lemma. In addition, different weights are assigned depending on their importance as regard semantics and/or fluency.

2.2 Morphological Similarity Module

This module uses the information provided by the Lexical Module in combination with Part-of-Speech (PoS) tags[2].

Similar to the Lexical Similarity Module, this module matches items in the hypothesis and reference segments and a set of weights is assigned to each type of match.

This module aims at making up for the broader coverage of the Lexical Module, thus preventing matches such as *invites* and *invite*, which although similar in meaning differ in their morphosyntactic features.

2.3 Dependency Similarity Module

The Dependency Module captures similarities beyond the external structure of a sentence and uses dependency structures to link syntax and semantics. Thus, this module allows for identifying sentences with the same meaning but different syntactic constructions (e.g. active – passive alternations), as well as changes in word order.

This module works at sentence level and follows the approach used by Owczarzack et al. (2007) and He et al. (2010) with some linguistic additions in order to adapt it to our metric combination. Similar to the Morphological Module, the Dependency Similarity metric also relies first on those matches established at lexical level − word-form, synonymy, hypernymy, hyponymy and lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form. Then, by means of flat triples with the form Label (Head, Mod) obtained from the parser[3], four different types of dependency matches are designed (i.e. complete, partial-no-label, partial-no-head, partial-no-mod) and weights are assigned to each type of match.

In addition, VERTa also enables the user to assign different weights to the dependency categories according to the type of evaluation performed.

Finally, a set of language-dependent rules has been implemented in order to a) widen the range of syntactically-different but semantically-equivalent expressions, and b) restrict certain dependency relations (e.g. subject, object).

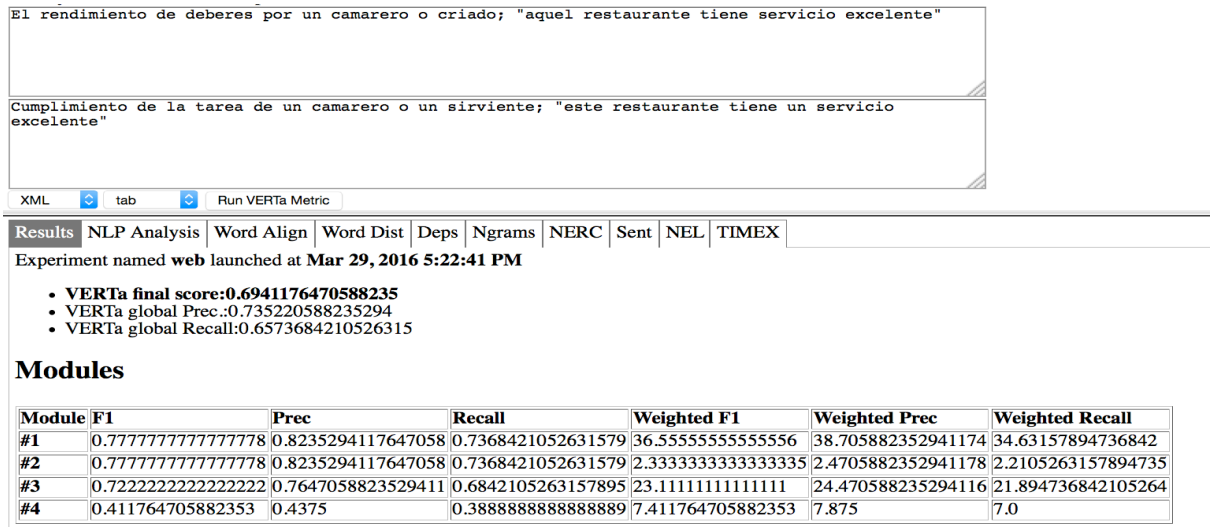
2.4 N-gram Similarity Module

This module matches chunks in the hypothesis and reference segments. N-grams can be calculated over lexical items (considering the information provided by the Lexical Module),

[1] Neither the Semantic Similarity Module nor the Language Model Module are available in the Spanish version of the metric.

[2] The text is tagged using Freeling (Padró and Stanilovsky, 2012).

[3] Both hypothesis and reference strings are annotated with dependency relations by means of Freeling dependency parsing (Lloberes et al., 2010).



Module	F1	Prec	Recall	Weighted F1	Weighted Prec	Weighted Recall
#1	0.7777777777777778	0.8235294117647058	0.7368421052631579	36.55555555555556	38.705882352941174	34.63157894736842
#2	0.7777777777777778	0.8235294117647058	0.7368421052631579	2.3333333333333335	2.4705882352941178	2.2105263157894735
#3	0.7222222222222222	0.7647058823529411	0.6842105263157895	23.11111111111111	24.470588235294116	21.894736842105264
#4	0.411764705882353	0.4375	0.3888888888888889	7.411764705882353	7.875	7.0

Figure 1: VERTa's home page, global score and scores per module

over PoS and over the combination of lexical items and PoS. The n-gram length can go from bigrams to sentence-length grams. This module is particularly useful when evaluating Fluency because it deals with word order.

3 VERTa GUI: a Graphical Interface

VERTa GUI allows users to visualize the similarity between two segments module by module. The reference and hypothesis segments are entered in different text boxes and VERTa GUI does not only return the global score but also the score per module (see Figure 1).

In addition, this visual interface also allows users to navigate the different modules in VERTa, by means of a set of tabs. By clicking on each tab (see Figure 2), users are taken to the corresponding module: Word align and word distance (Lexical and Morphological Modules), Deps (Dependency Module), Ngrams (N-gram Module), and NERC, Sent, NEL and TIMEX, corresponding to the Semantic Module.

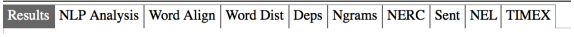


Figure 2: Tabs taking to each module in VERTa

The information contained in these tabs is not only useful in order to check the metric's performance, but also in order to identify possible MT errors.

3.1 A Practical Case

Although VERTa was initially developed to evaluate English a new version has been developed for Spanish. This version uses all modules in VERTa except for the Semantic Similarity Modules and the Language Module Modules. In the demo reported in this paper, VERTa is used to evaluate adequacy. To this aim, the combination of modules is the following:

- Lexical Module: 0.46
- PoS Module: 0.03
- Dependency Module: 0.32
- Ngram Module: 0.19

The metric can account for the semantic similarity between two sentences such as those in example 1 where the hypothesis segment conveys the meaning of the source segment, despite not being very natural.

SOURCE: *the performance of duties by a waiter or servant; "that restaurant has excellent service "*

HYP: *El rendimiento de deberes por un camarero o criado; "aquel restaurante tiene servicio excelente".*

REF: *Cumplimiento de la tarea de un camarero o un sirviente; "este restaurante tiene un servicio excelente"*

As shown in Figure 3, synonymy helps in matching *deberes* ("duties") and *tareas* ("tasks"), as well as *criado* ("manservant") and *sirviente* ("servant") in the Lexical Module. In addition, possible errors can also be identified by the elements not matched (coloured in red), such as the lack of determiners preceding *deberes* and *servicio* in the hypothesis segment.

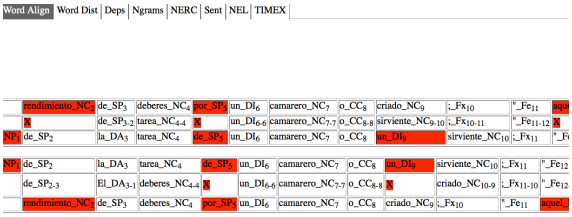


Figure 3: Example of matches in the lexical module

In addition, the n-gram module matches chunks in the hypothesis segments to those in the reference segments. Those chunks that can be matched are highlighted in green.

The demo can be accessed at http://grial.ub.edu:8080/VERTaDemo/.

4 Conclusions and Future Work

This paper has described the Spanish version of VERTa, a linguistically-motivated MT metric, and its graphical interface. The architecture of the metric and the modules in the Spanish version have been described. In addition, its graphical interface, VERTa GUI has been presented in order to show the metric's performance and the usefulness of the information provided by VERTa's modules.

In the future we're planning to add the Semantic Similarity Module and the Language Model Module to the Spanish version. In addition, we are also considering a better way to extract and display information on error analysis.

Acknowledgements

This work has been funded by the Spanish Government (project TUNER, TIN2015-65308-C5-1-R)

References

Comelles, E. and J. Atserias. 2015. VERTa: A Linguistically-Motivated Metric at the WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. EMNLP, pp. 366-372, Lisbon.

Comelles, E. 2015. *Automatic Machine Translation Evaluation: A Qualitative Approach*, University of Barcelona, Barcelona.

Gautam, S. and P. Bhattacharyya. 2014. LAYERED: Metric for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 387-393, Baltimore.

Giménez, J. and Ll. Márquez. 2010. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4):77–86.

González, M., A. Barrón-Cedeño, and Ll. Márquez. 2014. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. En *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 394-401, Baltimore.

He, Y., J. Du, A. Way and J. van Genabith. 2010. The DCU Dependency-based Metric in WMT-Metrics MATR 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics* MATR, WMT 2010, pp. 349-353, Uppsala.

Joty, S., F. Guzmán, Ll. Márquez and P. Nakov. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 402-408, Baltimore.

Lloberes, M., I. Castellón and Ll. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh conference on International Language Resources and Evaluation,* LREC'10, pp. 693-699, Malta.

Owczarzak, K., J. van Genabith and A. Way. 2007. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the ACL Workshop on Statistical Machine Translation,* ACL, pp. 104-111. Prague.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference,* LREC 2012, pp.2473-2479. Istanbul.

Yu, H., Q. Ma, X. Wu and W. Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, EMNLP, pp. 417-421, Lisbon.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 185-188 recibido 29-03-2016 revisado 11-04-2016 aceptado 23-04-2016

Pictogrammar, comunicación basada en pictogramas con conocimiento lingüístico*

Pictogrammar, pictograms based communication with grammatical support

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

Resumen: Diversos trastornos, como los trastornos del espectro autista (TEA), afectan a la capacidad de comunicación de las personas desde edades tempranas. Para muchos de estos casos se utilizan métodos de comunicación aumentativa y adaptativa (CAA) con el fin de desarrollar o recuperar la capacidad de comunicación. Pictogrammar es un sistema CAA completo que hace uso de una ontología propia, denominada PictOntology, con el fin de mejorar estos problemas de comunicación.
Palabras clave: TEA, comunicación, gramática semántica

Abstract: Several disorders, such as autism spectrum disorder (ASD) affect the ability of communication of people from an early age. For many of these cases and adaptive and augmentative communication methods (AAC) are used to develop or regain communication skills. Pictogrammar is a complete AAC system which uses an ontology itself, called PictOntology, in order to improve these communication problems.
Keywords: ASD, communication, semantic grammar

1 Introducción

Autismo, trastornos de espectro autista (TEA) o síndrome de Asperger son trastornos del desarrollo cerebral que suelen aparecer en edades tempranas y agrupan diversos diagnósticos: déficit en la comunicación, dificultades para integrarse socialmente, una exagerada dependencia a las rutinas y hábitos cotidianos o una alta intolerancia a cualquier cambio. Los déficits en la comprensión del lenguaje incluyen la dificultad de comprender direcciones simples, preguntas u órdenes (Lim, 2011). Además, se hace presente la falta de comunicación verbal o si está presente es muy inmadura "*quiero agua*", en lugar de "*quiero un vaso de agua, por favor*".

Entre los sistemas más usuales para paliar, al menos en parte, y mejorar la comunicación, encontramos los Sistemas Aumentativos y Alternativos de Comunicación (en inglés, SAAC). La comunicación y el lenguaje son esenciales para todo ser humano, son necesarios para aprender, para relacionarse con los demás, para disfrutar y para participar en la sociedad.

Se consideran alternativos aquellos sistemas que sustituyen totalmente al habla, mientras que se entiende por aumentativos aquellos sistemas que son un complemento al habla. Los sistemas alternativos se refieren más al lenguaje y los aumentativos al habla.

Entre las causas que pueden hacer necesario el uso de un SAAC encontramos los trastornos del espectro autista (TEA), la parálisis cerebral (PC), la discapacidad intelectual, las enfermedades neurológicas tales como la esclerosis lateral amiotrófica (ELA), la esclerosis múltiple (EM) o el párkinson, las distrofias musculares, los traumatismos cráneo-encefálicos, las afasias o las pluridiscapacidades de tipologías diversas.

Algunos de los SAACs más populares están basados en pictogramas, signos en forma de iconos dibujados que representan figurativamente, de forma más o menos realista, un objeto real o significado, tal como Pictogram Exchange Communication System (PECS)(Andy y Lori, 1994). La Figura 1 muestra ejemplos de algunos de los pictogramas utilizados en PECS.

PECS no es sólo un SAAC sino una meto-

* Este trabajo está parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto REDES (TIN2015-65136-C2-1-R) del gobierno de España.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

dología de trabajo para el aprendizaje comunicativo de cualquier persona con dificultades o trastornos de comunicación. En PECS se puede incluso añadir conceptos más descriptivos del lenguaje, tal como el tamaño, forma, color o número, de forma que los mensajes comunicativos son más específicos al combinar distintos símbolos. La Figura 1 muestra un ejemplo de una petición utilizando pictogramas.

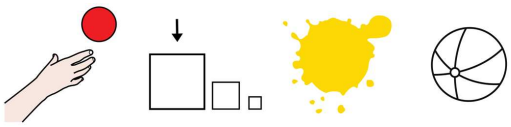


Figura 1: "Quiero la pelota grande amarilla" utilizando pictogramas

Encontramos diversos comunicadores basados en PECS, tal como Speak4Yourself[1], ARaSuite[2], SC@UT[3], CPA[4] o e-Mintza[5], sistemas con pictogramas categorizados por familias que varían entre aplicaciones. El principal problema es que las categorías son un poco aleatorias, basadas en tipos sintácticos (nombres, verbos, adjetivos), o en aspectos temáticos (alimentos, juguetes), y la gran cantidad de pictogramas disponibles hacen que no sean sistemas funcionales para los usuarios.

El principal objetivo de este trabajo es desarrollar Pictogrammar, un sistema AAC completo que trabaja sobre una ontología con una representación formal y un lenguaje controlado. La finalidad de dicha ontología es adecuar el espacio de trabajo de la herramienta al uso terapeútico y paliativo en el tratamiento de trastornos del lenguaje.

En la sección 2 se muestran las ontologías desarrolladas e incluidas en Pictogrammar. La sección 3 describe los detalles de Pictogrammar. En la sección 4 explicamos algunas de las ventajas de la aplicación de tales ontologías en Pictogrammar. Finalizamos este trabajo mostrando conclusiones y trabajo futuro.

[1]http://speakforyourself.org
[2]http://sourceforge.net/projects/arasuite/
[3]http://scaut.ugr.es/scaut/
[4]http://prezi.com/jcpr9qcmcnr-/cpa/
[5]http://fundacionorange.es/emintza.html

2 *Ontologías SUpO y PictOntology*

SUpO (Simple Upper Ontology)(Martínez-Santiago et al., 2015) modela de forma general y básica el conocimiento del mundo a partir de conceptos cotidianos, tal como comidas, juguetes, personas, etc. Existen otras ontologías generales disponibles, tal como SUMO(Pease, 2006), OpenCyC[6] o DOLCE(Masolo et al., 2003), ontologías con conceptos del mundo pero que no tienen detalles semánticos. SUpO se ha diseñado para modelar el nivel semántico del lenguaje, y es más adecuada para el modelado del lenguaje. Su modelo semántico es una adaptación de FrameNET, que incluye una taxonomía de conceptos utilizada en el diseño de SUpO.

Los módulos sintácticos y morfosintácticos han sido implementados con Grammatical Frameworks y con la librería Grammatical Framework Resource Grammar (Ranta, 2011), disponible en más de 20 idiomas.

Definimos PictOntology como una especialización de una ontología de recursos multimedia donde se han establecido propiedades adicionales y una taxonomía de pictogramas. Es así mismo, una versión de SUpO en la cual el vocabulario está formado por pictogramas, concretamento por 621 pictogramas de la colección SymbolStix[7], una colección de aproximadamente 12.000 pictogramas que cubre una gran variedad de categorías, tales como deportes, geografía, personas, salud, tecnología, etc.. La Tabla 1 muestra algunas de las propiedades de PictOntology.

La construcción de PictOntology es, en definitiva, un proceso de integración de una ontología de recursos multimedia. Por este motivo reutiliza varios conceptos y atributos de la ontología Exchangeable Image File Format (EXIF). A partir de ahí, se define un conjunto de metadatos para cada recurso multimedia, el pictograma, junto con el mapeo de dichos elementos a un conjunto de propiedades específicas de cada recurso.

En PictOntology, una categoría está formada por palabras que forman normalmente parte del mismo rol semántico. Pero, además, añadimos alguna restricción más: Una categoría en PictOntology agrupa los pictogramas que de forma nativa evocan distintas ideas sobre el mismo concepto, y que compar-

[6]available at http://sw.opencyc.org/
[7]https://www.n2y.com/products/symbolstix

ten la misma categoria léxica. La finalidad de estas restricciones es doble: Facilitar la integración de PictOntology con SUpO, ya que PictOntology es más restrictiva en su definición y permitir que el usuario pueda ampliar fácilmente la ontología siempre que respete las categorias ya predefinidas. Por ejemplo, es posible añadir un nuevo color, que debe ser necesariamente un adjetivo y que, además, solo podrá ser utilizado en el mismo contexto que el resto de los colores.

3 Pictogrammar

Pictogrammar es un systema AAC basado en los siguientes componentes:

- Un lenguaje natural controlado, que es el objeto que en definitiva debe ser enseñado y aprendido. En otras palabras, se trata del documento de especificación de la ontologia que gobierna Pictogrammar, SUpO.
- Los usuarios son el alumno y los terapeutas, familiares o tutores.
- SUpO, una ontología con conocimiento del mundo, factible al tener un vocabulario principal pequeño.
- PictOntology, una ontología formada por pictogramas y enlazada con SUpO.
- Un sistema SAAC basado en PictOntology entre alumnos y equipo de intervención. Es, en definitiva, una herramienta de autoría para manipular SUpO basada en los pictogramas de PictOntology.

Dado que Pictogrammar está enlazado con una ontología con conocimiento lingüístico, aporta diversos beneficios y aspectos novedosos a la hora de generar mensajes, tales como:

1. Cuando el usuario genera el mensaje: a) Expandible. Es posible incrementar el vocabulario del usuario, sin necesidad de reescribir la gramática. Esto es posible gracias a las restricciones impuestas en la definición de las categorias de PictOntology. b) Gramática semántica predictiva. El SAAC filtrará pictogramas de acuerdo con el contexto de la frase que está construyendo. Por ejemplo, el verbo "comer" solo permite como complementos aquello que es comestible, y a su vez la comida solo puede adjetivarse con propiedades aplicables a la comida, tales como el sabor o el color. c) Las frases son sintáctica y semánticamente correctas, no es posible generar frases al margen del lenguaje controlado definido. d) Adaptativo. El vocabulario y la complejidad sintáctica es completamente dependiente del usuario.

2. Cuando el sistema lee la frase construida: a) Las frases generadas con el SAAC son sintácticamente correctas, suenan naturales. Hay concordancia de género y número, persona y tiempo verbal, etc. Esta propiedad es consecuencia directa del uso de los marcos gramáticales. b) Traducción inmediata de las frases construidas a cualquiera de los idiomas que soportan los marcos gramaticales.

3. Cuando el equipo de intervención define el lenguaje: a) El mismo sistema puede sugerir cuáles son los siguientes conceptos mejor situados para ser aprendidos. Por ejemplo, si el alumno está trabando conceptos relativos a alimentos, el sistema puede sugererir adjetivos específicos de alimentos. b) Una ontología común hace posible compartir conocimiento sobre el modelo del lenguaje que los usuarios, equipo y alumno, son capaces de entender y generar. c) De igual modo, Pictogrammar puede constituirse como el pilar para un futuro ecosistema de aplicaciones terapéuticas y/o paliativas, todas ellas compartiendo un modelo del lenguaje común. Por ejemplo, se podría definir una aplicación orientada a lectoescritura o la generación de lenguaje oral o lenguaje espontáneo, etc. Todas ellas compartirían exactamente el mismo modelo del lenguaje.

4 Conclusiones y trabajo futuro

En este trabajo presentamos Pictogrammar, un SAAC basado en PictOntology, con una finalidad terapéutica y paliativa, que aplica diversas técnicas de las tecnologías del lenguaje humano. Además describe PictOntology, una ontología de pictogramas enlazada con SUpO.

Nombre	Tipo	Descripción
ma:identifier	identifier:URI, type:String	Nombre del fichero del pictograma
ma:title	title:String, type:String	Nombre en inglés del pictograma. Normalmente el equivalente a la expresión en inglés
ma:language	String	Normalmente los pictogramas son independientes del idioma
ma:creator	String	Creador del recurso
ma:contributor	identifier:URI—String, role:String	Identificador de la persona que añade el pictograma
ma:collection	URI—String	Nombre de la colección origen
ma:relation	identifier:URI, relation:String	"is-a" relación entre un pictograma y su categoría
pt:expressions	List of lang:String, expression: String	Traducción textual del pictograma
pt:level	"transparent" ó "learned" ó "abstract"	'transparent": pictogramas con un significado obvio por su ilustración / "learned": pictogramas cuyo significado tiene que aprenderse / "abstracts": pictogramas que no tienen un significado obvio
pt:learned_group	String	Esta etiqueta es compartida por todos los pictogramas relativos al mismo concepto
pt:SUpO_concepts	{identifier:URI, type:List of Strings}	Identificador de los conceptos SUpO relativos al pictograma

Tabla 1: Ejemplos de propiedades de PictOntology

Como trabajo inmediato es necesario obtener resultados del rendimiento del sistema con usuarios reales, en términos de tiempo de aprendizaje de los alumnos, tamaño del vocabulario adquirido, frecuencia y complejidad de frases, etc. Sobre la adquisición del lenguaje oral, pretendemos seguir la línea de trabajo de Kasari et al. (2014), que muestra que los dispositivos que generan comunicación mejoran este aspecto comunicativo, así como la espontaneidad y diversas habilidades comunicativas en los primeros años de colegio, especialmente en niños con TEA.

Bibliografía

Andy, B. y F. Lori. 1994. The picture exchange communication system. *Focus on Autism and Other Developmental Disabilities*, 9(3):1–19.

Kasari, C., A. Kaiser, K. Goods, J. Nietfeld, P. Mathy, R. Landa, S. Murphy, y D. Almirall. 2014. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6):635–646.

Lim, H. A. 2011. *Developmental speech-language training through music for children with autism spectrum disorders: Theory and clinical application.* Jessica Kingsley Publishers.

Martínez-Santiago, F., M. Díaz-Galiano, L. Ureña-López, y R. Mitkov. 2015. A semantic grammar for beginning communicators. *Knowledge-Based Systems*, 86:158–172.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, y L. Schneider. 2003. *The WonderWeb Library of Foundational Ontologies Preliminary Report.*

Pease, A. 2006. Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics. En *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts.* Mouton de Gruyter, New York.

Ranta, A. 2011. *Grammatical framework: Programming with multilingual grammars.* CSLI Publications, Center for the Study of Language and Information.

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 189-192 recibido 28-03-2016 revisado 10-04-2016 aceptado 23-04-2016

Evall: A Framework for Information Systems Evaluation*

Evall: Una Plataforma para la Evaluación de Sistemas de Información

Enrique Amigó, Jorge Carrillo-de-Albornoz, Julio Gonzalo and Felisa Verdejo
Universidad Nacional de Educación a Distancia (UNED)
Calle Juan del Rosal 16, Madrid
{enrique, jcalbornoz, julio, felisa}@lsi.uned.es

Abstract: In this paper, the **Evall** framework for the automatic evaluation of information systems task is presented. With just one click and providing the system outputs of the algorithms, Evall allows researchers to automatically generate a Latex report including the results of their algorithms, statistical significance tests, measures descriptions, and references.
Keywords: Evaluation Framework, Information Systems

Resumen: En este artículo presentamos **Evall**, un framework de evaluación para tareas de investigación en el area de Sistemas de Información. Con un simple click y las salidas de los algoritmos a evaluar, Evall genera un informe automático en Latex con los resultados de todos los sistemas, test de significacia estadística, descripción de las métricas, y referencias.
Palabras clave: Framework de Evaluacion, Sistemas de Información

1 Introduction

In computer science, specially in the area of information systems, a common approach to evaluate the accuracy of proposed methods is by comparing the output generated by the algorithm with a **gold standard** built by human experts. For instance, **Classification** tasks consist of predicting the labels assigned to items by a certain gold standard. **Clustering** tasks aim to group items in the same way than the gold. Finally, the objective of **Ranking** tasks is to order a set of items in correspondence with the gold.

A correct evaluation pursuits, at least, two main objectives: **interpretable**, so it is easy to determine the relative improvements between systems; **standardization and replicability**, so it is possible to replicate the process and to obtain the same results, thus allowing comparison between different systems. Also, the selection of the appropriate evaluation measure determines to a great extent the conclusions of the experimental work. However the evaluation, and specially, the measure selection, is not a trivial issue. First of all, the same problem can be evaluated using several measures, and selecting the best ones for the problem at hand is a challenging work. Second, some measures are complex mathematical formulations that are not easy to understand and to interpret. Third, in many cases there are no standard implementations of the measures, or they are not available, so the measure is implemented several times, which can induce in bugs and errors (specially when comparing results from different implementations). Fourth, the input formats mostly depends of the measure implementation, and usually vary substantially for the same measure. For this reasons, only a small set of measures are commonly used.

The evaluation process plays another very important role in research, since it allows the community to compare their approaches and encourages to overcome them. Up to day, this is a difficult and challenging task, as to evaluate the state of the art of a benchmark (or dataset) used in a workshop or evaluation camping implies several hours searching on Internet for papers published using that benchmark. Also, in many cases the evaluations are performed in different scenarios and under different conditions, so they are not strictly comparable.

Finally, there are other relevant points that highly influence the evaluation process and that are usually complex and become it a tedious process. For example, statistical significance tests are not usually integrated in the measures implementations, or results are not appropriately formatted (an output

* This research was supported by the Spanish Ministry of Science and Innovation (VoxPopuli Project, TIN2013-47090-C3-1-P). Mario Almagro and Javier Rodriguez have contritubed to the development.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

results, for instance, in Latex would strongly benefit the paper preparation).

2 What is Evall?

In this context, Evall aims to help researchers by proposing an easy-to-use **evaluation framework**, **transparent to the user**. Given a set of systems' outputs for a given task (i.e., Classification, Ranking, Clustering, etc.) and a gold standard, Evall produces an informative report in Latex format that includes measure descriptions and explanations, result tables, statistical significance tests, comparisons between systems' outputs, charts, references, etc., as well as a set of CSV files containing a more fine grained description of the results. The five main contributions of Evall are:

(i) Evall allows to evaluate systems outputs by **only indicating the task to be addressed** (i.e., Classification, Ranking, and Clustering). According to the task, Evall selects all available measures, checks their preconditions in the inputs (system outputs and gold standard), and generates the results. The selected measures are described, including a summary of its foundations and properties. Furthermore, indications about its limitations and relevant bibliography are also provided.

(ii) The **replicability of results** and the **comparison** between different systems' outputs is achieved. By using Evall, researches ensure that their evaluations are comparable with others that have used Evall too, and that the evaluations are free of errors.

(iii) Evall produces as output a **Latex report** that allows researchers to easily copy and paste tables, descriptions, or result analysis directly to future papers. Also, Evall produces a set of **CSV reports** containing all data in a more fine grained way, which allows researchers to do more experiments, further analysis, etc.

(iv) Evall is designed to **store benchmarks** (gold standards used in workshops, evaluation campaigns, conferences, or papers) and to evaluate new system outputs with the official measures (among others). This allow a researcher developing a new algorithm or approach to evaluate it just by producing the output in Evall format, avoiding the use of an evaluation library or the implementation of the desired measures. Evall also permits the comparison with all system outputs stored in the repository and addressing the same benchmark, and ensures a strict comparison under the same conditions. Similarly, this allows conference or workshops organizers to forget about evaluation by using Evall, ensuring an appropriate evaluation scenario.

(v) All **system outputs** evaluated against a benchmark using Evall can be **stored in the framework**, so future researchers can compare with them. Researchers can also associate some useful information such a brief description, a reference to a paper, etc.

Apart from these main features, Evall also includes other important features such as significance statistical tests, smoothing process, personalization of reports, manual selection of measures, standardization of the input format across tasks, or warnings and statistics about the system outputs. In summary, with a single click the user obtains edited information in Latex format with his/her results in terms of multiple measures, statistical significance tests, and system output data checking, as well as information about the categories, properties and limitation of the measures.

Up to day, and to the best of our knowledge, there are no standard evaluation frameworks or tools specially designed for this and covering a wide range of information systems tasks. There are some implementations of a small set of measures included in tools developed for another purpose, such as Weka[1], Gate[2] or Open NLP[3]. There are also some specific designed tools for evaluation for concrete problems such as machine translation (IQmt[4]), but there is not a universal and dedicated evaluation framework for information systems tasks.

3 What can you do in Evall?

In the design and development of Evall several scenarios have been taken into account:

Browsing scenario: The user is interested in exploring the existing tasks, measures, benchmarks or evaluation results stored in Evall by others researchers. The web interface allows the user to explore, learn and access to all relevant information stored

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] https://gate.ac.uk/
[3] https://opennlp.apache.org/
[4] http://www.cs.upc.edu/ nlp/IQMT/

in Evall, such as accessing the system outputs rankings associated to the benchmarks.

Evaluating against benchmarks included in Evall: The user is interested in comparing his/her results with the state of the art. Given a previously stored benchmark (i.e. Trec-2013), with only one click, the user can obtain a report comparing his/her own output with the baselines and best approaches stored in Evall for this benchmark. The report includes measures descriptions and suitability, evaluation results and salient aspects extracted from resulting data, as well as information about statistical significance tests. Furthermore, it compares the evaluation results against theoretical baseline approaches such as random systems. The user will need to (i) select an Evall benchmark and (ii) provide the system outputs in the Evall format.

Evaluating against his/her own benchmark: The user has defined his/her own benchmark. In this case the user has to (i) indicate the type of task (i.e. Classification), (ii) provide his/her gold standard in Evall format (iii), and provide the system outputs in Evall format.

Expert scenario: The user understands the nature and suitability of measures. He is interested in executing measures under a wide set of approaches and going beyond the capabilities of Evall standard reports by using the results obtained in the set of CSV files generated. Evall returns a set of CSV files containing the evaluation results for each system, test case and measure, as well as the aggregated results for each system. Evall gives also the possibility of customizing the evaluation results, by selecting measures and parameters.

System output contribution: The user can contribute with new system outputs against a benchmark included in Evall and store the results including some interesting information such as a brief description, a reference to a paper where the approach is better described, etc.

Benchmark contribution: A user, specially workshops and evaluation campaigns organizers, is interested in sharing his/her data under a common evaluation framework such as Evall and generating the results of the competition using Evall, allowing to compare the results achieved by the participants even when the evaluation campaign has been finished.

4 Evall inputs formats

Evall works under a general theoretical framework which categorizes problems in terms of measurement theory. Evall is based on the idea that system outputs or gold standards are measurements. That is, values assigned to items. This information can be captured with tuples containing the test case (i.e. queries in information retrieval), and an item and a value. For instance, the following table (Table 1) represents a relevance measurement for four documents in the context of two queries, (test cases), in a Ranking scenario.

Query_1	d31	0.5
Query_1	d52	0.2
Query_2	d31	0.7
Query_2	d25	0.3

Table 1: Example of Evall input format for ranking tasks

Given that the system output is a ranking (ordinal measurement), the last column can be avoided by considering directly the document order for the same query.

Query_1	d31	-
Query_1	d52	-
Query_2	d31	-
Query_2	d25	-

Table 2: Example of Evall input format for ranking tasks without absolute values

In the case of classification or clustering tasks, the measurement is nominal. Therefore, the relative order of values is not relevant, and the values can be strings as in the following example (Table 3):

Test_case_1	d21	Sport
Test_case_1	d23	World news
Test_case_2	d34	World news
Test_case_2	d43	World news

Table 3: Example of Evall input format for clustering and classification tasks

In the Evall framework, this is the common format for any output in any task: a three column CSV standard format. This for-

mat is used both for system outputs and gold standards.

5 Evall coverage: tasks and measures

In terms of tasks, Evall covers most of existing information access evaluation campaigns. For instance, examining the evaluation campaigns Semeval 2013 and 2014 and Clef 2014, we have checked that the Evall tasks cover 30 from 37 tasks or subtasks. The non covered problems include temporal intervals extraction, some text evaluation metrics (i.e. ROUGE), user based evaluation, and evaluation of structures. We expect to add some of them in future versions.

In terms of metrics, in particular the current Evall prototype covers the following sets:

Classification evaluation scenario consists in comparing the labels produced by systems for each item with the value provided by the gold. Both the gold and system outputs are nominal measurements. That is, the absolute difference or ordering relationships between values is not relevant for evaluation purposes. The classification measures provide with Evall includes: **Accuracy**, **Accurate Output**, **Weighted Accuracy**, **Utility** (Cormack and Lynam, 2005), **Lam%** (Hull, 1998), **Macro Average Accuracy**, **Kappa statistic** (Cohen, 1960), **Mutual Information**, **Precision**, **Recall**, **F-measure**, **Reliability** and **Sensitivity** (Amigó, Gonzalo, and Verdejo, 2013).

The **Ranking** evaluation scenario focuses on the priority relationships between items in both the gold and system outputs. That is, the priority relationships in the gold must be reflected in the system output. The measure set in Evall includes binary relevance measure such as **Precision at K**, **R-Precision**, **Mean Reciprocal Rank**, **Mean Average Precision**, and graded relevance measures such as **DCG** (Järvelin and Kekäläinen, 2002), **ERR** or **RBP** (Moffat and Zobel, 2008).

Clustering can be interpreted as the problem of predicting if two items belong or not to the same group. That is, predicting equality relationships between items. This corresponds with the fact of checking the relationship equivalence between two nominal measurements (the system output and the gold standard). The measures available in Evall are grouped into five categories: Set matching (**Purity** and **Inverse Purity**, **F-measure**), Entropy Based (**Class and Cluster entropy** (Steinbach, Karypis, and Kumar, 2000; Ghosh, 2003), **Mutual Information** (Xu, Liu, and Gong, 2003), **Counting Pairs** (Rand, Jaccar and F&M statistics (Halkidi, Batistakis, and Vazirgiannis, 2001; Meila, 2003)), **Editing Distance** and **Bcubed**.

References

Amigó, E., J. Gonzalo, and F. Verdejo. 2013. A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference*, pages 643–652, New York, USA.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Cormack, G. V. and T. R. Lynam. 2005. Trec 2005 spam track overview. In *TREC*.

Ghosh, J. 2003. Scalable clustering methods for data mining. In N. Ye, editor, *Handbook of Data Mining*. Lawrence Erlbaum.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis. 2001. On Clustering Validation Techniques. *J. of Int. Inf. Syst.*, 17(2-3):107–145.

Hull, D. A. 1998. The TREC-7 filtering track: description and analysis. In *Proc. of TREC-7, 7th Text Retrieval Conf.*, pages 33–56.

Järvelin, K. and J. Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October.

Meila, M. 2003. Comparing clusterings. In *Proceedings of COLT 03*.

Moffat, A. and J. Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December.

Steinbach, M., G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques.

Xu, W., X. Liu, and Y. Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proc. of the 26th annual Int. ACM SIGIR Conf.*, pages 267–273.

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/home-2/revista/instrucciones-autor/

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 : .....
IBAN

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 : .....

IBAN _____|_____|_____|_____|_____|_____

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 2*4 €* (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 Maillo
UNED
felisa@lsi.uned.es

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Antònia Martí Antonín	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Raquel Martínez	Universidad Nacional de Educación a Distancia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lidia Moreno	Universidad Politécnica de Valencia (España)
Lluís Padró	Universidad Politécnica de Cataluña (España)
Manuel Palomar	Universidad de Alicante (España)
Ferrán Pla	Universidad Politécnica de Valencia (España)
German Rigau	Universidad del País Vasco (España)
Horacio Rodríguez	Universidad Politécnica de Cataluña (España)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé	Universidad de Barcelona
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignon / Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maillo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/category/revista/consejo_redaccion/
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/home-2/revista/consejo-asesor/
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/socios/inscripcion-para-socios/

© 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

Proyectos

Demostraciones

Información General