sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Artículos

Tesis

Información General

ISSN: 1135-5948

Comité Editorial

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2013
Editores: Mariona Taulé Delor, Universidad de Barcelona, mtaule@ub.edu
Mª Teresa Martín Valdivia, Universidad de Jaén, maite@ujaen.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

Ferrán Pla	Universidad Politécnica de Valencia
German Rigau	Universidad del País Vasco
Horacio Rodríguez	Universidad Politécnica de Cataluña
Kepa Sarasola	Universidad del País Vasco
Emilio Sanchís	Universidad Politécnica de Valencia
Mariona Taulé	Universidad de Barcelona
L. Alfonso Ureña	Universidad de Jaén
Felisa Verdejo	UNED
Manuel Vilares	Universidad de A Coruña
Ruslan Mitkov	Universidad de Wolverhampton, UK
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues, France
Leonel Ruiz Miyares	Centro de Linguística Aplicada de Santiago de Cuba
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Alexander Gelbukh	Instituto Politécnico Nacional, México
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores, Portugal
Bernardo Magnini	Fondazione Bruno Kessler, Italia

Revisores adicionales

Emmanuel Anguiano Hernández	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Ester Boldrini	Universidad de Alicante, España
Lluis F. Hurtado	Universidad Politécnica de Valencia, España
Daniel Fernández González	Universidad de Vigo, España
Victor Manuel Flores Fonseca	Universidad Politécnica de Madrid, España
Miguel Ángel García Cumbreras	Universidad de Jaén, España
Elena Lloret Pastor	Universidad de Alicante, España
Eugenio Martínez Cámara	Universidad de Jaén, España
Fernando Martínez Santiago	Universidad de Jaén, España
Arturo Montejo Ráez	Universidad de Jaén, España
Adrián Pastor López Monroy	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Jesús Peral Cortés	Universidad de Alicante, España
Paolo Rosso	Universidad Politécnica de Valencia, España
David Vilares Calvo	Universidad de A Coruña, España
Francisco Viveros	Instituto Politécnico Nacional, México
Alisa Zhila	Instituto Politécnico Nacional, México

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

comunicaciones científicas y resúmenes de tesis. 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 28 trabajos para este número de los cuales 24 eran artículos científicos y 4 correspondían a resúmenes de tesis. De entre los 24 artículos recibidos 11 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 45,8%.

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.

Marzo de 2013
Los editores

sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Preamble

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

The subject areas addressed are the following:

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

The 50th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by the traditional peer reviewed

process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Twenty-eight papers were submitted for this issue of which twenty-four were scientific papers and four dissertation summaries. From these twenty-four papers, we selected eleven (45.8% 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.

March 2013
Editorial board

sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 13-20 recibido 23-10-12 revisado 14-01-13 aceptado 19-02-13

Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias

Polarity classification of opinionated Spanish texts using dependency parsing

David Vilares, Miguel A. Alonso y **Carlos Gómez-Rodríguez**
Departamento de Computación, Universidade da Coruña
Campus de Elviña, 15011 A Coruña
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Resumen: En este artículo se describe un sistema de minería de opiniones que clasifica la polaridad de textos en español. Se propone una aproximación basada en PLN que conlleva realizar una segmentación, tokenización y etiquetación de los textos para a continuación obtener la estructura sintáctica de las oraciones mediante algoritmos de análisis de dependencias. La estructura sintáctica se emplea entonces para tratar tres de las construcciones lingüísticas más significativas en el ámbito que nos ocupa: la intensificación, las oraciones subordinadas adversativas y la negación. Los resultados experimentales muestran una mejora del rendimiento con respecto a los sistemas puramente léxicos y refuerzan la idea de que el análisis sintáctico es necesario para lograr un análisis del sentimiento robusto y fiable.
Palabras clave: Minería de opiniones, Análisis del sentimiento, Análisis sintáctico de dependencias

Abstract: This article describes an opinion mining system that classifies the polarity of Spanish texts. We propose a NLP-based approach which performs segmentation, tokenization and POS tagging of texts to then obtain the syntactic structure of sentences by means of a dependency parser. The syntactic structure is then used to address three of the most significant linguistic constructions in the area in question: intensification, adversative subordinate clauses and negation. Experimental results show an improvement in performance with respect to purely lexical approaches and reinforce the idea that parsing is required to achieve a robust and reliable sentiment analysis system.
Keywords: Opinion Mining, Sentiment Analysis, Dependency Parsing

1. Introducción

El auge en los últimos años de los blogs, los foros y las redes sociales ha hecho que millones de usuarios utilicen estos recursos para expresar sus opiniones sobre toda una variedad de temas. La diversidad y cantidad de críticas presentes en la web resultan de gran utilidad a fabricantes y vendedores, que ven en ellas un mecanismo para conocer de primera mano cómo sus artículos son percibidos por los consumidores. Los beneficios asociados a conocer toda esta información, sumados a la complejidad técnica del análisis de las opiniones, han provocado que se hayan comenzado a demandar soluciones capaces de monitorizar este flujo ingente de reseñas.

Todo ello ha contribuido a que la minería de opiniones (MO), también conocida como análisis del sentimiento, esté jugando un papel importante como ámbito de investigación en los últimos años. La MO se centra en tratar automáticamente información con opinión, lo que permite, entre otras cosas, extraer la polaridad (positiva, negativa, neutra o mixta) de un texto (Pang y Lee, 2008).

En este artículo presentamos un sistema de clasificación de polaridad para textos escritos en español, cuyas principales características son la utilización de diccionarios semánticos y de la estructura sintáctica de las oraciones para clasificar un texto subjetivo como positivo o negativo. La utilidad práctica de esta aproximación viene avalada por los resultados experimentales presentados, que muestran una mejora en precisión de más de cuatro puntos porcentuales con respecto a un sistema reciente que no hace uso de la sintaxis.

ISSN 1135-5948

El resto del artículo se organiza como sigue. En la sección 2 se revisa brevemente la situación actual de la MO, centrándose en lo referido a la detección de la polaridad. En la sección 3 se describe la propuesta planteada y se detallan los aspectos sintácticos tratados. En la sección 4 se muestran detalles de implementación y los resultados de los experimentos realizados. Por último, en la sección 5 se presentan las conclusiones y las principales líneas de trabajo futuras.

2. Estado del arte

Una parte importante de los esfuerzos actuales relacionados con la MO se están realizando en tareas relativas a la clasificación de la polaridad, problema que ha sido abordado desde dos enfoques principales. El primero asume esta tarea como un proceso genérico de clasificación (Pang, Lee, y Vaithyanathan, 2002): a partir de un conjunto de entrenamiento, donde los textos son anotados con su polaridad, se construye un clasificador mediante aprendizaje automático (AA). El segundo enfoque se apoya en la orientación semántica (OS) de las palabras, donde cada término que expresa opinión es anotado con un valor que representa su polaridad (Turney, 2002). Este segundo enfoque es el que tomaremos como base para el desarrollo de nuestro trabajo.

La mayor parte de los sistemas de MO se centran en el tratamiento de textos en inglés. En el caso de textos escritos en español, probablemente el sistema más relevante sea *The Spanish SO Calculator* (Brooke, Tofiloski, y Taboada, 2009), desarrollado en la Universidad Simon Fraser de Canadá. Este sistema, además de resolver la OS almacenada a nivel individual en adjetivos, sustantivos, verbos y adverbios; trata modificadores de la polaridad como son la negación o los intensificadores (*"muy", "poco", "bastante", ...*). También detecta y descarta el sentimiento reflejado en el contenido no fáctico del texto, representado, por ejemplo, mediante expresiones condicionales o subjuntivas.

La manera más habitual de tratar todas estas construcciones lingüísticas es a nivel léxico y en este aspecto The Spanish SO Calculator no es una excepción. En lo que respecta al tratamiento de la negación, (Taboada et al., 2011) utiliza información morfológica para identificar el alcance de la negación, mientras que (Yang, 2008) considera dicho alcance como los términos situados a la derecha de la negación y en (Fernández Anta et al., 2012) se emplea una heurística que asume que los tres elementos a continuación de una negación son los que deben cambiar su polaridad. Para la intensificación, (Fernández Anta et al., 2012) considera de nuevo que los tres términos a la derecha son los que deben variar su polaridad. (Taboada et al., 2011) además de los intensificadores propiamente dichos, trata como tales aspectos del discurso como la conjunción *"pero"* o las mayúsculas.

Nuestra propuesta sigue una estrategia distinta, que se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüísticas e identificar los elementos de la frase que están implicados en ellas. A este respecto, trabajos anteriores (Jia, Yu, y Meng, 2009) ya han mostrado los beneficios de utilizar la estructura sintáctica de la frase en aquellos textos en los que aparecen ocurrencias de términos negativos.

Un problema adicional al que se enfrentan los sistemas de MO es la calidad ortográfica de los textos a analizar. Cuando éstos provienen de la web, debe tenerse en cuenta que es frecuente que sus autores omitan acentos, letras o vocablos; o empleen tanto abreviaturas no reconocidas como oraciones agramaticales. La solución más utilizada consiste en emplear patrones heurísticos para adaptar el texto (Saralegi Urizar y San Vicente Roncal, 2012; Martínez Cámara et al., 2012) .

3. Clasificación de opiniones basada en dependencias sintácticas

En contraste con las propuestas léxicas dominantes hasta el momento, en este trabajo proponemos la utilización de la estructura sintáctica de la frase para obtener la OS de un texto. Como primer paso, es necesario preprocesar los textos, para ello se ha diseñado un preprocesador *ad-hoc* que trata los siguientes aspectos:

- La unificación de expresiones compuestas, que actúan como una sola unidad de significado (*"a menos que", "en absoluto",...*).
- La normalización de los signos de puntuación. En un entorno web es común obviar las normas ortográficas respecto a la colocación de signos como el punto o la coma, lo que puede afectar negativamente al resto del procesado.

A continuación, se procede a segmentar el texto en oraciones y a dividir cada una de ellas en *tokens* (principalmente para palabras, pero también signos de puntuación, números, etc.) para después realizar la etiquetación morfosintáctica de cada una de las palabras del texto.

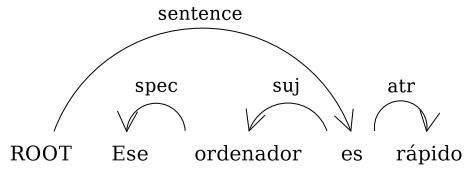

Figura 1: Ejemplo de árbol de dependencias

El siguiente paso consiste en realizar el análisis sintáctico de dependencias mediante el cual se identifican relaciones binarias padre/dependiente entre los términos de una oración. Se incluye un enlace con un elemento artificial inicial (ROOT) para facilitar las definiciones formales e implementaciones. Cada uno de esos vínculos binarios constituye una dependencia, que se anota con la función sintáctica que relaciona los dos términos. A la estructura obtenida se le denomina árbol de dependencias. En la Figura 1 se ilustra un ejemplo sencillo de este tipo de análisis. Como corpus de referencia para la definición de las relaciones de dependencia se ha utilizado Ancora (Taulé, Martí, y Recasens, 2008).

Finalmente, para la realización del análisis semántico, nuestra propuesta se apoya en el SODictionariesV1.11Spa (Brooke, Tofiloski, y Taboada, 2009). Se trata un conjunto de diccionarios de polaridad para adjetivos, sustantivos, verbos, adverbios e intensificadores; cuyo contenido se resume en la Tabla 1. Cada término se encuentra anotado con un valor entre -5 y 5, donde -5 es lo más negativo y 5 lo más positivo. El valor asignado a cada palabra se corresponde con una orientación semántica genérica, independientemente del dominio o contexto en el que se utilice. Así, por ejemplo, al adjetivo *"rápido"* o al verbo *"recomendar"* se les asocia una polaridad de valor 2. Es importante señalar que los valores numéricos asociados a los intensificadores tienen un significado distinto, ya que representan el porcentaje (positivo o negativo) por el que modifican el sentimiento de la expresión a la que afectan.

Diccionario	**Nº términos**
adjetivos	2,049
sustantivos	1,324
verbos	739
adverbios	548
intensificadores	157

Tabla 1: Contenido del SODictionariesV1.11Spa

3.1. Propuesta base

Nuestra versión inicial determina la polaridad de un texto únicamente a partir de la combinación de la OS de sustantivos, adjetivos, verbos y adverbios; esto es, sin considerar ninguna construcción lingüística compleja, lo que equivale a ignorar la estructura sintáctica del texto. En la Figura 2 se ilustra un ejemplo de análisis de la OS sobre el árbol de dependencias correspondiente a la oración *"Ese ordenador es muy rápido, pero no recomiendo comprarlo"*. Podemos observar que la propuesta base establece una OS muy positiva para un texto que intuitivamente se percibe como ligeramente negativo. Se trata de un ejemplo didáctico que refleja los problemas de obviar fenómenos como la intensificación, los nexos adversativos o la negación a la hora de extraer completamente la polaridad.

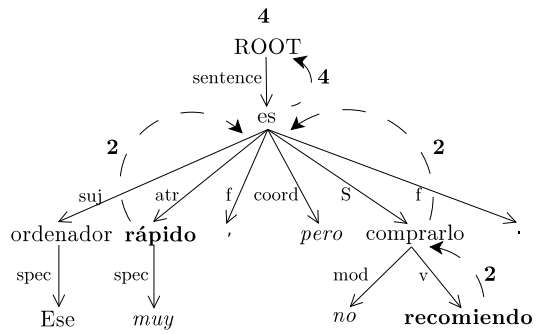

Figura 2: Análisis semántico sobre árbol de dependencias

3.2. Tratamiento de la intensificación

Los intensificadores son términos o expresiones que modifican la polaridad de ciertas palabras. Consideraremos dos tipos: *amplificadores*, si permiten aumentar la polaridad (*"muy", "bastante",...*), y *decrementadores* si la disminuyen (*"poco", "en absoluto",...*). Para modelar esta construcción se asocia a cada

intensificador un factor de ponderación. Así, basándonos en el SODictionariesV1.11Spa, al amplificador *"muy"* se le asocia el valor 0,25 y al decrementador *"en absoluto"*, -1. La principal diferencia radica en que nuestra propuesta utiliza el árbol de dependencias para determinar la parte de la frase que se ve afectada por tal modificación, considerando las dependencias anotadas en Ancora como *spec, espec, cc* o *sadv*.

Para el ejemplo presentado en la Figura 2, la OS de *"muy rápido"* se obtendría incrementando en un 25 % la OS de *"rápido"*: $2 * (1 + 0,25) = 2,5$. En caso de que haya varios intensificadores, se combinan todos sus porcentajes de intensificación antes de que actúen sobre el término afectado. Por ejemplo, si la expresión intensificada fuese *"en absoluto muy rápido"* la OS se obtendría como $2 * (1 + (-1 + 0,25)) = 0,5$.

En un entorno web existen otras formas de enfatizar opiniones, como son el empleo de mayúsculas o de exclamaciones. Hemos tratado estas peculiaridades siguiendo un enfoque similar al del resto de intensificadores.

3.3. Tratamiento de las oraciones adversativas

Los nexos adversativos permiten contraponer hechos expresados en dos oraciones. En un entorno de MO este tipo de frases se emplean para restringir o excluir opiniones, lo que puede ser considerado como un caso especial de intensificación. Disponer de un árbol de dependencias resulta de gran utilidad en este caso, ya que nos permite identificar con precisión tanto la oración subordinada como la subordinante. Desafortunadamente, el corpus de Ancora representa sintácticamente este tipo de oraciones de forma distinta según el nexo concreto utilizado, por lo que el tratamiento realizado para este tipo de cláusulas no ha sido todo lo completo que nos hubiera gustado. Hemos optado por centrarnos en los nexos más relevantes que Ancora representa de manera uniforme. Se han dividido estos nexos en dos grupos: los *restrictivos*, que reducen la OS de la oración principal y donde destaca la conjunción *"pero"*; y los *excluyentes*, que eliminan enteramente lo expresado en la primera oración, entre los que se encuadran conjunciones como *"sino"*. Así, según la clase de nexo, se pondera el sentimiento acumulado, tanto en la oración subordinante como en la subordinada, de forma distinta. En la Tabla 2 se ilustran los factores de ponderación $F_{principal}$ y $F_{subordinada}$, establecidos mediante una evaluación empírica del SFU Spanish Review Corpus, cuyo contenido se detalla en la sección 4.2.

Nexo	$F_{principal}$	$F_{subordinada}$
Restrictivo	0,75	1,4
Excluyente	0	1

Tabla 2: Factores de ponderación según el tipo de nexo adversativo

Para homogeneizar en un futuro la estructura sintáctica de otras subordinadas adversativas, y para simplificar la ponderación de estas oraciones; se optó por reestructurarlas en el árbol de dependencias. En la Figura 3 se ilustra la estructura esquemática de una oración adversativa una vez reorganizada. Se observa que en el nivel superior de la cláusula subordinada se incluye un nodo de apoyo, representado por **. Se crea también un nuevo tipo de de dependencia, *art_rel_adversative*, para identificar sintácticamente el inicio de una cláusula de este tipo. Si se retoma el ejemplo de la Figura 2, donde aparecen dos oraciones conectadas por la conjunción adversativa *"pero"*, la estructura sintáctica reorganizada sería la ilustrada en la Figura 4.

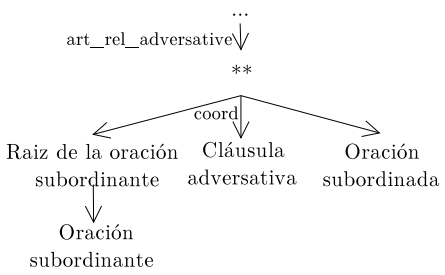

Figura 3: Reestructuración de oraciones adversativas

3.4. Tratamiento de la negación

Son muchos los términos o expresiones que permiten negar una opinión. Sin embargo, la frontera entre un negador como tal y un intensificador decrementador es difusa. En este trabajo se ha restringido el tratamiento de este fenómeno a los términos *"no"*, *"nunca"* y *"sin"*. Otras expresiones negadoras, como *"lo menos"* o *"en absoluto"*, han sido tratadas como intensificadores. Para ello, se ha aprovechado la información semántica proporcio-

nada por el SODictionariesV1.11Spa para este tipo de locuciones.

Para resolver el sentimiento de una oración con ocurrencias de términos negativos es necesario realizar dos pasos: identificar el alcance de la negación y modificar la polaridad del fragmento de la oración correspondiente.

3.4.1. Identificación del alcance de la negación

Nuestra estrategia para identificar el alcance de la negación se basa en la propuesta de (Jia, Yu, y Meng, 2009). Sin embargo, el procedimiento ha sido adaptado a las peculiaridades del análisis sintáctico realizado. Las características del árbol de dependencias permiten definir un procedimiento estrictamente sintáctico, basado en las relaciones entre elementos, sin necesidad de localizar delimitadores léxicos.

La forma de identificar ese alcance difiere según el negador utilizado. Cuando se emplea el término *"sin"*, el árbol de dependencias nos asegura que la rama descendiente constituye el alcance de ese negador, sin necesidad de analizar el tipo de relación. Por contra, la estructura sintáctica utilizada para representar los elementos *"no"* y *"nunca"*, requiere identificar dependencias concretas como *neg* o *mod*, e iniciar un proceso más complejo. En primer lugar, se establece un alcance candidato, formado tanto por el padre del negador como por sus hermanos. A continuación se corrige dicho alcance aplicando una serie de reglas heurísticas, que son procesadas en orden hasta que una cumpla los requisitos:

1. *Regla del padre subjetivo*: Si el padre del negador aparece en los diccionarios semánticos, entonces sólo él constituye el alcance corregido de la negación.

2. *Regla del atributo o complemento directo*: Si alguno de los hermanos desempeña una de estas funciones sintácticas, entonces dicho hermano forma parte del alcance de la negación.

3. *Regla del complemento circunstancial más cercano*: Si alguna rama al mismo nivel del negador actúa cómo complemento circunstancial, entonces dicha rama forma el alcance corregido. En caso de varios complementos circunstanciales, sólo se incorpora el más cercano físicamente al negador.

Si ninguna regla se cumple, entonces se asume el alcance candidato (salvo el nodo padre) como el corregido. En el ejemplo de la Figura 4, para la negación *"no recomiendo comprarlo"*, ninguna de la reglas se cumple, por lo que el alcance corregido estaría formado sólo por el verbo *"recomiendo"*.

3.4.2. Modificación de la polaridad

Nuestra propuesta para resolver la modificación de la polaridad que implica una negación es similar a la empleada en trabajos como (Taboada et al., 2011). Una vez obtenido el alcance corregido, se extrae su polaridad, y a continuación, el valor obtenido es modificado en una cantidad preestablecida de signo contrario. Para los negadores *"no"* y *"nunca"*, dicho valor es 4, mientras que para *"sin"* el valor es menor, 3,5, para ajustarse a su carácter más local. Así, en el ejemplo de la Figura 4, se observa como para la negación de *"recomiendo"* se obtiene una OS de -2.

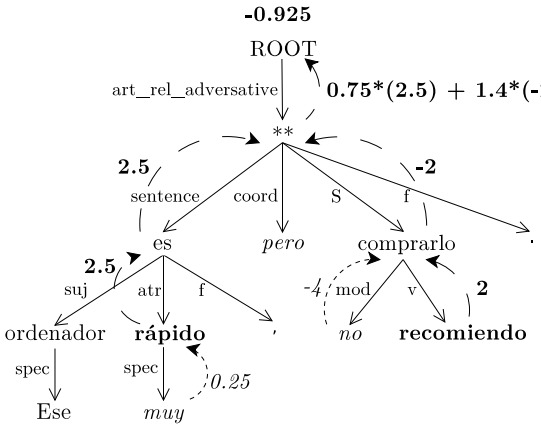

Figura 4: Análisis final de la OS sobre el árbol de dependencias reestructurado

4. Resultados experimentales

4.1. Implementación

Nuestra propuesta para la clasificación de la polaridad se ha implementado en Python, apoyado en el *toolkit* NLTK[1] para las tareas de segmentación, tokenización y etiquetación. En concreto, para la tarea de etiquetación se ha aplicado el algoritmo de Brill utilizando el corpus Ancora (Taulé, Martí, y Recasens, 2008) para el entrenamiento (se ha utilizado el 90 % del corpus para el entrenamiento y el 10 % restante para la evaluación). Para mejorar el rendimiento práctico del etiquetador

[1] http://nltk.org/

sobre el análisis de textos de la web, donde se obvian los acentos en muchas palabras, el fragmento del corpus destinado al aprendizaje fue ampliado de forma que cada oración dispusiese de su equivalente sin palabras acentuadas gráficamente. Los resultados de la evaluación del etiquetador, mostrados en la Tabla 3, sugieren que las ambigüedades creadas por esta duplicación apenas afectan a la precisión teórica del etiquetador y, sin embargo, se comprobó empíricamente que mejoraba la anotación sobre textos no acentuados.

Corpus	Precisión
Original	0,9586
Ampliado	0,9571

Tabla 3: Precisión del etiquetador de Brill

La tarea del análisis sintáctico de dependencias se ha realizado con el algoritmo *Nivre arc-eager* (Nivre, 2008) generado con MaltParser[2] (Nivre et al., 2007) mediante aprendizaje automático a partir del corpus Ancora.

En la sección anterior se comentó cómo se han tratado algunas construcciones de naturaleza sintáctica, sin embargo, hay aspectos que no pueden resolverse a ese nivel. Ejemplo de ello es la mayor importancia de las oraciones finales de una opinión. Para modelar esta peculiaridad, en nuestra propuesta se ha optado por aumentar en un 75 % la OS de las tres últimas frases de una crítica.

Otro aspecto a tener en cuenta es el introducido en (Kennedy y Inkpen, 2006), donde se habla del problema de la tendencia positiva del lenguaje humano. Al expresar una opinión negativa, es frecuente utilizar negaciones de términos positivos en lugar de los correspondientes antónimos; *"no barato"* en vez de *"caro"* o *"no bueno"* en vez de *"malo"* son dos ejemplos de esta situación. Para compensar dicha desviación, muchas aproximaciones léxicas incrementan la OS de los términos negativos, mejorando notablemente su rendimiento. Sin embargo, el empleo de esta técnica en nuestra propuesta resultó contraproducente. Sí se consiguió mejorar la precisión del sistema aumentando la dispersión de las OS de sustantivos, adjetivos, verbos y adverbios del SODictionariesV1.11Spa en un 20 %, esto es, que sus polaridades comprendan valores entre -6 y 6. Todos los aspectos que incrementaron el rendimiento se incluyeron en la versión final de nuestro sistema.

4.2. Evaluación

Para la evaluación de nuestra propuesta se ha empleado un corpus formado por 400 documentos: el SFU Spanish Review Corpus (Brooke, Tofiloski, y Taboada, 2009). Contiene reseñas de productos y servicios de ocho categorías distintas: lavadoras, hoteles, películas, coches, ordenadores, libros, música y móviles. Para cada categoría se dispone de un total de 50 documentos, donde en 25 de ellos se expresa una opinión positiva, mientras los otros 25 expresan una negativa.

Nuestra propuesta procesa cada texto y obtiene como resultado su OS, si ésta es mayor que 0 el texto se cataloga como positivo, en caso contrario como negativo. En la Tabla 4 se ilustra la precisión para distintas configuraciones. Todas las construcciones lingüísticas tratadas han mejorado el rendimiento. Especialmente significativo es el incremento obtenido con la incorporación de la negación. Se realizaron test chi-cuadrado ($p < 0,01$), comparando con las polaridades correctas. Con un * se ilustran las configuraciones para las que se obtuvieron polaridades que no difieren de manera estadísticamente significativa de las correctas.

Propuesta	Precisión
Base	0,618
+ intensificación	0,660
+ adversativas	0,670
+ negación	0,755*
Final	0,785*

Tabla 4: Precisión al incorporar distintas funcionalidades

Haber utilizado para la evaluación el mismo corpus y los mismos diccionarios semánticos que la solución léxica The Spanish SO-Calculator, permite comparar nuestra alternativa sintáctica con ella. En la Tabla 5 se contrasta el rendimiento. Nuestra propuesta incrementa en un 5,72 % el rendimiento obtenido por The Spanish SO-CAL. También se construyó un clasificador SVM, basado en AA, empleando para ello WEKA[3]. Para su desarrollo, se utilizó el SFU Spanish Review Corpus y como método de evaluación se optó por

[2] http://www.maltparser.org/

[3] http://www.cs.waikato.ac.nz/ml/weka/index.html

una validación cruzada de 10 iteraciones. Todos los términos se cambiaron a su forma minúscula y se utilizó su frecuencia absoluta de aparición. (Brooke, Tofiloski, y Taboada, 2009) también propone un sistema de AA, incluyendo PLN, pero sus resultados no mejoran los presentados con nuestra configuración.

Método	Precisión (%)
Nuestra propuesta	**78,50**
The Spanish SO-CAL	74,25
SVM	72,50

Tabla 5: Precisión para distintos métodos

En la Tabla 6 se muestra la precisión de la versión final del sistema, desglosada para las categorías del corpus. Para los ámbitos considerados de entretenimiento, como las películas o los libros; el rendimiento es peor que la media. Hay dos razones posibles. La primera es referida al empleo de OS genéricas. Términos como *"guerra"* o *"asesino"* son manifiestamente negativos, sin embargo, en dominios relacionados con las novelas o las películas, probablemente describan la temática o el argumento, sin afectar a la calidad del producto. El segundo motivo está relacionado con los gustos personales, lo que complica clasificar la polaridad de ciertos términos en estos ámbitos. Por el contrario, se obtienen mejores resultados en dominios donde los criterios de calidad están claramente establecidos, como es el caso de los hoteles o los ordenadores.

Categoría	Neg	Pos	Total
Lavadoras	0,79	0,86	0,82
Hoteles	0,88	0,92	0,90
Películas	0,67	0,65	0,66
Coches	0,77	0,71	0,74
Ordenadores	0,91	0,82	0,86
Libros	0,80	0,70	0,74
Música	0,84	0,71	0,76
Móviles	0,86	0,76	0,80

Tabla 6: Precisión según categoría

El sistema, con la misma configuración, se evaluó también sobre HOpinion[4] (críticas de hoteles) y sobre CorpusCine (Cruz, Troyano, y Ortega, 2008), para los que se obtuvo una precisión global de 0,89 y 0,64, respectivamente. Es interesante reseñar que estos resultados son similares a los obtenidos para las categorías de hoteles y películas, respectivamente, en el SFU Spanish Review.

5. Conclusiones y trabajo futuro

Este artículo describe una estrategia para resolver la OS de textos con opinión empleando técnicas de análisis de dependencias. Los experimentos realizados confirman que la utilización de la sintaxis resulta útil a la hora de tratar construcciones lingüísticas en un entorno de MO, como son la negación, la intensificación y las frases adversativas. A este respecto, el análisis que se ha hecho de la negación evita contrarrestar artificialmente la tendencia positiva del lenguaje humano. Esto nos sugiere que se está realizando una identificación fiable del alcance de la negación.

En busca de futuras mejoras, tratar las expresiones y construcciones desiderativas es una línea de trabajo que nos gustaría explorar. También se ha planeado realizar una evaluación más exhaustiva con otros algoritmos de análisis sintáctico de dependencias, como el 2-planar (Gómez-Rodríguez y Nivre, 2010).

La evaluación de nuestra propuesta se realizó sobre un corpus de textos extensos creado por (Brooke, Tofiloski, y Taboada, 2009). Al respecto, el éxito de redes como Twitter ha aumentado el interés por analizar textos cortos (Villena-Román et al., 2013), por lo que sería interesante poder evaluar y adaptar nuestro sistema a las características de este tipo de documentos.

Ciertos factores que afectan a la clasificación de la polaridad no se han considerado. Por ejemplo, el problema de la polaridad cambiante para determinados términos según el dominio en el que aparezcan (Pang y Lee, 2008). La ironía o el sarcasmo son dos figuras literarias que se utilizan para expresar una opinión de una forma mucho más creativa y sutil, lo que dificulta su tratamiento y su identificación. A este respecto, en (Reyes y Rosso, 2011) se describe una aproximación para detectar la ironía que podría ser utilizada para enriquecer nuestra propuesta.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y FEDER (TIN2010-18552-C03-02) y por la Xunta de Galicia (CN2012/008,

[4]http://clic.ub.edu/corpus/hopinion

CN 2012/319).

Bibliografía

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54, Borovets, Bulgaria. ACL.

Cruz, F., J. A. Troyano, y J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. En *Procesamiento de lenguaje natural, 41*, páginas 81–87.

Fernández Anta, A., P. Morere, L. Núñez Chiroque, y A. Santos. 2012. Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report. En *TASS 2012 Working Notes*, Castellón, Spain.

Gómez-Rodríguez, C. y J. Nivre. 2010. A transition-based parser for 2-planar dependency structures. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, páginas 1492–1501, Stroudsburg, PA, USA. ACL.

Jia, L., C. Yu, y W. Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. En *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM'09, páginas 1827–1830, New York, NY, USA. ACM.

Kennedy, A. y D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Martínez Cámara, E., M. T. Martín Valdivia, M. A. García Cumbreras, y L. A. Ureña López. 2012. SINAI at TASS 2012. En *TASS 2012 Working Notes*, Castellón, Spain.

Nivre, J. 2008. Algorithms for deterministic incremental dependency parsing. *Compuational Linguistics*, 34(4):513–553.

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(2):95–135.

Pang, B. y L. Lee. 2008. *Opinion Mining and Sentiment Analysis*. now Publishers Inc., Hanover, MA, USA.

Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proceedings of EMNLP*, páginas 79–86.

Reyes, A. y P. Rosso. 2011. Mining subjective knowledge from customer reviews: a specific case of irony detection. En *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, páginas 118–124, Stroudsburg, PA, USA. ACL.

Saralegi Urizar, X. y I. San Vicente Roncal. 2012. Detecting Sentiments in Spanish Tweets. En *TASS 2012 Working Notes*, Castellón, Spain.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Taulé, M., M. A. Martí, y M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. En Nicoletta Calzolari Khalid Choukri Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis, y Daniel Tapias, editores, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, páginas 417–424, Stroudsburg, PA, USA. ACL.

Villena-Román, J., S. Lana-Serrano, J.C. González Cristóbal, y E. Martínez-Cámara. 2013. TASS Worshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural, 50.*

Yang, K.. 2008. WIDIT in TREC 2008 blog track: Leveraging multiple sources of opinion evidence. En E. M. Voorhees y Lori P. Buckland, editores, *NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008).*

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 21-28 recibido 19-11-12 revisado 08-02-13 aceptado 19-02-13

Análisis de similitud basado en grafos: Una nueva aproximación a la detección de plagio translingüe*

Graph-Based Similarity Analysis: A New Approach to Cross-Language Plagiarism Detection

Marc Franco-Salvador, Parth Gupta y **Paolo Rosso**
Natural Language Engineering Lab - ELiRF
Departamento de sistemas informáticos y computación
Universitat Politècnica de València
{mfranco,pgupta,prosso}@dsic.upv.es

Resumen: La variante translingüe de la detección de plagio automática trata de detectar plagio entre documentos en diferentes idiomas. En los últimos años se han propuesto una serie de aproximaciones que hacen uso de tesauros, modelos de alineamiento o diccionarios estadísticos para lidiar con la similitud a través de idiomas. En este trabajo proponemos una nueva aproximación a la detección de plagio translingüe que hace uso de una red semántica multilingüe para generar grafos de conocimiento, obteniendo un modelo de contexto para cada documento, de lo cual carecen otros métodos. Para evaluar nuestra propuesta, utilizamos las particiones español-inglés y alemán-inglés del corpus PAN-PC'11, comparando nuestros resultados con dos de las aproximaciones del estado del arte. Los resultados experimentales indican su potencial como alternativa para el análisis de similitud en detección de plagio translingüe.
Palabras clave: Detección de plagio translingüe, similitud textual, red semántica multilingüe, BabelNet, grafos de conocimiento.

Abstract: Cross-language variant of automatic plagiarism detection tries to detect plagiarism among documents across language pairs. In recent years a few approaches are proposed that use thesauri, alignment models or statistical dictionaries to deal with the similarity across languages. We propose a new approach to the cross-language plagiarism detection that makes use of a multilingual semantic network to generate knowledge graphs, obtaining a context model for each document which the other methods lack. To evaluate the proposed method, we use the Spanish-English and German-English partitions of the PAN-PC'11 corpus and compare our results with two state-of-the-art approaches. Experimental results indicate its potential to be a new alternative for similarity analysis in cross-language plagiarism detection.
Keywords: Cross-language plagiarism detection, textual similarity, multilingual semantic network, BabelNet, knowledge graphs.

1 Introducción

El plagio translingüe es definido como el uso no autorizado del contenido original de la obra de otros autores desde una fuente en otro idioma. Actualmente es un grave problema para los autores que además se ha complicado a causa de Internet. Éste pone a nuestra disposición, de forma gratuita y sencilla, una gran fuente de información y las herramientas necesarias para traducir y copiar contenidos originales. La investigación dentro del campo de la detección de plagio translingüe está justificada. En una encuesta realizada recientemente sobre las actitudes y prácticas de los estudiantes (Barrón-Cedeño, 2012), se pone de manifiesto que el plagio translingüe es un problema real: un 63.75 % de los estudiantes opina que copiar y traducir fragmentos de

* Agradecer a la Conselleria d'Educació, Formació i Ocupació de la Generalitat Valenciana por la financiación por parte del programa Gerónimo Forteza, sin el cual no hubiera sido posible llevar a cabo la investigación del primer autor que ha llevado a esta publicación. Este trabajo se ha hecho dentro del ámbito del *VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems* y como parte del proyecto de la Comisión Europea WIQ-EI IRSES (no. 269180).

ISSN 1135-5948

texto desde otros documentos y incluirlos en sus trabajos no es plagio.

La detección de plagio puede ser realizada de forma manual, pero dada la gran cantidad de obras publicadas, es muy complicado detectar los casos, aun más si la fuente del plagio proviene de otro idioma. Existen una serie de aproximaciones para llevar a cabo la detección de plagio translingüe de forma automática. Éstas hacen uso de tesauros, modelos de alineamiento o diccionarios estadísticos para detectar la similitud a nivel translingüe. *Cross-language character n-gram* (CL-CNG) (Mcnamee y Mayfield, 2004) es un modelo que se basa en la sintaxis de los documentos, haciendo uso de n-gramas, que ofrece un rendimiento notable para lenguajes con similitudes sintácticas. *Cross-language explicit semantic analysis* (CL-ESA) (Potthast et al., 2011a) es un modelo de análisis de semejanzas de colecciones relativas, lo que significa que un documento está representado por sus similitudes con una colección de documentos, las cuales son comparadas con un modelo de detección de similitud monolingüe. *Cross-language alignment-based similarity analysis* (CL-ASA) (Barrón-Cedeño et al., 2008; Pinto et al., 2009) se basa en la tecnología de máquinas de traducción estadística, la cual combina traducciones estadísticas, usando diccionarios estadísticos, y análisis de similitud. Los anteriores modelos han sido comparados (Potthast et al., 2011a), ofreciendo CL-ASA y CL-CNG el mejor desempeño. Por ese motivo, en nuestra evaluación comparamos nuestra aproximación con éstos.

Nuestra nueva aproximación, llamada *cross-language knowledge graphs analysis* (CL-KGA), proporciona un modelo de contexto de los documentos sospechosos y fuente a comparar. Para ello utiliza grafos de conocimiento generados por una red semántica multilingüe, los cuales expanden y relacionan los conceptos originales del texto. Así, la similitud entre documentos se mide mediante un método de análisis de similitud entre grafos.

Para la evaluación de los modelos utilizamos el corpus del PAN-PC'11 (Potthast et al., 2011b)[1], la competición internacional celebrada de forma anual en el marco de *Uncovering Plagiarism Authorship and Social Software Misuse* (PAN)[2], en la cual se presentan y ponen a prueba aproximaciones para la detección de plagio a nivel monolingüe y translingüe. Para nuestra evaluación utilizamos su partición de detección de plagio translingüe.

La estructura de la publicación es la siguiente: En la sección 2 explicamos en que consiste una red semántica multilingüe. En la sección 3 presentamos el modelo CL-KGA de análisis de similitud, y describimos los modelos con los que lo comparamos: CL-CNG y CL-ASA. En la sección 4 evaluamos nuestra aproximación utilizando los casos español-inglés (es-en) y alemán-inglés (de-en) de la tarea de detección de plagio externo del corpus del PAN-PC'11, comparando nuestros resultados con los obtenidos por los otros dos modelos. Finalmente, en la sección 5 presentamos nuestras conclusiones y trabajos futuros.

2 Red semántica multilingüe

Una red semántica multilingüe (RSM) consiste en un grafo dirigido y ponderado donde los nodos representan conceptos y nombres de entidades, y las aristas representan relaciones entre ellos. Además, cada uno de los nodos tiene una dimesión multilingüe con el conjunto de las lexicalizaciones del concepto en diferentes idiomas. En este trabajo, a partir de fragmentos de texto, vamos a utilizar una RSM para construir grafos de conocimiento, y compararlos entre ellos para detectar plagio translingüe.

La aproximación que describimos en la sección 3 es genérica y puede ser utilizada con cualquier RSM como ConceptNet[3] o EuroWordNet[4], pero para nuestros experimentos hemos elegido BabelNet (Navigli y Ponzetto, 2010). Ésta está formada por una base de conocimiento de gran tamaño, con el conjunto de lexicalizaciones de los conceptos disponibles en los siguientes idiomas: alemán, catalán, español, francés, inglés e italiano. Sus relaciones y conceptos provienen de WordNet, la mayor red semántica disponible, y de las entradas multilingüe de la Wikipedia[5], así BabelNet combina información lexicográfica con conocimiento enciclopédico. La lista de conceptos está formada por to-

[1] URL: http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-11.html

[2] URL: http://pan.webis.de/

[3] URL: http://csc.media.mit.edu/conceptnet/

[4] URL: http://www.illc.uva.nl/EuroWordNet/

[5] URL: http://www.wikipedia.org/

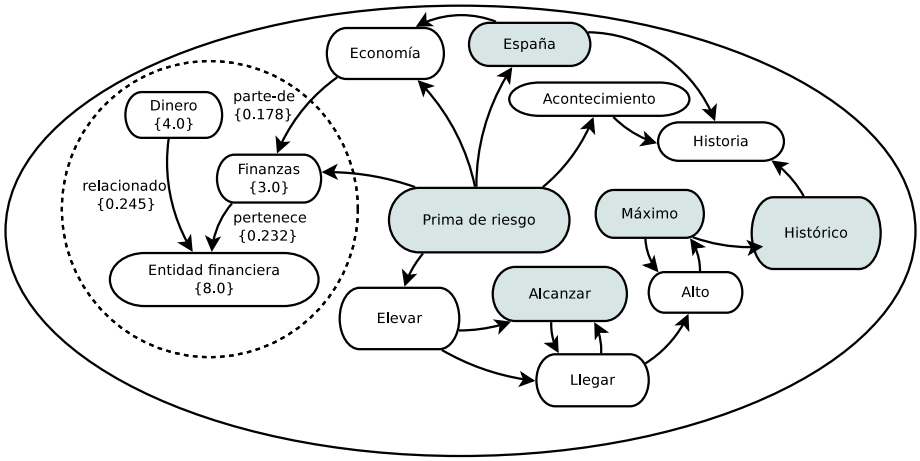

Figura 1: Ejemplo simplificado, sin dimensión multilingüe, del grafo de conocimiento de la frase "la prima de riesgo española alcanza máximos históricos" (las relaciones y los pesos de nodos y aristas se incluyen sólo dentro de la línea discontinua).

dos los significados de palabra en WordNet y de las entradas etiquetadas de la Wikipedia, por otro lado las relaciones entre conceptos las forman los punteros semánticos entre conceptos en WordNet y los enlaces entre entradas en la Wikipedia. Por último, las lexicalizaciones multilingüe se obtienen a partir de las entradas en diferentes idiomas de la Wikipedia.

El API de BabelNet permite utilizar esta RSM, entre otros usos, como diccionario estadístico, traductor, para desambiguación del sentido de las palabras (Navigli y Ponzetto, 2012), y para construir grafos de conocimiento.

2.1 Grafos de conocimiento

Un grafo de conocimiento consiste en un grafo dirigido y ponderado, generado a partir de un conjunto de palabras como las de una frase, que contiene los conceptos originales expandidos y relacionados entre ellos, dando lugar a un "modelo de contexto" de la frase o conjunto de palabras original. El peso de un concepto es el número de relaciones salientes, mientras que el peso de una relación está en función del número de relaciones que conectan con sus conceptos origen y destino[6]. Para comprender mejor en que consiste un grafo de conocimiento, vamos a poner un ejemplo. Supongamos que tenemos la frase "La prima de riesgo española alcanza máximos históricos". Los conceptos de la frase son $C = \{$prima de riesgo, españa, alcanzar, máximo, histórico$\}$. En BabelNet podemos construir un grafo de conocimiento g a partir de C, el cual contendrá un nuevo listado de conceptos $C_g = (C \cup C')$, siendo la lista de conceptos expandidos $C' = \{$economía, finanzas, historia...$\}$. Además, entre los conceptos de C_g, aparecerán una serie de conexiones R que los relacionan, $R \in \{$relacionado, parte de, pertenece, equivalente, opuesto...$\}$. Además, como se ha comentado anteriormente, cada uno de los conceptos y relaciones del grafo tiene una dimensión multilingüe, por lo que dos fragmentos de texto similares en diferentes lenguajes, deberían tener unos grafos también similares. En la fig. 1 podemos ver el contenido del grafo g en español.

3 Modelos de detección de plagio translingüe

De acuerdo a su paradigma de resolución, existen diferentes categorías de modelos de análisis de similitud que pueden ser utilizados para detección de plagio translingüe: (i) modelos que hacen uso de diccionarios, *gazetteers*, reglas o tesauro lingüísticos para realizar las traducciones de los conceptos desde un lenguaje origen L a uno destino L'. En esta categoría tenemos por ejemplo CL-VSM (Steinberger, Pouliquen, y Ignat, 2004), que utiliza modelos de espacio vectorial (Stein y Anderka, 2009) o CL-CTS (Gupta, Barrón-Cedeño, y Rosso, 2012), que usa el tesauro conceptual Eurovoc para las traducciones. (ii) Modelos que se basan en la sintaxis del documento y su estructura para comparar los documentos. El CL-CNG (Mcna-

[6]En (Navigli y Ponzetto, 2010) tenemos la fórmula de estimación de pesos de relaciones.

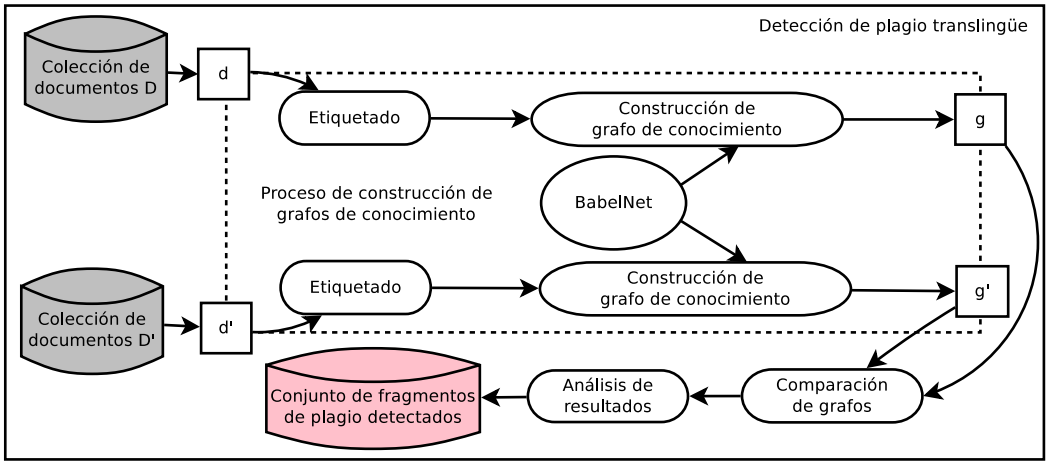

Figura 2: Proceso de detección de plagio translingüe utilizando grafos de conocimiento.

mee y Mayfield, 2004) está incluido en esta categoría. (iii) Modelos que utilizan corpus comparables, como CL-ESA (Potthast et al., 2011a). Éste utiliza corpus alineados por tema y idioma, como la enciclopedia de la Wikipedia, y analiza la similitud con un modelo monolingüe como los modelos de espacio vectorial. (iv) Los modelos basados en un corpus paralelo alinean los corpus en diferentes idiomas a nivel de documento y palabra. Los modelos CL-ASA (Barrón-Cedeño et al., 2008; Barrón-Cedeño, 2012), CL-LSI (Dumais et al., 1997) y CL-KCCA (Vinokourov, Shawe-Taylor, y Cristianini, 2003), quedan dentro de esta categoría. Además, existen modelos que pueden utilizar combinaciones de las categorías anteriores, para lo cual nuestra aproximación CL-KGA es el ejemplo perfecto, siendo la RSM BabelNet la unión de (i) un tesauro (WordNet) y de (iii) un corpus comparable (Wikipedia).

Dejando de lado aproximaciones como CL-LSI y CL-KCCA que ofrecen un elevado rendimiento a un alto coste computacional, existen trabajos (Potthast et al., 2011a; Gupta, Barrón-Cedeño, y Rosso, 2012) que han comparado algunos de los anteriores modelos: CL-ASA, CL-ESA, CL-CNG y CL-CTS. En sus resultados se refleja como CL-CNG es un buen *baseline* para tomar como partida en la detección de plagio translingüe, y CL-ASA ofrece en promedio los mejores resultados. Por esa razón hemos elegido CL-CNG y CL-ASA como las aproximaciones a comparar, en la evaluación, con nuestro modelo.

A continuación vamos a describir nuestra nueva propuesta, CL-KGA, y los dos modelos con los que la comparamos.

3.1 Análisis de similitud basado en grafos de conocimiento

La aproximación que proponemos en esta publicación, CL-KGA, utiliza grafos de conocimiento generados a partir de una RSM para obtener una similitud entre dos textos, como por ejemplo documentos o fragmentos de texto. Dado un conjunto de documentos D en un lenguaje L_1 y un conjunto de documentos D' en un lenguaje L_2, para comparar dos documentos $d \in D$ y $d' \in D'$, en primer lugar debemos realizar un procesado previo del texto para extraer y etiquetar morfológicamente sus conceptos. Además, es conveniente lematizar el texto. Para todas estas tareas, en nuestra investigación hemos hecho uso de la herramienta TreeTagger [7]. Una vez procesado el texto, podemos construir, utilizando la RSM BabelNet, los grafos de conocimiento g y g' a partir de los documentos d y d'. En la fig. 2 podemos ver un esquema del proceso de detección de plagio translingüe utilizando grafos de conocimiento. Para obtener una similitud $S(g, g')$ entre g y g', tomando como base la aproximación de comparación flexible de grafos conceptuales[8] (Montes y Gómez et al., 2001), hemos propuesto la ecuación 1 para trabajar con grafos de conocimiento.

$$S(g, g') = S_c(g, g') * (a + b * S_r(g, g')) \qquad (1)$$

[7] URL: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[8] Un grafo conceptual es un grafo finito dirigido bipartido con dos clases de nodos: conceptos y relaciones (Sowa, 1984; Sowa, 1999).

$$S_c(g,g') = \frac{\left(2 * \sum_{c \in g_u} w(c)\right)}{\left(\sum_{c \in g} w(c) + \sum_{c \in g'} w(c)\right)} \quad (2)$$

$$S_r(g,g') = \frac{\left(2 * \sum_{r \in N(c,g_u)} w(r)\right)}{\left(\sum_{r \in N(c,g)} w(r) + \sum_{r \in N(c,g')} w(r)\right)} \quad (3)$$

donde $S_c(g,g')$ es la similitud entre los conceptos de los grafos, $S_r(g,g')$ es la similitud entre las relaciones, g_u es el grafo resultante de la intersección de g y g', c es un concepto, r es una relación, $w(c)$ y $w(r)$ son sus pesos, y $N(c,g_i)$ es el conjunto de relaciones conectadas al concepto c en el grafo g_i. Las variables a y b se utilizan en la ecuación 1 para dar la apropiada relevancia a los conceptos y relaciones, ya que sus pesos no se calculan del mismo modo y, por tanto, valores de similitud iguales no tienen porqué tener el mismo significado. Además, para la resolución de determinados problemas, no son igual de relevantes conceptos que relaciones, por este motivo se suele utilizar la regla $a + b = 1$, y se toman a y b como porcentajes de relevancia. En la sección 4 analizaremos cuales son los porcentajes de relevancia adecuados para conceptos y relaciones en detección de plagio translingüe utilizando BabelNet.

Es importante señalar que después de la intersección $g_u = (g \cap g')$, los pesos del grafo g_u tendrán que ser recalculados. El cálculo del peso de un concepto es trivial, pues es el número de relaciones salientes. Recalcular el peso de las relaciones requiere de coste cúbico siguiendo su proceso de creación en BabelNet, ya que para cada relación sería necesario recorrer todos los conceptos dos veces[9]. Por ese motivo, en la ecuación 4 proponemos un algoritmo genérico de reestimación del peso $w(r,c,g_u)$, siendo r una relación saliente de un concepto c en el grafo de intersección g_u. El nuevo peso se calcula en función del antiguo y del nuevo valor del peso de c en los grafos g, g' y g_u,

$$w(r,c,g_u) = \frac{w(c,g) * d(c,g,g_u) + w(c,g') * d(c,g',g_u)}{2} \quad (4)$$

[9]Donde $t(n,m) \in O(n^2 * m)$, siendo n el número de conceptos y m el número de relaciones entre ellos.

$$d(x,g_1,g_2) = \frac{|R(g_1,x)|}{|R(g_2,x)|} \quad (5)$$

donde $w(c,g_i)$ es el peso del concepto c en el grafo g_i, y $R(g_i,x)$ es el conjunto de relaciones salientes del concepto x en el grafo g_i.

3.2 Análisis de silimitud basado en n-gramas de caracteres

El modelo CL-CNG, *cross-language character n-gram*, ha demostrado ofrecer un rendimiento elevado para lenguajes europeos con similitudes sintácticas y hace uso de n-gramas a nivel de caracteres para comparar los documentos en diferentes idiomas. En este modelo se utilizan normalmente trigramas de caracteres (CL-C3G) (Potthast et al., 2011a).

Dado un documento fuente d en un lenguaje L_1 y un documento sospechoso d' en un lenguaje L_2, la similitud $S(d,d')$ entre los dos documentos se mide como se muestra en la ecuación 6:

$$S(d,d') = \frac{\vec{d}.\vec{d'}}{|d|.|d'|} \quad (6)$$

donde $\vec{d}$ y $\vec{d'}$ son las proyecciones vectoriales de d y d' en un espacio de n-gramas de carácter.

3.3 Análisis de similitud basado en alineamiento

El modelo CL-ASA mide la similitud entre dos documentos d y d', en dos idiomas diferentes L_1 y L_2, alineandolos a nivel de palabra, determinando la probabilidad de que un documento d' sea una una traducción del documento d. La similitud $S(d,d')$ se calcula haciendo uso de la ecuación 7:

$$S(d,d') = l(d,d') * t(d|d') \quad (7)$$

donde $l(d,d')$ es el factor de longitud definido en (Pouliquen, Steinberger, y Ignat, 2003) y $t(d|d')$ es el modelo de traducción definido en la ecuación 8:

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x,y) \quad (8)$$

donde $p(x,y)$ es la probabilidad de que una palabra x en el lenguaje L_1 sea una traducción de la palabra y del lenguaje L_2. Dichas probabilidades de traducción pueden obtenerse mediante un diccionario estadístico. Para nuestros experimentos se ha entrenado

un diccionario estadístico alemán-inglés y español-inglés haciendo uso del modelo de alineamiento de palabras IBM M1 (Brown et al., 1993; Och y Ney, 2003), sobre el corpus paralelo multilingüe JRC-Acquis (Steinberger et al., 2006), además de probar también el diccionario estadístico de la RSM BabelNet.

4 Experimentos y evaluación

En esta sección vamos a evaluar el rendimiento de nuestro modelo CL-KGA, para la tarea de detección de plagio translingüe es-en y de-en, utilizando la RSM BabelNet como base de conocimiento, frente a los modelos estado del arte CL-ASA y CL-C3G. Para el modelo CL-ASA realizaremos las pruebas con dos diccionarios estadísticos diferentes: un diccionario entrenado con el modelo de alineamiento IBM M1, y el diccionario estadístico de BabelNet (BN_{dict}), que ya ha demostrado anteriormente ofrecer un buen rendimiento para la tarea de detección de plagio translingüe (Franco-Salvador, Gupta, y Rosso, 2012). Además, previamente a la comparación de los modelos, vamos a realizar unos experimentos para determinar cual es la relación de porcentajes adecuados para los valores de relevancia de conceptos y relaciones en el CL-KGA.

4.1 Corpus y definición de la tarea

Del corpus PAN-PC'11, tomamos las particiones es-en y de-en para su tarea de detección de plagio externo: dado un conjunto de documentos fuente D en el lenguaje L_1 y un conjunto de documentos sospechosos D' en el lenguaje L_2, la tarea es determinar los fragmentos concretos de los documentos fuente que están presentes en los sospechosos. Para ello utilizamos una ventana deslizante de cinco oraciones de longitud sobre pares de documentos (d,d'), $d \in D$ y $d' \in D'$, y detectamos plagio translingüe sobre ellos con los modelos comentados anteriormente. En la Tabla 1 podemos ver las estadísticas de los documentos utilizados para la evaluación.

Documentos es-en		Documentos de-en	
Sospechosos	304	Sospechosos	251
Fuentes	202	Fuentes	348
Casos de plagio {es,de}-en			
Traducción automática	5.142		
Traducción automática + corrección manual	433		

Cuadro 1: Estadísticas de la tarea de detección de plagio externo del corpus PAN-PC'11

4.2 Unidades de medida

Para medir la calidad de los resultados vamos a tomar las medidas utilizadas en la competición del PAN: *recall* (rec.) y *precision* (prec.) a nivel de carácter, además de *granularity* (gran.), la cual tiene en cuenta el hecho de que en ocasiones los detectores solapan o deportan multiples detecciones para un mismo caso de plagio. Las tres medidas son combinadas con el objetivo de obtener una medida global de la detección de plagio, el *plagdet*:

$$plagdet(S,R) = \frac{F_1}{log_2(1 + granularity(S,R))}$$

Donde S es el conjunto de casos de plagio del corpus, R es el conjunto de casos de plagio reportados por el detector, y F_1 es la media armónica de *precision* y *recall* ponderadas equitativamente. [10]

4.3 Experimento 1

En primer lugar vamos a comparar el rendimiento del modelo CL-KGA utilizando la RSM BabelNet, según sus valores de relevancia para conceptos y relaciones. Para ello hemos diseñado un experimento, midiendo solamente el *plagdet*, utilizando una porción aleatoria del 20 % del corpus PAN-PC'11, tanto para es-en como de-en, en el que probaremos los siguientes porcentajes de relevancia para conceptos (c) y relaciones (r): $(c,r) \in \{(100,0),(80,20),(50,50),(20,80),(0,100)\}$.

% (c,r)	Plagdet(es-en)	Plagdet(de-en)
(100,0)	0.617	**0.636**
(80,20)	0.616	0.6247
(50,50)	**0.655**	0.620
(20,80)	0.642	0.581
(0,100)	0.612	0.522

Cuadro 2: Relevancia de conceptos y relaciones en el modelo CL-KGA.

En vista de los resultados de la tabla 2, podemos deducir que las relaciones utilizando la RSM BabelNet son prácticamente igual de importantes que los conceptos para es-en, mientras que para de-en tienen poca o ninguna importancia. La diferencia puede estar producida por unos conceptos muy conectados en grafos es-en, mientras que en de-en podemos estar ante un elevado número de conceptos parte de un grafo menos conexo. Para nuestros siguientes experimentos tomaremos las mejores configuraciones de ambas particiones.

[10] Una descripcción más detallada de las medidas se puede encontrar en (Potthast et al., 2010).

4.4 Experimento 2

En este experimento vamos a comparar CL-KGA, con los modelos descritos anteriormente, para las particiones completas es-en y de-en del corpus del PAN-PC'11.

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.594**	**0.518**	**0.706**	**1.008**
CL-ASA$_{BNdict}$	0.567	0.499	0.662	1.015
CL-ASA$_{IBMM1}$	0.517	0.448	0.689	1.071
CL-C3G	0.170	0.128	0.617	1.372

Cuadro 3: Resultados de la detección de plagio translingüe es-en

En la tabla 3 podemos observar como CL-KGA ha superado en todos los valores al resto de modelos para la detección de plagio translingüe es-en. El modelo CL-ASA que más se le ha aproximado -utilizando el diccionario del propio BabelNet- tiene un *plagdet* un 4.7 % inferior. Además, aparte de observar el aumento de los valores de *precision* y *recall*, es importante señalar que se ha alcanzado un valor de *granularity* muy próximo a 1, lo cual es el mejor valor posible, e indica que no existen solapamientos en la detección interpretando una sección de plagio como varias, o viceversa.

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.514**	**0.443**	**0.631**	**1.018**
CL-ASA$_{IBMM1}$	0.406	0.344	0.604	1.113
CL-ASA$_{BNdict}$	0.289	0.222	0.595	1.171
CL-C3G	0.078	0.047	0.330	1.089

Cuadro 4: Resultados de la detección de plagio translingüe de-en

En la tabla 4 vemos también unos buenos resultados para de-en en nuestro modelo. CL-KGA ha superado al CL-ASA$_{IBMM1}$, el más cercano, en un valor de *plagdet* del 26.6 %, lo cual supone una excelente mejora respecto al estado del arte actual. Los otros valores también han mejorado, destacando un incremento del *recall* de un 28 %, lo cual indica un considerable aumento en el número de detecciones positivas. En esta ocasión el diccionario de BabelNet no se ha comportado tan bien como para es-en[11].

En vista de los resultados anteriores, podemos afirmar cómo hacer uso de grafos de conocimiento es una buena alternativa para la detección de plagio translingüe.

[11]Lo cual viene justificado en (Franco-Salvador, Gupta, y Rosso, 2012) como consecuencia del procesamiento previo de las palabras en alemán al construir BabelNet

5 Conclusiones y trabajos futuros

En este trabajo hemos presentado un nuevo modelo para el análisis de similitud a nivel translingüe, el CL-KGA, que hace uso de una RSM para construir grafos de conocimiento a modo de modelos de contexto de documentos. El modelo propuesto ha demostrado ofrecer un rendimiento superior a otros modelos estado del arte como CL-ASA y CL-CNG, evaluados sobre la partición translingüe del corpus PAN-PC'11.

En futuras investigaciones se seguirá investigando en el campo de la detección de plagio translingüe para extender nuestro modelo con otras RSM que nos proporcionen una mayor variedad de lenguajes compatibles, además de investigar el potencial de nuestro nuevo modelo para análisis de similitud a nivel monolingüe.

Bibliografía

Barrón-Cedeño, Alberto. 2012. *On the mono- and cross-language detection of text re-use and plagiarism.* Ph.D. thesis, Universitat Politènica de València.

Barrón-Cedeño, Alberto, Paolo Rosso, David Pinto, y Alfons Juan. 2008. On cross-lingual plagiarism analysis using a statistical model. En *Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, PAN'08.

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

Dumais, S. T., T. A. Letsche, M. L. Littman, y T. K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. En *Proc. AAAI-97 spring symposium series: Cross-language text and speech retrieval*, páginas 18–24. Hull & D. Oard (Eds.).

Franco-Salvador, Marc, Parth Gupta, y Paolo Rosso. 2012. Cross-language plagiarism detection using BabelNet's statistical dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación*, 16(4):383–390.

Gupta, Parth, Alberto Barrón-Cedeño, y Paolo Rosso. 2012. Cross-language high

similarity search using a conceptual thesaurus. En *Proc. 3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics*. CLEF 2012.

Mcnamee, Paul y James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1):73–97.

Montes y Gómez, Manuel, Alexander F. Gelbukh, Aurelio López-López, y Ricardo A. Baeza-Yates. 2001. Flexible comparison of conceptual graphs. En *Proc. DEXA*, páginas 102–111.

Navigli, Roberto y Simone Paolo Ponzetto. 2010. Babelnet: building a very large multilingual semantic network. En *Proc. of the 48th annual meeting of the association for computational linguistics*, ACL '10, páginas 216–225, Stroudsburg, PA, USA.

Navigli, Roberto y Simone Paolo Ponzetto. 2012. Multilingual wsd with just a few lines of code: The babelnet api. En *Proc. 50th annual meeting of the association for Computational Linguistics*.

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

Pinto, D., J. Civera, A. Barrón-Cedeño, A. Juan, y P. Rosso. 2009. A statistical approach to crosslingual natural language tasks. *Journal of algorithms*, 64(1):51–60.

Potthast, M., A. Barrón-Cedeño, B. Stein, y P. Rosso. 2010. An evaluation framework for plagiarism detection. En *Proc. of the 23rd Int. Conf. on Computational Linguistics*, COLING-2010, páginas 997–1005, Beijing, China.

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, y Paolo Rosso. 2011a. Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1):45–62.

Potthast, Martin, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, y Paolo Rosso. 2011b. Overview of the 3rd int. competition on plagiarism detection. En *CLEF (Notebook Papers/Labs/Workshop)*.

Pouliquen, B., R. Steinberger, y C. Ignat. 2003. Automatic linking of similar texts across languages. En *Proc. Recent Advances in Natural Language Processing III*, páginas 307–316. RANLP'03.

Sowa, J. F. 1984. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman.

Sowa, J. F. 1999. *Knowledge representation: logical, philosophical and computational foundations*. Brooks/Cole Publishing Co.

Stein, B. y M. Anderka. 2009. Collection-relative representations: A unifying view to retrieval models. En *Proc. 20th Int. Conf. on database and expert systems applications*, DEXA'09, páginas 383–387. A. M. Tjoa & R. R. Wagner (Eds.).

Steinberger, R., B. Pouliquen, y C. Ignat. 2004. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. En *Proc. 4th Slovenian language technology conference*, IS'2004. Information Society.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, y D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. En *Proc. 5th Int. Conf. on language resources and evaluation*. LREC'2006.

Vinokourov, A., J. Shawe-Taylor, y N. Cristianini. 2003. Inferring a semantic representation of text via cross-language correlation analysis. En *Proc. NIPS-02: Advances in neural information processing systems*, páginas 1473–1480. S. Becker, S. Thrun, & K. Obermayer (Eds.).

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 29-35 recibido 29-11-12 revisado 11-02-13 aceptado 19-02-13

WeFeelFine as Resource for Unsupervised Polarity Classification *

WeFeelFine *como recurso para clasificación de la polaridad no supervisada*

Arturo Montejo-Ráez
Departamento de Informática
Universidad de Jaén
Las Lagunillas s/n, Jaén - 23071
amontejo@ujaen.es

Resumen: En este trabajo se presenta una solución no supervisada al problema de la clasificación de la polaridad en micro-blogs. La propuesta no sólo no necesita de entrenamiento, sino que se construye a partir de las propias publicaciones de millones de usuarios en la web. Los resultados muestran la efectividad de esta propuesta, abriendo la puerta a una nueva forma de afrontar el análisis de sentimientos en micro-blogs.
Palabras clave: Análisis de emociones, clasificacin de la polaridad, Twitter, micro-blogging

Abstract: This papers shows the results obtained by a non supervised method in the task of sentiment polarity detection on micro-blogs. This method does not need of training, but it also is self-constructed from millions of publications on the web. The results show the effectiveness of the proposal, openining a new way of facing sentiment analysis in micro-blogs.
Keywords: Sentiment Analysis, polarity classification, Twitter, micro-blogging

1 Introduction

Twitter has become a key service in web-based communication. Its growth rate in terms of content and users has focused the attention of many other services, companies, communities and, of course, scientists. The amount of messages from Twitter users that floods the Internet turns this service into a very useful source of information about the topics on which people focus their interests. Nowadays, proper filtering, extraction and understanding of this overwhelming stream of text is the main subject of study for Natural Language Processing research. Besides, Sentiment Analysis on tweets is one of the most active topic of research taking place (Asiaee T. et al., 2012).

This work presents a novel unsupervised approach to tackle Sentiment Analysis on Twitter by associating to each tweet a list of "feelings" obtained by means of search over a corpus of micro-blogging publications gathered by the WeFeelFine project (Kamvar and Harris, 2011). In this way, tweets are caracterized by the most similar feelings associated by performing a retrieval over the sentences related to each tweet in WeFeelFine data. Then, a final measure of polarity is computed according to the list of feelings obtained. Our results show that this approach outperforms many state-of-the-art unsupervised solutions and that, due to its simplicity, may open a new way of understanding sentiment analysis of micro-blogs by using micro-blogs themselves.

The paper is organized as follows: first a brief introduction to the polarity classification problem is given. Then, the WeFeelFine project is described, with pointers to related research based on its data. Next, our approach is unveiled, describing the prepara-

* This work is partly funded by the European Comission, under the VII Framework Program (FP7 - 2007-2013), within the FIRST project (FP7-287607) and by the Spanish Government, within the TEXT-COOL project (TIN2009-13391-C04-02).

ISSN 1135-5948

tion of the system and its components. Experimental setup and results follows, to end with final conclusions and reflections on future lines to explore.

2 The polarity classification problem

Sentiment Analysis is one of the most active research areas in Natural Language Processing nowadays (Pang and Lee, 2008), with special interest in the classification of texts into positive, negative or neutral. This latter task is known as the Polarity Classification problem, and attracts the attention of the research community and also companies, politicians or personalities, due to the relevance in the study of reputation of products, people or any other item based on opinions of users in the web.

Polarity Classification is solved using both supervised and non-supervised approaches. Supervised strategies have reported the best results since early works (Pang, Lee, and Vaithyanathan, 2002) and it is still the choice for many solutions, from Information Theory based features (with SVM classifier) (Lin et al., 2012) to more complex learned rules (Tan et al., 2012). Unsupervised approaches have relied mainly on the use of lexicons where words are associated with polarity scores (Boldrini et al., 2010), although more avanced solutions using intensive lexical analysis are proposed (Chen et al., 2012). In any case, a value of 70% for F-score seems to be, still, far from these methods.

Turney (Turney, 2002), instead of manually generating a corpus of emotional words, used Altavista search engine to compute the *Semantic Orientation* (SO) of a phrase according to the proximity of well known emotional words (like *excellent* or *poor*) in millions of web pages. Unfortunately, the complexity of modern ranking algorithms used by main search engines has opened a gap between word statistics in the web and the actual results obtained, so the validity of such approach can be argued. Anyhow, the proposal of using the implicit knowledge in millions of texts has been an inspiration in our work.

Sentiment Analysis has been specially focused on Twitter, due to its relevance as social media (Martínez-Cámara et al., In press) and despite the inherent challenges of subjective micro-blogging, like irony (Reyes, Rosso, and Buscaldi, 2012). Our experiments are performed on tweets from this popular service, as explained later.

3 WeFeelFine

Since 2005, the website WeFeelFine[1] has been harvesting from social media millions of sentences containing "I feel" or "I am feeling" expressions, creating a huge database of sentences related to feelings or emotions (Kamvar and Harris, 2011). Although the main goal of the project is to serve as a monitor of human state at a global level, we found that the collected data could be useful in sentiment analysis. The authors of the website, indeed, perform this kind of analysis in order to produce semantic related data. Thanks to its API, it is possible to download a bunch of sentences (up to a limit of 1,500 imposed by the site) per each of the defined feelings. The current list of feelings stored contains 2,178 different feelings, although the 200 most frequent ones hold 70% from a total of almost 2 millions sentences. We can see the feelings with higher presence in the database along with the percentage over the total of sentences in Figure 1.

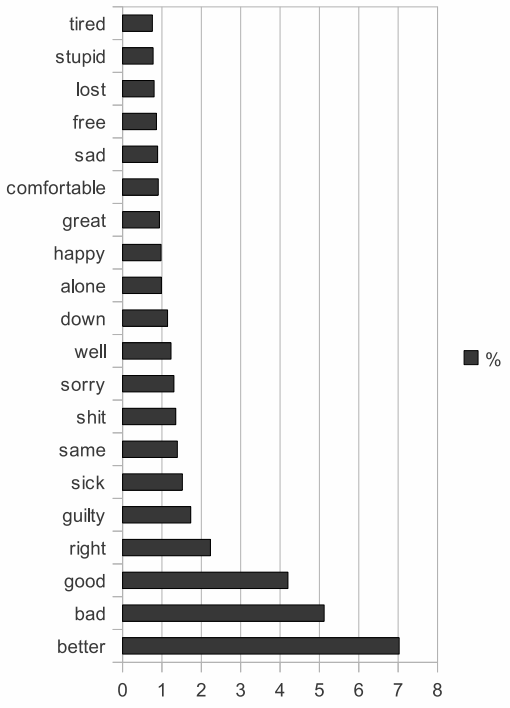

Figure 1: 20 most frequent feelings in WeFeelFine

[1] http://wefeelfine.org

WeFeelFine offers interactive tools to explore the data and relates the feelings with profile information like gender or age, as author details are also extracted from the web (see snapshots at Figure 2 and Figure 3. In our experiments, this information related to author profiles has been discarded, as such data had to be extracted when using other sources[2], although it could represent informative attributes (Schler et al., 2006). WeFeelFine is a very interesting project and its continuous crawling of data could represent a valuable resource in sentiment analysis, as considered by previous studies (Agarwal et al., 2008), where a bag of sentiment words is created using WeFeelFine list of feelings and augmented with synonyms and antonyms from Thesaurus[3].

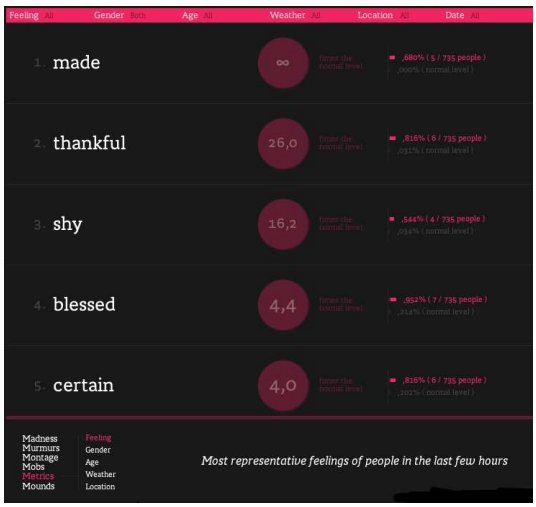

Figure 2: State monitor of feelings in We Feel Fine applet

4 *System architecture*

Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) is an interesting alternative to document modelling. Instead of generating a vector of word statistics, it produces a vector of related Wikipedia articles, being the similarity of the original document to each article the weight for that dimension. Therefore, the document is used as a query against a search engine over the whole Wikipedia, returning a list of articles ranked by their similarity. The novelty here is that the index is generated from data gathered from the blogosphere in a continous flow. So,

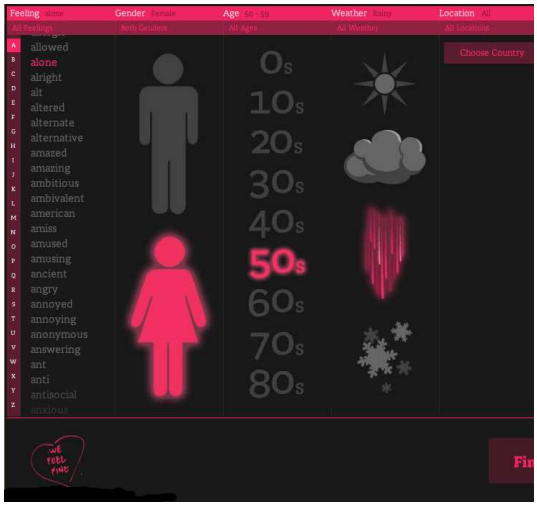

Figure 3: Exploring sentiments by gender, weather, age, country...

we can consider that the proposed approach is a mixed model of Gabrilovich and Turney (Turney, 2002) models.

Our approach is similar in that we represent each tweet by a vector of feelings. A manual labelling of the polarity (with +1 or -1 values) of those 200 feelings (which took just few minutes), is used to compute the final semantic orientation of the tweet by associating to it the list of most related feelings according to a search over the collection built from WeFeelFine data downloaded usign its API. Thus, the system can be splited into two different modules: the indexing module and the classification module.

4.1 Index generation

By means of the WeFeelFine API, we have generated a collection of 200 documents, corresponding to the most frequent feelings in this web, according to the state of its database on 10th October of 2012. Thus, for each feeling, there exists a document containing 1,500 sentences. These documents are indexed, as visualized in the whole process given in Figure 4. For indexing and retrieval the Lucene[4] engine has been used with default configuration (version 3.6.1). For both, sentences to feelings and testing tweets, hashtags and mentions ('#' and '@' strings) have been removed, along with URLs.

4.2 Search

Once the index is generated, we can take a tweet and ask the search engine with it

[2]PAN task on Author Profiling: `http://pan.webis.de`

[3]http://thesaurus.com

[4]http://lucene.apache.org/

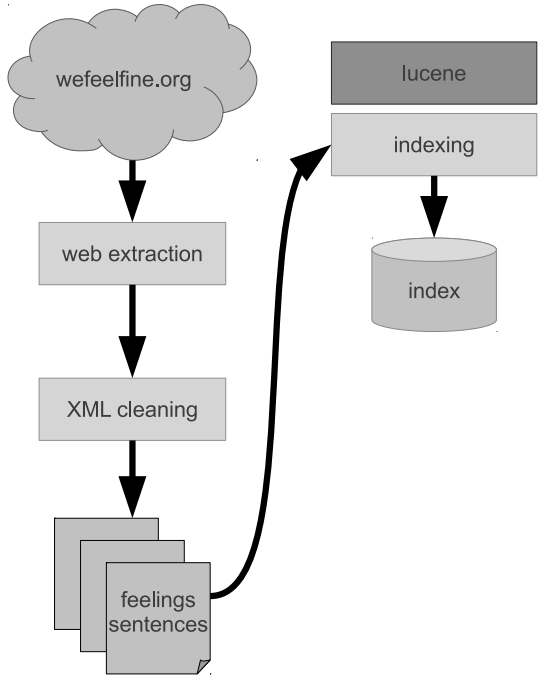

Figure 4: Indexing process

as a query, retrieving the closest feelings, as shown in Figure 5. Finally, from the ranked list of feelings, the final polarity of the tweet is computed based on the polarity value manually assigned to each feelilng. The only parameter that has to be specified is the number of results to be used before averaging, which determines the number of feelings to be taken into account when computing the polarity according to one of the two possible equations defined below.

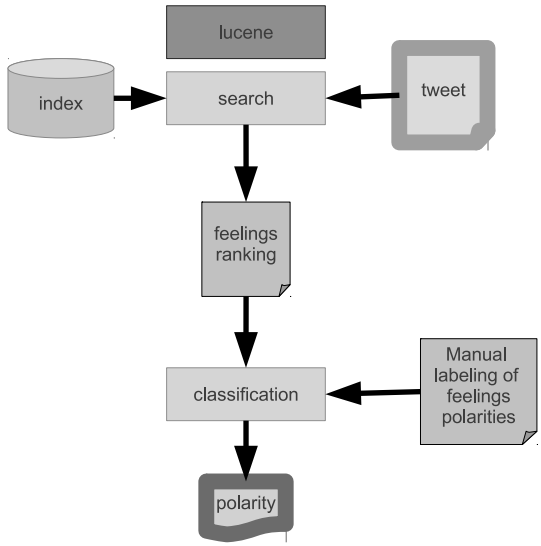

Figure 5: Classification process

As we will see later, the *Ranking Status Value* (RSV) computed by Lucene can also be useful when computing the final polarity. By using this ranking value (which reflects a distance between the tweet and a feeling), we can perform a weighted summatory. Therefore, two possible equations are proposed:

$$p(t) = \frac{1}{|R|} \sum_{r \in R} l_r \tag{1}$$

where
$p(t)$ is the polarity of tweet t
R is the list of retrieved feelings
l_r is the polarity label of feeling r

In the case of considering the RSV, the formula is very similar, but with RSV_r weighting the polarity.

$$p(t) = \frac{1}{|R|} \sum_{r \in R} RSV_r \cdot l_r \tag{2}$$

For example, the tweet "*The Nike Training Club beta iPhone app looks very interesting*" returns the top ten results with the given RSVs shown in Table 1. The table shows how feelings, according to the sentences that made up the representant document, are close to the tweet as query. The scoring formula of Lucene combines cosine and boolean similarities, but in summary is fully based on TF.IDF values, with other factors like document length. We have not changed the default *practical scoring formula* of the engine, altough an adjustment to Twitter nature is foreseen.

Rank	RSV	feeling	polarity
1	0.05166112	cool	+1
2	0.040141936	dumb	-1
3	0.03140159	lucky	+1
4	0.030341815	awesome	+1
5	0.029633064	fine	+1
6	0.029432593	used	-1
7	0.028811168	low	-1
8	0.027871676	missing	-1
9	0.027096074	complete	+1
10	0.026837287	proud	+1

Table 1: Resulting list from a Lucene search

5 *Experiments and results*

To prove this approach, we have taken the Emoticon data set from Stanford University (Go, Bhayani, and Huang, 2009). To enable the comparison of results with other approaches, only the test set is considered. It contains 177 negative tweets and 182 positive

tweets, manually labelled. Therefore, a total of 359 queries have been launched against Lucene, generating a list of results (feelings) for every tweet.

In order to explore the effect in the number of results considered, a range from 1 to 100 top results were taken into account, obtaining corresponding values of precision, recall, F-score and accuracy as performance scores.

5.1 Plain averaging

Results obtained applying Equation 1 are given in Table 2. As can be seen, an impressive F-score of 70.03% is reached when 55 top results are used as feelings to average the final polarity, although the performance of other values near 55 top results are small. Graphically represented in Figure 6, the effect of the number of feelings on the performance is clear. It is visible a constant increase in performance up to 20-30 results. We believe that this is due to the fact that semantic charge of a tweet (even if it is composed by few words) needs of a fined-grain representation under the shape of a list of feelings. Thus, when more feelings are considered, the tweet is modelled more properly. But also, a bit of performance drops beyond that number of 55 top feelings, this can be due to the integration of noise for larger list of results.

# results	accuracy	precision	recall	f-score
1	0.607242	0.677321	0.606584	0.640004
5	0.571031	0.643939	0.569551	0.604465
10	0.584958	0.680557	0.583132	0.628089
15	0.618384	0.700917	0.617030	0.656304
20	0.657382	0.717188	0.655724	0.685080
25	0.665738	**0.725959**	0.664509	0.693876
30	0.679666	0.720909	0.678323	0.698968
35	0.676880	0.723591	0.675421	0.698677
40	0.665738	0.715360	0.664199	0.688831
45	0.662953	0.711189	0.661296	0.685336
50	0.662953	0.714360	0.661374	0.686847
55	**0.682451**	0.721333	**0.680605**	**0.700377**
60	0.660167	0.695295	0.658471	0.676382
65	0.660167	0.699042	0.658704	0.678274
70	0.654596	0.699104	0.653287	0.675419
75	0.665738	0.703054	0.664432	0.683198
80	0.662953	0.698754	0.661995	0.679878
85	0.674095	0.699647	0.673061	0.686097
90	0.674095	0.704569	0.672984	0.688414
95	0.657382	0.697619	0.656267	0.676312
100	0.674095	0.708827	0.672984	0.690441

Table 2: Results obtained with plain averaging

5.2 RSV weighting

These configuration does not take into account the RSV when retrieving the feelings closest to the given tweet. The RSV is a useful measurement of the similarity between

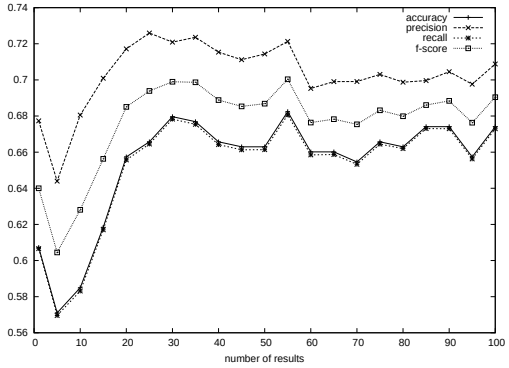

Figure 6: Effect of the number of results on performance for plain averaging

the tweet and the feeling. Thus, using it as weighting value in a linear combination to obtain the final polarity could lead to better performance values. Table 3 shows the results obtained when Equation 2 is applied. Our intuition is confirmed, with a 73% in F-score reached (again, with 55 results). Again this time, as can be observed in Figure 7, the usage of more results leads to better performance scores, with a constant increase up to 20-30 results. Both approaches are visually compared in Figure 8. From this graph we can conclude that applying the RSV as weight on the polarity of associated feelings leads to a better performances independently from the number of results considered.

# results	accuracy	precision	recall	f-score
1	0.607242	0.677321	0.606584	0.640004
5	0.635097	0.644961	0.633669	0.639265
10	0.654596	0.673526	0.652511	0.662852
15	0.662953	0.677771	0.661219	0.669392
20	0.676880	0.695655	0.674955	0.685149
25	0.710306	0.726939	0.708698	0.717702
30	0.701950	0.717462	0.700379	0.708817
35	0.713092	0.729327	0.711523	0.720315
40	0.699164	0.719044	0.697476	0.708096
45	0.713092	0.737645	0.711213	0.724187
50	0.715877	0.734548	0.714348	0.724307
55	**0.718663**	**0.743753**	**0.716785**	**0.730020**
60	0.701950	0.721432	0.700301	0.710710
65	0.693593	0.710775	0.692059	0.701292
70	0.685237	0.702374	0.683662	0.692892
75	0.685237	0.700415	0.683818	0.692017
80	0.696379	0.712209	0.694962	0.703480
85	0.690808	0.705399	0.689467	0.697342
90	0.693593	0.709749	0.692137	0.700832
95	0.690808	0.705399	0.689467	0.697342
100	0.704735	0.721573	0.703281	0.712310

Table 3: Results obtained with RSV weighting

6 Conclusions and further work

Being an unsupervised approach, the results obtained look very promising. Although supervised methods outperforms unsupervised ones, the need of a training corpus is a main drawback in the former approaches. Every

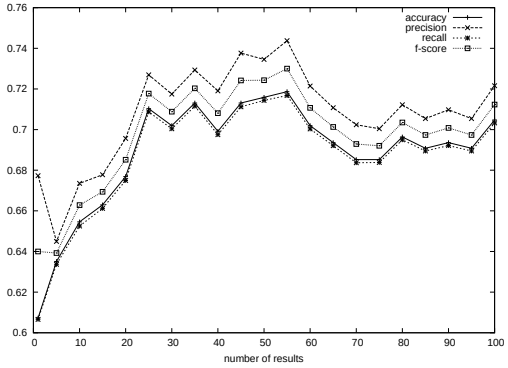

Figure 7: Effect of the number of results on performance for RSV weighting

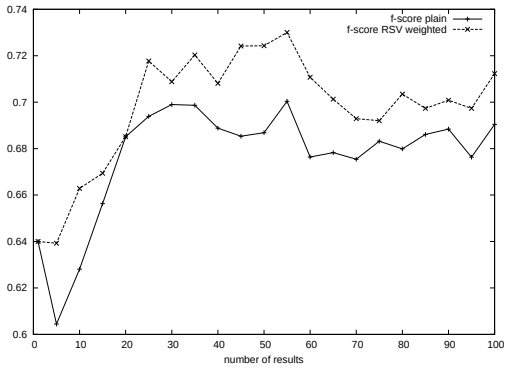

Figure 8: Plain averaging Vs. RSV weighting on F-score

day, millions of tweets flow from their authors to the web, tons of blogs are written and commented and, in many of them, feelings and emotions are expressed. The use of all the huge flow of data to semantically tag the emotions from the same flow of data represents an innovative solution to the attractive problem of sentiment polarity classification. Our experiments open a new way of tackling the problem.

Further experimentation is planned, including applying SentiWordNet (Baccianella, Esuli, and Sebastiani, 2010) as resource to determine the polarity of the crawled feelings without the need of manual intervention. Also, the method should be tested on additional data, like the i-Sieve corpus (Kouloumpis, Wilson, and Moore, 2011). Another open question is how to determine the optimal number of results from Lucene. The behaviour of Lucene RSVs could provide some clue on this issue. The normalization of the final RSVs values will be also studied. Besides, this method only performs a binary classification (positive/negative) and this is insufficient in many scenarios, where neutral or objective lables are also expected, along with a level of "intensity" in polarity values.

Despite the results derived from the experimentation to come, our approach can be easily moved to other languages. Current approaches on Multilingual Sentiment Analysis (Balahur and Turchi, 2012) rely on the traslation of lexicons or resources. In our case, a crawler of emotional publications by means of simple regular expression matching, as is done by WeFeelFine, would allow us to target any other language. This, also, is our intention in the case of Spanish, and the generation of a collection of tweets is undergoing.

References

Agarwal, N., H. Liu, J. Salerno, and S. Sundarajan. 2008. Understanding group interaction in blogosphere: a case study. In *Proc 2nd international conference on computational cultural dynamics (ICCCD), September*, pages 15–16.

Asiaee T., Amir, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1602–1606, New York, NY, USA. ACM.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Balahur, A. and M. Turchi. 2012. Multilingual sentiment analysis using machine translation? *WASSA 2012*, page 52.

Boldrini, Ester, Alexandra Balahur, Patricio Martínez-Barco, and Andrés Montoyo. 2010. Emotiblog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 1–10, Stroudsburg,

PA, USA. Association for Computational Linguistics.

Chen, L., W. Wang, M. Nagarajan, S. Wang, and A.P. Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 50–57.

Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Kamvar, Sepandar D. and Jonathan Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 117–126, New York, NY, USA. ACM.

Kouloumpis, E., T. Wilson, and J. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Lin, Y., J. Zhang, X. Wang, and A. Zhou. 2012. An information theoretic approach to sentiment polarity classification. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 35–40. ACM.

Martínez-Cámara, E., M.T. Martín-Valdivia, L.A. Ureña López, and A. Montejo-Ráez. In press. Sentiment analysis in twitter. *Natural Language Engineering*.

Pang, B. and L. Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers.

Pang, Bo, Lillian Lee, and Shivakumar 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*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reyes, A., P. Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*, 74:1–12. cited By (since 1996) 0.

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, March.

Tan, L.K.W., J.C. Na, Y.L. Theng, and K. Chang. 2012. Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *Journal of Computer Science and Technology*, 27(3):650–666.

Turney, Peter 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. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 37-44 recibido 28-11-12 revisado 14-02-13 aceptado 19-02-13

TASS - Workshop on Sentiment Analysis at SEPLN

TASS - Taller de Análisis de Sentimientos en la SEPLN

Julio Villena-Román
DAEDALUS
jvillena@daedalus.es

Sara Lana-Serrano
DIATEL - Universidad Politécnica de Madrid
slana@diatel.upm.es

Eugenio Martínez-Cámara
SINAI - Universidad de Jaén
emcamara@ujaen.es

José Carlos González-Cristóbal
GSI - Universidad Politécnica de Madrid
jgonzalez@dit.upm.es

Resumen: Este artículo describe el desarrollo de TASS, taller de evaluación experimental en el contexto de la SEPLN para fomentar la investigación en el campo del análisis de sentimiento en los medios sociales, específicamente centrado en el idioma español. El principal objetivo es promover el diseño de nuevas técnicas y algoritmos y la aplicación de los ya existentes para la implementación de complejos sistemas capaces de realizar un análisis de sentimientos basados en opiniones de textos cortos extraídos de medios sociales (concretamente Twitter). Este artículo describe las tareas propuestas, el contenido, formato y las estadísticas más importantes del corpus generado, los participantes y los diferentes enfoques planteados, así como los resultados generales obtenidos.
Palabras clave: TASS, análisis de reputación, análisis de sentimientos, medios sociales.

Abstract: This paper describes TASS, an experimental evaluation workshop within SEPLN to foster the research in the field of sentiment analysis in social media, specifically focused on Spanish language. The main objective is to promote the application of existing state-of-the-art algorithms and techniques and the design of new ones for the implementation of complex systems able to perform a sentiment analysis based on short text opinions extracted from social media messages (specifically Twitter) published by representative personalities. The paper presents the proposed tasks, the contents, format and main statistics of the generated corpus, the participant groups and their different approaches, and, finally, the overall results achieved.
Keywords: TASS, reputation analysis, sentiment analysis, social media.

1 Introduction

According to Merriam-Webster dictionary,[1] **reputation** is the overall quality or character of a given person or organization as seen or judged by people in general, or, in other words, the general recognition by other people of some characteristics or abilities for a given entity.

For the economic implications, reputation is especially important in business, where it refers to the perception or attitudes that customers, stakeholders, employees, competitors and any other agent exhibit about that organization. Reputation includes aspects such as customer satisfaction about the company's product and services, commitment and loyalty from employees, partners' trust on agreements and obligations, support from investors, etc.

In turn, **reputation analysis** is the process of tracking, investigating and reporting other entities' opinions about the entity's actions. It covers many factors to calculate the market value of reputation. Reputation analysis can be used by companies as a tool to improve competitiveness in the complex marketplace of relationships among people and companies.

Currently market research using user surveys is typically performed. However, the rise of social media such as blogs and social networks and the increasing amount of user-generated contents in the form of reviews, recommendations, ratings, etc., has led to

[1] http://www.merriam-webster.com/

ISSN 1135-5948

creation of an emerging trend towards the use of online reputation analysis.

The so-called **sentiment analysis**, i.e., the application of natural language processing and text analytics to identify and extract subjective information from texts, which is the first step towards online reputation analysis, is becoming a promising topic in the field of customer relationship management, as the social media and its associated word-of-mouth effect is turning out to be the most important source of information for companies about their customers' sentiments towards their products.

Sentiment analysis is a major technological challenge. The task is so hard that even humans often disagree on the categorization on the positive or negative sentiment that is supposed to be expressed on a given text, either in a specific segment of the text or as a global property of the full text. The fact that issues that one individual finds acceptable may not be the same to others, along with multilingual aspects, cultural factors and different contexts make it very hard to categorize a text written in a natural language into a positive or negative sentiment, even with a training based on a given user model and a context for analysis. And the shorter the text is, for example, when analyzing Twitter messages or short comments in Facebook, the harder the task becomes.

Within this context, TASS,[2] which stands for *Taller de Análisis de Sentimientos en la SEPLN* (*Workshop on Sentiment Analysis at SEPLN*, in English) is an experimental evaluation workshop, organized as a satellite event of the SEPLN 2012 Conference, held on September 7th, 2012 in Jaume I University at Castellón de la Plana, Comunidad Valenciana, Spain, to promote the research in the field of sentiment analysis in social media, initially focused on Spanish though it could be extended to any language.

The main objective was to encourage participants to improve the existing techniques and algorithms and even design new ones in order to perform a sentiment analysis in short text opinions extracted from social media messages (specifically Twitter) published by a series of important personalities. Moreover, the sentiment extraction is complemented with a text categorization, thus researching on the whole process of reputation analysis.

[2] http://www.daedalus.es/TASS

The challenge task is intended to provide a benchmark forum for comparing the latest approaches in this field. In addition, with the creation and release of the fully tagged corpus, we aim to provide a benchmark dataset that enables researchers to compare their algorithms and systems.

2 Description of tasks

Two tasks were proposed for the participants in this first edition: a first task focused on **sentiment analysis** and a second task about text categorization, which was called **trending topic coverage**. Groups could participate in both tasks or just in one of them.

Along with the submission of the results of their experiments, participants were encouraged to submit a paper to the workshop in order to describe their systems to the audience in a regular workshop session. Submitted papers were reviewed by the program committee.

2.1 Task 1: Sentiment Analysis

This task consists on performing an automatic sentiment analysis to determine the polarity of each message in the test corpus.

The evaluation metrics to evaluate and compare the different systems are the usual measurements of precision (1), recall (2) and F-measure (3) calculated over the full test set.

$$\text{precision} = \frac{\text{N(correct classifications)}}{\text{N(all classifications)}} \quad (1)$$

$$\text{recall} = \frac{\text{N(retrieved documents)}}{\text{N(all documents)}} \quad (2)$$

$$F(\beta) = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \quad (3)$$

2.2 Task 2: Trending topic coverage

In this case, the technological challenge is to build a classifier to identify the topic of the text, and then apply the polarity analysis to get the assessment for each topic.

The evaluation metrics are the same as in Task 1 (precision, recall and F-measure).

3 Corpus

The corpus provided to participants contains over 70,000 tweets, written in Spanish by nearly 200 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, between November 2011 and March 2012.

Although the context of extraction has a Spain-focused bias, the diverse origin of the authors, including people from Spain, Mexico, Colombia, Puerto Rico, USA and many other countries, makes the corpus reach a global coverage in the Spanish-speaking world.

Due to restrictions in the Twitter API Terms of Service,[3] it is forbidden to redistribute a corpus that includes text contents or information about users. However, it is valid if those fields are removed and instead IDs (including Tweet IDs and user IDs) are provided. The actual message content can be easily obtained by making queries to the Twitter API using the Tweet ID. In addition, using the user ID, it is possible to extract information about the user's name, registration date, geographical information of their location, and many other fields, which may allow to perform experiments for instance on the different varieties of Spanish.

Thus each Twitter message includes its Tweet ID (`twitid`), the user ID (`user`) and the creation date (`date`). Each message is annotated with its global polarity for sentiment, i.e., an indication of whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. 5 polarity levels have been defined: *strong positive* (`P+`), *positive* (`P`), *neutral* (`NEU`), negative (`N`), strong negative (`N+`) and one additional *no sentiment* label (`NONE`).

Moreover, in those cases where applicable, this same polarity levels are annotated but related to the entities that are mentioned in the text. This is done when the sentiment is referring to the identified entity but not in cases where an entity appears in the text but is not involved in the expressed sentiment.

There is also an indication of the level of *agreement* or *disagreement* of the expressed sentiment within the content. This is especially useful to make out whether a neutral sentiment comes from a neutral set of keywords (i.e., slightly positive or negative) or else the text contains positive and negative sentiments at the same time. For example, "*Peter is a very good friend but I cannot stand John*" could be considered `NEU` with `DISAGREEMENT` where *Peter* is regarded as `P+` and *John* as `N+`.

On the other hand, a selection of a set of 10 topics has been made based on the thematic areas covered by the corpus, such as *politics* (*política*), *soccer* (*fútbol*), *literature* (*literatura*) or *entertainment* (*entretenimiento*).

Each message of the corpus has been semiautomatically assigned to one or several of these topics. A baseline model (Villena-Román *et al.*, 2011) was used to obtain the candidate topics that then were manually revised.

This tagged corpus has been divided into two sets: training and test. The *training corpus* was released along with the corresponding tags so that participants may train and validate their models for classification and sentiment analysis. The *test corpus* was provided without any tag and was used to evaluate the results provided by the different systems.

Table 1 shows a summary of the training and test data provided to participants.

Attribute	Train corpus	Test corpus
Tweets	7 219	60 798
Topics	10	10
Tweet languages	1	1
Users	154	158
User types	3	3
User languages	1	1
Date start	2011-12-02 T00:47:55	2011-12-02 T00:03:32
Date end	2012-04-10 T23:40:36	2012-04-10 T23:47:55

Table 1: Train and test corpus

There were 3 user types: journalists (*periodistas*), politicians (*políticos)* or celebrities (*famosos*). The only language involved this year was Spanish (*es)*.

The list of selected topics is shown in Table 2, sorted by frequency in the test corpus.

Topic	Frequency
Politics (*política*)	3 119
Other (*otros*)	2 337
Entertainment (*entretenimiento*)	1 677
Economy (*economía*)	942
Music (*música*)	566
Soccer (*fútbol*)	252
Films (*cine*)	245
Technology (*tecnología*)	217
Sports (*deportes*)	113
Literature (*literatura*)	99

Table 2: Topic list

The corpus is encoded in XML in which the text of the content entity has been removed to

[3] https://dev.twitter.com/terms/api-terms

follow the Twitter restrictions. Two sample tweets are shown in Figure 1. The second one is tagged with both the global polarity of the message and the polarity associated to each one of the entities that appears in the text ("*UPyD*" and "*Foro Asturias*"), whereas the first tweet is only tagged with the global polarity as the text contains no mentions to any entity.

```
<twit>
    <twitid>0000000000</twitid>
    <user>usuario0</user>
    <content><![CDATA['Conozco a alguien q es adicto al drama!
        Ja ja ja te suena d algo!]]></content>
    <date>2011-12-02T02:59:03</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P+</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>entretenimiento</topic>
    </topics>
</twit>

<twit>
    <twitid>0000000001</twitid>
    <user>usuario1</user>
    <content><![CDATA['UPyD contará casi seguro con grupo gracias
        al Foro Asturias.]]></content>
    <date>2011-12-02T00:21:01</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>UPyD</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>Foro_Asturias</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>politica</topic>
    </topics>
</twit>
```

Figure 1: Sample tweet messages

The full corpus was made public after the workshop[4] so that any group interested in the field of sentiment analysis in Spanish could make use of it in their own research.

4 Participants

Participants were required to register for the task(s) in order to obtain the corpus.

Results should be submitted in a plain text file with the following format:

```
twitid \t polarity \t topic
```

Where `twitid` is the Tweet ID for every message in the test corpus, the `polarity` contains one of the 6 valid tags (`P+`, `P`, `NEU`, `N`, `N+` and `NONE`), and the same for `topic`.

Although the polarity level should be classified into those 5 levels and results were primarily evaluated for them, the evaluation also included metrics with just 3 levels (*positive*, *neutral* and *negative*).

Participants could submit results for one or both tasks. Several results for the same task were allowed too.

15 groups registered and finally 8 groups sent their submissions for one of the two tasks. The list of active groups is shown in Table 3. All of them submitted results for the sentiment analysis tasks and most of them (6 out of 8, 75%) participated in both tasks.

Group	Task 1	Task 2
Elhuyar Fundazioa	Yes	No
IMDEA	Yes	Yes
L2F - INESC	Yes	Yes
La Salle - URL	Yes	Yes
LSI UNED	Yes	Yes
LSI UNED 2	Yes	Yes
SINAI - UJAEN	Yes	Yes
UMA	Yes	No

Table 3: Participant groups

There was another group at Delft University of Technology that submitted experiments for both tasks but finally did not submit a report for the workshop, so their results are not included.

The next sections briefly describe the approaches for the different groups.

4.1 Elhuyar Fundazioa

In their paper, Saralegi and San Vicente (2012) describe their supervised approach that includes some linguistic knowledge-based processing for preparing the features.

The processing comprises lemmatization, part-of-speech tagging, tagging of polarity words, treatment of emoticons, negation, and weighting of polarity words depending on syntactic nesting level. A pre-processing step for handling spelling errors was also performed.

Detection of polarity words is done according to a polarity lexicon built in two ways: projection to Spanish of an English lexicon, and extraction of divergent words of positive and negative tweets of training corpus.

Evaluation results show a good performance and also good robustness of the system both for

[4] http://www.daedalus.es/TASS/corpus.php

the fine granularity (65% of accuracy) as well as for coarse granularity polarity detection (71% of accuracy).

4.2 IMDEA

The IMDEA (Instituto Madrileño de Estudios Avanzados) team state that sentiment analysis and topic detection are new problems that are at the intersection of natural language processing and data mining (Fernandez Anta *et al.*, 2012).

An interesting comparative analysis of different approaches and classification techniques for these problems is presented.

The data is preprocessed using well-known techniques and tools proposed in the literature, together with others specifically proposed here that take into account the characteristics of Twitter. Then, popular classifiers have been used, in particular, most popular classifiers of WEKA (Hall *et al.*, 2009) were evaluated. Their report describes some of the results obtained in their preliminary research.

4.3 L2F – INESC

The strategy used by the L2F (Laboratório de sistemas de Língua Falada) team at INESC (Instituto de Engenharia de Sistemas e Computadores) for performing automatic sentiment analysis and topic classification over Spanish Twitter data is described by Batista and Ribeiro (2012).

They have decided to consider both tasks as classification tasks, thus sharing the same method. Their most successful and recent experiments in this field cast the problem as a binary classification problem, which aims at discriminating between two possible classes. Binary decisions are stated as "document matches/does not match category A", and a binary classifier exists for each polarity level and topic. Binary classifiers are easier to develop, offer faster convergence ratios, and can be executed in parallel. The final results are then generated by combining all the different binary classifiers.

Specifically, they have adopted an approach based on logistic regression classification models, which corresponds to the maximum entropy classification for independent events.

As described in their paper, the L2F system achieved the best results for the topic classification contest, and the second place in terms of sentiment analysis.

4.4 La Salle – Universitat Ramon Llull

Trilla and Alías (2012) describe how they adapt a text classification scheme based on multinomial naive Bayes. The multinomial naive Bayes is a probabilistic generative approach that builds a language model assuming conditional independence among the linguistic features. Therefore, no sense of history, sequence nor order is introduced in this model. This approach achieves a good result in terms of the evaluation metrics.

4.5 LSI – UNED

Martín-Wanton and Carrillo de Albornoz, (2012) present the participation of the LSI (Lenguajes y Sistemas Informáticos) group at UNED (Universidad Nacional de Educación a Distancia) in TASS. For polarity classification, they propose an emotional concept-based method. The original method makes use of an affective lexicon to represent the text as the set of emotional meanings it expresses, along with advanced syntactic techniques to identify negations and intensifiers, their scope and their effect on the emotions affected by them.

Besides, the method addresses the problem of word ambiguity, taking into account the contextual meaning of terms by using a word sense disambiguation algorithm. On the other hand, for topic detection, their system is based on a probabilistic model called Twitter-LDA, based on Latent Dirichlet Allocation technique. They first build for each topic of the task a lexicon of words that best describe it, thus representing each topic as a ranking of discriminative words. Moreover, a set of events is retrieved based on a probabilistic approach adapted to the characteristics of Twitter.

To determine which of the topics corresponds to each event, the topic with the highest statistical correlation was obtained comparing the ranking of words of each topic and the ranking of words most likely to belong to the event.

The experimental results achieved show the adequacy of their approach for the task.

4.6 LSI – UNED 2

Castellano, Cigarrán and García-Serrano (2012a) describe the research done for the workshop by the second team component of the second group from LSI at UNED.

Their proposal addresses the sentiment and topic detection from an information retrieval perspective, based on language divergences. Kullback-Liebler divergence (computed against the testing corpus) is used to generate both, polarity and topic models, which will be used in the information retrieval process (Castellano, Cigarrán and García-Serrano, 2012b).

In order to improve the accuracy of the results, they propose several approaches focused on carry out language models, not only considering the textual content associated to each tweet but, as an alternative, the named entities or adjectives detected as well.

Results show that modeling the tweets set using named entities and adjectives improves the final precision results and, as a consequence, their representativeness in the model compared with the use of common terms.

General results are promising (fifth and fourth position in each of the proposed tasks), indicating that an IR and language models based approach may be an alternative to other classical proposals focused on the application of classification techniques.

4.7 SINAI – Universidad de Jaén

The participation of the SINAI (Sistemas Inteligentes de Acceso a la Información) research group of the University of Jaén is described by Martínez Cámara *et al.* (2012).

For the first task, they have chosen a supervised machine learning approach, in which they have used support vector machines (SVM) for classifying the polarity. Text features included are unigrams, emoticons, positive and negative words and intensity markers.

In the second task, they have also used SVM for the topic classification but several bags of words (BoW) have been used with the goal of improving the classification performance.

One BoW has been obtained using Google Adwords Keyword Tool,[5] which allows to enter a term and directly returns the top *N* related concepts. The second BoW has generated based on the hash tags of the training tweets, per each category.

4.8 Universidad de Málaga

Moreno-Ortiz and Pérez-Hernández (2012) describe the participation of the group at Facultad de Filosofía y Letras in Universidad de Málaga (UMA). They use a lexicon-based approach to sentiment analysis. These approaches differ from the more common machine-learning based approaches in that the former rely solely on previously generated lexical resources that store polarity information for lexical items, which are then identified in the texts, assigned a polarity tag, and finally weighed, to come up with an overall score for the text.

Such systems have been proved to perform on par with supervised, statistical systems, with the added benefit of not requiring a training set. However, it remains to be seen whether such lexically-motivated systems can cope equally well with extremely short texts, as generated on social networking sites, such as Twitter.

In their paper they perform such an evaluation using Sentitext, a lexicon-based sentiment analysis tool for Spanish. One conclusion is that Sentitext's Global Sentiment Value is strongly affected by the number of lexical units available in the text (or the lack of them, rather). On the other hand, they also confirm Sentitext's tendency to assign middle-of-the-scale ratings, or at least avoid extreme values, which is reflected on its poor performance for the `N+` and `P+` classes, most of which were assigned to the more neutral `N` and `P` classes.

Another interesting conclusion which is drawn from their analysis of the average number of polarity lexical segments and Sentitext's Affect Intensity (an internal measure similar to the polarity level) is that Twitter users employ highly emotional language.

5 Results

The gold standard has been generated by first pooling all submissions, then a voting scheme has been applied and finally an extensive human review of the ambiguous decisions (thousands of them). Due to the high volume of data, this was the only way to generate a tagged set; unfortunately, this is subject to errors and misclassifications. Obviously, if all annotators are consistently wrong, the gold standard will end up with a wrong label, and accuracy figures will then be an upper bound of actual accuracy.

A manual evaluation of a part of the gold standard to assess its quality has not been done yet, although this task is planned for future editions of the workshop.

[5] https://adwords.google.com/o/KeywordTool

Both tasks have been evaluated as a single label classification. This specifically affects to the topic classification, where the most restrictive criterion has been applied: a "success" is achieved only when all the test labels have been returned. Participants were welcome to discuss and reevaluate their experiments with a less restrictive strategy in their papers.

Regarding Task 1, 17 different experiments were submitted. Results are listed in the tables below. All tables show the precision value achieved in each experiment (recall and F-measure were also evaluated and provided to participants but are omitted here).

Table 4 considers 5 levels of sentiments (P+, P, NEU, N, N+) and no sentiment (NONE).

Precision values range from 65.3% to 16.7%. Only 8 from 20 submissions achieve figures higher than 50% and specifically 5 of the 9 groups have at least one submission above this value. Besides, results for different submissions from the same group are typically very similar except for the SINAI group.

Run Id	Group	Precision
pol-elhuyar-1-5l	Elhuyar Fund.	65.3%
pol-l2f-1-5l	L2F - INESC	63.4%
pol-l2f-3-5l	L2F - INESC	63.3%
pol-l2f-2-5l	L2F - INESC	62.2%
pol-atrilla-1-5l	La Salle - URL	57.0%
pol-sinai-4-5l	SINAI - UJAEN	54.7%
pol-uned1-2-5l	LSI UNED	53.9%
pol-uned1-1-5l	LSI UNED	52.5%
pol-uned2-2-5l	LSI UNED 2	40.4%
pol-uned2-1-5l	LSI UNED 2	40.0%
pol-uned2-3-5l	LSI UNED 2	39.5%
pol-uned2-4-5l	LSI UNED 2	38.6%
pol-imdea-1-5l	IMDEA	36.0%
pol-sinai-2-5l	SINAI - UJAEN	35.7%
pol-sinai-1-5l	SINAI - UJAEN	35.3%
pol-sinai-3-5l	SINAI - UJAEN	35.0%
pol-uma-1-5l	UMA	16.7%

Table 4: Results for task 1 (Sentiment Analysis) with 5 levels + NONE

In order to perform a supplementary evaluation, Table 5 gives results considering the classification only in 3 levels (POS, NEU, NEG) and no sentiment (NONE) merging P and P+ in only one category, as well as N and N+ in another one.

In this case, precision values improve, as expected. The precision obtained now ranges from 71.1% to 35.1%. In this case, 9 submissions have a precision value over 50% and 6 groups have at least one result over this percent.

Run Id	Group	Precision
pol-elhuyar-1-3l	Elhuyar Fund.	71.1%
pol-l2f-1-3l	L2F - INESC	69.0%
pol-l2f-3-3l	L2F - INESC	69.0%
pol-l2f-2-3l	L2F - INESC	67.6%
pol-atrilla-1-3l	La Salle - URL	62.0%
pol-sinai-4-3l	SINAI - UJAEN	60.6%
pol-uned1-1-3l	LSI UNED	59.0%
pol-uned1-2-3l	LSI UNED	58.8%
pol-uned2-1-3l	LSI UNED 2	50.1%
pol-imdea-1-3l	IMDEA	46.0%
pol-uned2-2-3l	LSI UNED 2	43.6%
pol-uned2-4-3l	LSI UNED 2	41.2%
pol-uned2-3-3l	LSI UNED 2	40.4%
pol-uma-1-3l	UMA	37.6%
pol-sinai-2-3l	SINAI - UJAEN	35.8%
pol-sinai-1-3l	SINAI - UJAEN	35.6%
pol-sinai-3-3l	SINAI - UJAEN	35.1%

Table 5: Results task 1 (Sentiment Analysis) with 3 levels + NONE

Table 6 shows the results for Task 2. 13 experiments were submitted in.

Run Id	Group	Precision
top-l2f-2	L2F – INESC	65.4%
top-l2f-1y3	L2F – INESC	64.9%
top-atrilla-1	La Salle - URL	60.1%
pol-uned2-5a8	LSI UNED 2	45.3%
top-imdea-1	IMDEA	45.2%
pol-uned2-9a12	LSI UNED 2	42.2%
pol-uned2-1a4	LSI UNED 2	40.5%
top-sinai-5	SINAI - UJAEN	39.4%
top-sinai-4	SINAI - UJAEN	37.8%
top-sinai-2	SINAI - UJAEN	34.8%
top-sinai-3	SINAI - UJAEN	34.1%
top-sinai-1	SINAI - UJAEN	32.3%
pol-uned1-1y2	LSI UNED	31.0%

Table 6: Results for task 2 (Trending topic coverage)

In this task, precision ranges from 65.4% to 31.0% and only 4 submissions are above 50% (2 groups). As in task 1, different submissions from the same group usually get a similar precision, thus showing that the variations from each baseline do not affect much.

6 Conclusions and Future Work

TASS has been the first workshop about sentiment analysis in the context of SEPLN. We expected to attract a certain interest in the proposed tasks, as many groups around the world are currently carrying out an intense research in sentiment/opinion analysis in general and using short-texts in particular. We think that the number of participants, the quality of their work and their reports, and the good results achieved in such hard tasks, has met and gone beyond all our expectations.

The diversity of groups coming from different fields and areas of expertise including information retrieval, natural language processing, computational linguistics, machine learning, data or text mining, text analytics, and even semantic web, has shown that the sentiment analysis is becoming a trending topic within the information technology field.

Some participants expressed in their papers and during the workshop some concerns about the quality of both the annotation of the training corpus and also of the gold standard (the test corpus). In case of future editions of TASS and the reuse of the corpus, more effort must be invested in filtering errors and improving the annotation of the corpora.

Furthermore, as expressed by Moreno-Ortiz and Pérez-Hernández (2012), there is a need of further discussion about whether differentiating between neutral and no polarity is the best decision, since it is not always clear what the difference is, and, moreover, if this distinction is interesting from a practical perspective.

In future editions of the workshop, it would be interesting to extend the corpus to other languages to compare the performance of the different approaches on different languages.

References

Villena-Román, Julio; Collada-Pérez, Sonia; Lana-Serrano, Sara; González-Cristóbal; José Carlos. 2011. *Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization*. Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press 2011.

Saralegi Urizar, Xabier; San Vicente Roncal, Iñaki. 2012. *TASS: Detecting Sentiments in Spanish Tweets*. TASS 2012 Working Notes. Castellón, September 2012.

Fernández Anta, Antonio; Morere, Philippe; Núñez Chiroque, Luis; and Santos, Agustín. 2012. *Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report*. TASS 2012 Working Notes. Castellón, September 2012.

Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

Batista, Fernando; Ribeiro, Ricardo. 2012. *The L2F Strategy for Sentiment Analysis and Topic Classification*. TASS 2012 Working Notes. Castellón, September 2012.

Trilla, Alexandre; Alías, Francesc. 2012. *Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes*. TASS 2012 Working Notes. Castellón, September 2012.

Martín-Wanton, Tamara; Carrillo de Albornoz, Jorge. 2012. *UNED en TASS 2012: Sistema para la Clasificación de la Polaridad y Seguimiento de Temas*. TASS 2012 Working Notes. Castellón, September 2012.

Castellanos, Ángel; Cigarrán, Juan; García-Serrano, Ana. 2012. *Generación de un corpus de usuarios basado en divergencias del Lenguaje*. II Congreso Español de Recuperación de Información. Valencia, June 2012.

Castellano, Angel; Cigarrán, Juan; García-Serrano, Ana. 2012. *UNED @ TASS: Using Information Retrieval techniques for topic-based sentiment analysis through divergence models*. TASS 2012 Working Notes. Castellón, September 2012.

Martínez Cámara, Eugenio; García Cumbreras, M. Ángel. Martín Valdivia, M. Teresa; Ureña López, L. Alfonso. 2012. *SINAI at TASS 2012*. TASS 2012 Working Notes. Castellón, September 2012.

Moreno-Ortiz, Antonio; Pérez-Hernández, Chantal. 2012. *Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish*. TASS 2012 Working Notes. Castellón, September 2012.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 45-52 recibido 28-11-12 revisado 13-02-13 aceptado 19-02-13

Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques *

Análisis de sentimientos y detección de asunto de tweets en español: un estudio comparativo de técnicas de PLN

Antonio Fernández Anta[1] **Luis Núñez Chiroque**[1]
Philippe Morere[2†] **Agustín Santos**[1]
[1] Institute IMDEA Networks, Madrid, Spain
[2] ENSEIRB-MATMECA, Bordeaux, France
{antonio.fernandez, luisfelipe.nunez, philippe.morere, agustin.santos}@imdea.es

Resumen: Se está invirtiendo mucho esfuerzo en la construcción de soluciones efectivas para el análisis de sentimientos y detección de asunto, pero principalmente para textos en inglés. Usando un corpus de tweets en español, presentamos aquí un análisis comparativo de diversas aproximaciones y técnicas de clasificación para estos problemas.
Palabras clave: Análisis de sentimientos, detección de asunto.

Abstract: A significant amount of effort is been invested in constructing effective solutions for sentiment analysis and topic detection, but mostly for English texts. Using a corpus of Spanish tweets, we present a comparative analysis of different approaches and classification techniques for these problems.
Keywords: Sentiment analysis, topic detection.

1 Introduction

With the proliferation of online reviews, ratings, recommendations, and other forms of online opinion expression, there is a growing interest in techniques for automatically extracting the information they embody. Two of the problems that have been posed to achieve this are *sentiment analysis* and *topic detection*, which are at the intersection of natural language processing (NLP) and data mining. Research in both problems is very active, and a number of methods and techniques have been proposed in the literature to solve them. Most of these techniques focus on English texts and study large documents. In our work, we are interested in languages different from English and micro-texts. In particular, we are interested in sentiment and topic classification applied to Spanish Twitter micro-blogs. Spanish is increasingly present over the Internet, and Twitter has become a popular method to publish thoughts and information with its own characteristics. For instance, publications in Twitter take the form of *tweets* (i.e., Twitter messages), which are micro-texts with a maximum of 140 characters. In Spanish tweets, it is common to find specific Spanish elements (SMS abbreviations, hashtags, slang). The combination of these two aspects makes this a distinctive research topic, with potentially deep industrial applications.

The motivation of our research is twofold. On the one hand, we would like to know whether usual approaches that have been proved to be effective with English text are also so with Spanish tweets. On the other hand, we would like to identify the best (or at least a good) technique for processing Spanish tweets. For this second question, we would like to evaluate those techniques proposed in the literature, and possibly propose new ad hoc techniques for our specific context. In our study, we try to sketch out a comparative study of several schemes on term weighting, linguistic preprocessing (stemming and lemmatization), term definition (e.g., based on uni-grams or n-grams), the combination of several dictionaries (sentiment, SMS abbreviations, emoticons, spell, etc.) and the use of several classification methods.

1.1 Related Work

Sentiment analysis is a challenging NLP problem. Due to its tremendous value for

* Partially funded by Factory Holding Company 25, S.L. The authors want to thank Fernando Pérez for useful discussions.
† Partially done while at Institute IMDEA Networks.

ISSN 1135-5948

practical applications, it has experienced a lot of attention, and it is perhaps one of the most widely studied topic in the NLP field. Pang and Lee (2008) have a comprehensive survey of sentiment analysis and opinion mining research. Liu (2010), on his hand, reviews and discusses a wide collection of related works. Although most of the research conducted focuses on English texts, the number of papers on the treatment of other languages is increasing every day. Examples of research papers on Spanish texts are (Brooke, Tofiloski, and Taboada, 2009; Martínez-Cámara, Martín-Valdivia, and Ureña-López, 2011; Sidorov et al., 2012).

Most of the algorithms for sentiment analysis and topic detection use a collection of data to train a classifier, which is later used to process the real data. Data is preprocessed before being used in the classifier in order to correct errors and extract the main features. Many different techniques have been proposed for these two phases. For instance, different classification methods have been proposed, like Naive Bayes, Maximum Entropy, Support Vector Machines (SVM), BBR, KNN, or C4.5. In fact, there is no final agreement on which of these classifiers is the best. For instance, Go, Bhayani, and Huang (2009) report similar accuracies with classifiers based on Naive Bayes, Maximum Entropy, and SVM.

Regarding preprocessing the data, Laboreiro et al. (2010) explore tweets tokenization (or symbol segmentation) as the first key task for text processing. Once single words or terms are available, typical choices are using uni-grams, bi-grams, n-gram, or parts-of-speech (POS) as basic terms. Again, there is no clear conclusion on which is the best option, since Pak and Paroubek (2010) report the best performance with bi-grams, while Go, Bhayani, and Huang (2009) present better results with unigrams. The preprocessing phase may also involve word processing the input texts: stemming, spelling and/or semantic analysis. Tweets are usually very short, having emoticons like :) or :-), or abbreviated (SMS) words like *Bss* for *Besos* (*kisses*). Agarwal et al. (2011) propose the use of several dictionaries: an emoticon dictionary and an acronym dictionary. Other preprocessing tasks that have been proposed are contextual spell-checking and name normalization (Kukich, 1992).

One important question is whether the algorithms and techniques proposed for other types of data can be directly applied to tweets. Twitter data poses new and different challenges, as discussed by Agarwal et al. (2011) when reviewing some early and recent results on sentiment analysis of Twitter data (e.g., (Go, Bhayani, and Huang, 2009; Bermingham and Smeaton, 2010; Pak and Paroubek, 2010)). Engström (2004) has also shown that the bag-of-features approach is topic-dependent and Read (2005) demonstrated how models are also domain-dependent.

These papers, as expected, use a broad spectrum of tools for the extraction and classification processes. For feature extraction, *FreeLing* (Padró et al., 2010) has been proposed, which is a powerful open-source language processing software. We use it as analyzer and for lemmatization. For classification, Justin et al. (2010) report very good results using WEKA, which is one of the most widely used tools for the classification phase. Other authors proposed the use of additional libraries like LibSVM (Chang and Lin, 2011).

Most of the references above have to do with sentiment analysis, due to its popularity. However, the problem of topic detection is becoming also popular (Sriram et al., 2010), among other reasons, to identify trending topics (Allan, 2002; Bermingham and Smeaton, 2010; Lee et al., 2011). Due to the the real time nature of Twitter data, most works (Mathioudakis and Koudas, 2010; Vakali, Giatsoglou, and Antaris, 2012) are interested in breaking news detection and tracking. They propose methods for the classification of tweets in an open (dynamic) set of topics. Instead, we are interested in a closed (fixed) set of topics. However, we explore all the indexing and clustering techniques proposed, since most of them could also be applied to sentiment analysis.

1.2 Contributions

In this paper we have explored the performance of several preprocessing, feature extraction, and classification methods in a corpus of Spanish tweets, both for sentiment analysis and for topic detection. The different methods considered can be classified into almost orthogonal families, so that a different method can be selected from each family

to form a different configuration. In particular, we have explored the following families of methods.

Term definition and counting. In this family it is decided what constitutes a basic term to be considered by the classification algorithm. The different alternatives are using single words (uni-grams), or groups of words (bi-grams, tri-grams, n-grams) as basic terms.

Stemming and lemmatization. One of the main difference between Spanish and English is that the latter is a weakly inflected language in contrast to Spanish, a highly inflected one. One interesting questions is to compare how well the usual stemming and lemmatization processes perform with Spanish words.

Word processing and correction. We have used several dictionaries in order to correct the words and replace emoticons, SMS abbreviations, and slang terms by their meaning in correct Spanish. Finally, it is possible to use a morphological analyzer to determine the type of each word. Thus, a word-type filter can be applied to tweets.

Valence shifters. An alternative to the usual direct term-counting method is the processing of valence shifters and negative words (*not, neither, very, little*, etc). We think that those words are useful for sentiment classification since they change and/or revert the strength of a neighboring term.

Tweet semantics. The above approaches can be improved by processing specific tweet artifacts such as author tags, or hashtags and URLs (links), provided in the text. The author tags act like a history of the tweets of a specific person. Additionaly, the hashtags are a great indicator of the topic of a tweet, whereas retrieving keywords from the webpage linked within a tweet allows to overpass the limit of the 140 characters and thus improves the efficiency of the estimation. Another way to overpass this limit is to investigate the keywords of a tweet in a search-engine to retrieve other words of the same context.

Classification methods. In addition to these variants, we have explored the full spectrum of classification methods provided by WEKA.

The rest of the paper is structured as follows. In Section 2 we describe in detail the different techniques that we have implemented or used. In Section 3 we describe our evaluation scenario and the results we have obtained. Finally, in Section 4 we present some conclusions and open problems.

2 Methodology

In this section we give the details of how the different methods considered have been implemented in our system.

2.1 Attributes Definition and Preprocessing

n-grams As we mentioned, classifiers will consider sets of n words (n-grams), with unigrams as a special case. The value of n could be defined in our algorithm. When using n-grams, n is a parameter that highly influences performance. We found that, in practice, having n larger than 3 did not improve the results, so we limit n by that value.

Of course, it is possible to combine n-grams with several values of n. The drawback of this is the high number of entries in the final attribute list. Hence, when doing this, a threshold is used to remove all the attributes that appear too few times in the data set, as they are considered as noise. We force that the attribute appears at least 5 times in the data set to be considered. Also, a second threshold is used to remove ambiguous attributes. This threshold has been set to 85%, which means that more than 85% of the occurrences of an attribute have to be for a specific topic or sentiment.

Processing Terms The processing of terms involves first building the list of attributes, which is the list of different terms that appear in the data set of interest. In principle, the data set used to identify attributes is formed at least by all the tweets that are provided as input to the algorithm, but there are cases in which we do not use them. For instance, when using an affective dictionary (see below) we may not use the input data. Moreover, even if the input data is processes, we may filter it and only keep some of it. For instance, we may decide to use only nouns. In summary, the list of attributes is built from the input data (if so decided) preprocessed as determined and, potentially, by additional data (like the affective dictionary). Once this process is completed, the list of attributes and the list of vectors obtained from the tweets are passed to the classifier.

Stemming and Lemmatization When creating the list of attributes, typically only

the root of the words is used in the attribute list. The root can take the form of the lemma or the stem of the word (lemmatization or stemming, respectively). We have used the FreeLing software to perform the lemmatization process. The Snowball software stemmer has been used in our experiments. We have decided to always use one of the two processes.

Word Processing and Correction As mentioned above, one of the possible preprocessing steps before extracting attributes and vectors is to correct spelling errors. If correction is done, the algorithm uses the Hunspell dictionary to perform it.

Another optional preprocessing step expands the emoticons, shorthand notations, and slang commonly used in SMS messages which is not understandable by the Hunspell dictionary. The use of these abbreviations is common in tweets, given the limitation to 140 characters. An SMS dictionary is used to do the preprocessing. It transforms the SMS notations into words understandable by the main dictionary. Also, the emoticons are replaced by words that describe their meaning. For example `:-)` is replaced by *feliz* (*happy*).

We have observed that the information of a sentence is mainly located in a few keywords. These keywords have a different type according to the information we are interested in. For topic estimation, the keywords are mainly nouns and verbs, whereas for sentiment analysis they are adjectives and verbs. For example, in the sentence *La pelicula es buena* (*The movie is good*), the only word that is carrying the topic information is the noun *pelicula*, which is very specific to the cinema topic. Besides, the word that best reflects the sentiment of the sentence is the adjective *buena*, which is positive. Also, in the sentence *El equipo ganó el partido* (*The team won the match*), the verb *ganó* is carrying information for both topic and sentiment analysis: the verb *ganar* is used very often in the soccer and sport topics, and has a positive sentiment. We allow to filter the words of the input data using their type. The filtering is done using the FreeLing software, which is used to extract the type of each word.

When performing sentiment analysis, we have found useful to have an *affective dictionary*. This dictionary consist of a list of words that have a positive or negative meaning, expanded by their polarity "P" or "N" and their strength "+" or "-". For example, the words *bueno* (*good*) and *malo* (*bad*) are respectively positive and negative with no strength whereas the words *mejor* (*best*) and *peor* (*worse*) are respectively positive and negative with a positive strength. As a first approach, we have not intensively used the polarity and the strength of the affective words in the dictionary. Its use only forces the words that contain it to be added as attributes. This has the advantage of drastically reducing the size of the attribute list, specially if the input data is filtered. Observe that the use of this dictionary for sentiment analysis is very pertinent, since the affective words carry the tweet polarity information. In a more advanced future approach, the characteristics of the words could be used to compute weights. Since not all the words in our affective dictionary may appear in the corpus we have used, we have built *artificial* vectors for the learning machine. There is one artificial vector per sentiment analysis category (positive+, positive, negative, negative+, none), which has been built counting one occurrence of those words whose polarity and strength match with the appropriate category.

Valence Shifters There are two different aspects of valence shifting that are used in our methods. First, we may take into account negations that can invert the sentiment of positive and negative terms in a tweet. Second, we may take weighted words, which are intensifiers or weakeners, into account.

Negations are words that reverse the sentiment of other words. For example, in the sentence *La pelicula* ***no*** *es buena* (*The movie is* ***not*** *good*), the word *buena* is positive whereas it should be negative because of the negation *no*. The way we process negations is as follows. Whenever a negative word is found, the sign of the 3 terms that follow it is reversed. This allows us to differentiate a positive *buena* from a negative *buena*. The area of effect of the negation is restricted to avoid false negative words in more sophisticated sentences.

Other valence shifters are words that change the degree of the expressed sentiment. Examples of these are, for instance *muy* (*very*), which increases the degree, or *poco* (*little*), which decreases it. If the valence shifter is positive, the weight is multiplied by 3, while if it is negative by 0.5.

Twitter Artifacts It has been noticed that with the previous methods, not all the potential data contained in the tweets is used. There are several frequent element in tweets that carry a significant amount of information. Among others we have the following:
Hashtags(any word which starts with "#"). They are used for identify messages about the same topic since some of them may carry more topic information than the rest of the tweet. For example, if a tweet contains `#BAR`, which is the hashtag of the Barcelona soccer team, it can almost doubtlessly be classified in a soccer tweet.
References (a "@" followed by the username of the referenced user). References are interesting because some users appear more frequently in certain topics and will more likely tweet about them. A similar behaviour can be found for sentiment.
Links (a URL). Because of the character limitation of the tweets, users often include URLs of webpages where more details about the message can be found. This may help obtaining more context, specially for topic detection.

In our algorithms, we have the possibility of including hashtags and references as attributes. We believe that these options are just a complement to previous methods and cannot be used alone, because we have found that the number of hashtags and references in the tweets is too small. We also provide the possibility of adding to the terms of a tweet the terms obtained from the web pages linked from the tweet. A first approach could have been retrieving the whole source code of the linked page, get all the terms it contains, and keep the ones that match the attribute list. Unfortunately, there are too many terms, and the menus in the pages induce an unexpected noise which degrades the results. The approach we have chosen is to keep only the keywords of the pages. We chose to retrieve only the text within the HTML tags `h1`, `h2`, `h3` and `title`. The results with this second method are much better since the keywords are directly related to the topic.

Because of the short length of the tweets, our estimations often suffer from a lack of words. We found a solution to this problem in several papers (Banerjee, Ramanathan, and Gupta, 2007; Gabrilovich and Markovitch, 2005; Rahimtoroghi and Shakery, 2011) that use web sources (like Wikipedia or the Open Directory) to complete tweets. The web is a mine of information and search-engines can be used to retrieve it. We have used this technique to obtain many keywords and a context from just a few words taken from the tweets. For implementation reasons, Bing was chosen for the process. The title and description of the 10 first results of the search are kept and processed in the same way as the words of the tweet. We found out that we have better results by searching in Bing with only the nouns contained in the tweet; therefore, this is the option we chose.

2.2 Classification Methods

For classification, we use WEKA[1], which is a collection of machine learning algorithms that can be used for classification and clustering. It includes algorithms for classification, regression, clustering attribute selection and association rule mining. Almost all popular classification algorithms are included (Bayesian methods, decision tree learners, random trees and forests, etc.).

For each experiment we set up a configuration that tells our algorithm which attributes to choose and how to create vectors of attributes. The output of this algorithm is a WEKA file for a specific configuration and the input data. Once this file is available, we are able to run all the available classification algorithms that WEKA provides. However, due to space limit we will below concentrate on only a few.

3 Experimental Results

3.1 Data Sets and Experiments Configurations

In our experiments we have used a corpus of tweets provided for the TASS workshop at the SEPLN 2012 conference as input data set. This set contains about 70,000 tweets. Additionaly, over 7,000 of the tweets were given as a small training set with both topic and sentiment classification. The data set was shuffled for the topics and sentiments to be randomly distributed. Due to the large time taken by the experiments with the large data set, most of the experiments presented have used the small data set, using 5,000 tweets for training and 2,000 for evaluation.

[1] http://www.cs.waikato.ac.nz/ml/weka, accessed August 2012.

Configuration number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Parameters														
n-gram	1	1	1	1	1	1	1	2	1	1	1	1	2	1
Lemma/Stem (**L**/**S**)	L	L	S	L	L	L	L	L	L	L	L	L	L	L
SMS					X						X		X	
Word types (Nouns C&P)	X		X	X	X	X	X	X	X	X			X	X
Correct words				X										
Hashtags	X	X	X	X	X		X	X	X	X	X	X		X
Author Tags	X	X	X	X	X	X		X	X		X	X	X	X
Links									X	X				
Classifiers (Accuracy)														
Ibk	36.62	30.54	36.37	36.62	36.77	31.17	**37.97**	32.64	**38.57**	32.47	30.49	30.54	33.83	36.62
ComplementNaiveBayes	56.75	**58.45**	56.25	56.75	**57**	55.75	53.66	53.56	53.56	51.67	**58.25**	**58.45**	52.02	56.75
NaiveBayesMultinomial	56.35	**57.1**	55.61	56.35	56.25	55.46	53.71	55.61	54.11	53.26	**56.95**	**57.1**	56	56.35
RandomCommittee	53.56	52.47	52.62	53.56	**53.91**	**53.66**	52.52	**55.06**	52.72	52.27	51.92	52.47	38.15	53.56
SMO	56.3	55.06	55.95	56.3	56.55	55.51	55.26	55.9	55.16	54.21	42.38	55.06	54.81	56.3

Figure 1: Accuracy (%) of different configurations for topic detection in the small data set.

Precision	Recall	F-Measure	Class
0.468	0.619	0.533	música
0.316	0.318	0.317	economía
0.565	0.503	0.532	entretenimiento
0.721	0.814	0.765	política
0.386	0.354	0.37	cine
0.175	0.241	0.203	literatura
0.551	0.442	0.491	otros
0.194	0.162	0.176	tecnología
0.419	0.5	0.456	deportes
0.5	0.409	0.45	fútbol
0.579	0.584	0.578	Weighted Avg.

Table 1: Detail of Configuration 2 of topic detection with Complement Naive Bayes.

For the TASS workshop we tested multiple configurations with all the WEKA classifiers to choose the one with the highest accuracy. Different configurations gave the best results for sentiment analysis and topic detection. As described, our initial approach was to compare every possible configuration and all classification methods of WEKA. Unfortunately, it was unfeasible to execute all possible configurations with all possible classification methods. Hence, we made some decisions to limit the number of experiments.

In this paper, we have chosen to present only five classification algorithms from those provided by WEKA. In particular, we have chosen the methods Ibk, Complement Naive Bayes, Naive Bayes Multinomial, Random Committee, and SMO. This set tries to cover the most popular classification techniques. Then, we have chosen for each of the two problems (topic and sentiment) a basic configuration similar to the one submitted to the TASS workshop. Starting from this basic configuration a sequence of derived configurations are tested. In each derived configuration, one of the parameters of the basic configuration was changed, in order to explore the effect of that parameter in the performance. Finally, for each classification method a new configuration is created and tested with the parameter settings that maximized the accuracy.

The accuracy values computed in each of the configurations with the five methods with the small data set are presented in Figures 1 and 2. In both figures, Configuration 1 is the basic configuration. The derived configurations are numbered 2 to 9. (Observe that each accuracy value that improves over the accuracy with the basic configuration is shown on boldface.) Finally, the last 5 configurations of each figure correspond to the parameters settings that gave highest accuracy in the prior configurations for a method (in the order Ibk, Complement Naive Bayes, Naive Bayes Multinomial, Random Committee, and SMO).

3.2 Topic Estimation Results

As mentioned, Figure 1 presents the accuracy results for topic detection on the small data set, under the basic configuration (Configuration 1), configurations derived from this one by toggling one by one every parameter (Configurations 2 to 9), and the seemingly best parameter settings for each classification method (Configurations 10 to 14). Observe that no configuration uses a search engine. This is because we found that the ARFF file generated iafter searching the web as described above (even for the small data set) was extremely large and the experiment

Configuration number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Parameters														
N-gram	1	1	1	1	1	1	1	1	2	2	1	1	2	1
Lemma/Stem (**L**/**S**)	L	L	L	S	L	L	L	L	L	L	L	S	S	L
Affective dictionary	X		X	X	X	X	X	X	X	X	X			X
SMS	X	X	X	X	X	X		X	X		X	X		X
Word types (Adj, Verb)	X	X		X	X	X	X	X	X	X	X			
Correct words					X								X	
Weight						X					X	X		
Negation	X	X	X	X	X	X	X		X	X	X	X		X
Classifiers (Accuracy)														
Ibk	31.32	31.32	29.78	31.32	31.32	31.32	**32.47**	31.32	**31.52**	**32.47**	31.32	28.78	29.08	29.78
ComplementNaiveBayes	30.18	29.88	17.93	28.74	30.13	**30.23**	28.49	30.18	28.74	28.49	**30.23**	16.88	**39.49**	17.93
NaiveBayesMultinomial	32.82	**32.97**	**32.97**	**33.37**	32.77	**32.87**	32.52	32.82	32.87	32.52	**32.87**	32.52	**42.38**	**32.97**
RandomCommittee	33.72	**34.16**	**38.24**	**34.61**	**34.31**	33.67	**34.41**	**34.36**	**34.01**	**34.41**	33.67	**38.34**	**38.14**	**38.24**
SMO	39.79	39.64	**41.93**	38.94	39.59	39.6	29.24	39.74	38.3	39.24	39.6	**41.38**	**41.43**	**41.93**

Figure 2: Accuracy (%) of different configurations for sentiment analysis in the small data set.

could not be completed

The first fact to be observed in Figure 1 is that Configuration 1, which is supposed to be similar to the one submitted to TASS, seems to have a better accuracy with some methods (more than 56% versus 45.24%). However, it must be noted that this accuracy has been computed with the small data set (while the value of 45.24% was obtained with the large one). A second observation is that in the derived configurations there is no parameter that by changing its setting drastically improves the accuracy. This also applies to the rightmost configurations, that combine the best collection of parameter settings. Finally, it can be observed that the largest accuracy is obtained by Configuration 2 with Complement Naive Bayes. This configuration is obtained from the basic one by simply removing the word filter that allows only nouns. Looking more closely at this combination of parameter configuration and method, we can obtain other performance parameters, presented in Table 1. The meaning of these can be found in the WEKA documentation. This combination has a 58.45% of correctly classified instances, and a relative absolute error of 54.07%.

3.3 Sentiment Estimation Results

Figure 2, on its turn, shows the accuracy computed for the basic configuration (Configuration 1), the derived configurations (2 to 9), and the best settings per classification method (10 to 14) for sentiment analysis with the small data set. As before, it can be observed that the accuracy of Configuration 1 with SMO is better than the reported accuracy of the results submitted (39.79% versus 36.04%). It also holds that no parameter seems to make a huge difference. However in this case the combination of parameters seem to have some impact, since the best combination, formed by Configuration 13 and method Naive Bayes Multinomial, has significant better accuracy than any other configuration with the same method. However, other methods (e.g., SMO) has a more homogenous set of values.

Precision	Recall	F-Measure	Class
0.368	0.285	0.321	negative+
0.354	0.43	0.389	negative
0.145	0.064	0.089	neutral
0.317	0.14	0.194	positive
0.461	0.715	0.561	positive+
0.525	0.469	0.495	none
0.404	0.424	0.4	Weighted Avg.

Table 2: Detail of Configuration 13 of sentiment analysis with Naive Bayes Multinomial.

As before, we take a closer look at the best combination in Table 2. This combination is able to classify correctly 851 instances (and incorrectly 1157), with an accuracy of 42.38%, and relative absolute error of 77.29%.

4 Conclusions

We have presented a comprehensive set of experiments classifying Spanish tweets according to sentiment and topic. In these experiments we have evaluated the use of stemmers and lemmatizers, n-grams, word types, negations, valence shifters, link processing, search engines, special Twitter semantics (hashtags), and different classification methods. This collection of techniques

represent a thorough study.

The first conclusion of our study is that none of the techniques explored is the silver bullet for Spanish tweet classification. None made a clear difference when introduced in the algorithm. The second conclusion is that tweets are very hard to deal with, mostly due to their brevity and lack of context. The results of our experiments are encouraging though, since they show that it is possible to use classical methods for analyzing Spanish texts. The largest accuracy obtained (58% for topics and 42% for sentiment) are not too far from other values reported in the TASS workshop. However, these values reflect that there is still a lot of room for improvement, justifying further efforts.

References

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In LSM '11, pp. 30–38.

Allan, James. 2002. Topic detection and tracking. Kluwer Academic Publishers.

Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. 2007. Clustering short texts using wikipedia. In SIGIR'07, pp. 787–788.

Bermingham, Adam and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *CIKM 2010*, pp. 1833–1836.

Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *RANLP 2009*.

Chang, Chih-Chung and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27.

Cruz, Fermín L, Jose A Troyano, Fernando Enriquez, and Javier Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Proc. del Lenguaje Natural*, 41:73–80.

Engström, Charlotta. 2004. Topic dependence in sentiment classification. Master thesis, University of Cambridge.

Gabrilovich, Evgeniy and Shaul Markovitch. 2005. Feature generation for text categorization using world knowledge. In IJCAI'05, pp. 1048–1053.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pp. 1–6.

Justin, T., R. Gajsek, V. Struc, and S. Dobrisek. 2010. Comparison of different classification methods for emotion recognition. In *MIPRO 2010*, pp. 700 –703.

Kukich, Karen. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439.

Laboreiro, Gustavo, Luís Sarmento, Jorge Teixeira, and Eugénio Oliveira. 2010. Tokenizing micro-blogging messages using a text classification approach. In AND '10, pp. 81–88.

Lee, K., D. Palsetia, R. Narayanan, M.M.A. Patwary, A. Agrawal, and A. Choudhary. 2011. Twitter trending topic classification. In *ICDMW 2011*, pp. 251–258.

Liu, Bing. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, 2nd Edition*, Taylor & Francis Group.

Martínez-Cámara, Eugenio, M. Martín-Valdivia, and L. Ureña-López. 2011. Opinion classification techniques applied to a spanish corpus. In *NLDB 2011*, pp. 169–176.

Mathioudakis, Michael and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *SIGMOD'10*, pp. 1155–1158.

Padró, Lluís, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In *Global Wordnet Conference 2010*, pp. 99–105.

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC'10*.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Rahimtoroghi, Elahe and Azadeh Shakery. 2011. Wikipedia-based smoothing for enhancing text clustering. In *AIRS'11*, pp. 327–339.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACLstudent'05*, pp. 43–48.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. 2012. Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets In *MICAI'12*.

Sriram, Bharath, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *SIGIR '10*, pp. 841–842.

Vakali, Athena, Maria Giatsoglou, and Stefanos Antaris. 2012. Social networking trends and dynamics detection via a cloud-based framework design. In *WWW '12 Companion*, pp. 1213–1220.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 53-60 recibido 29-11-12 revisado 21-12-12 aceptado 19-02-13

SINAI en TASS 2012

SINAI at TASS 2012

Eugenio Martínez Cámara, M. Ángel García Cumbreras, M. Teresa Martín Valdivia, L. Alfonso Ureña López
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 – Jaén
{emcamara, magc, maite, laurena}@ujaen.es

Resumen: En el presente artículo se describe la participación del grupo de investigación SINAI de la Universidad de Jaén en la primera edición del taller sobre Análisis de Sentimientos en el congreso de la SEPLN (TASS 2012). El Taller propone dos tareas, una centrada en la determinación de la polaridad de *tweets* en español, y una segunda en la que hay que identificar los temas a los que pertenecen los *tweets*. Para la primera tarea se ha optado por una estrategia de aprendizaje automático supervisado, siendo SVM el algoritmo elegido. En cuanto a la segunda tarea, también se ha utilizado SVM, y con el fin de mejorar el resultado de la clasificación se ha combinado con bolsas de palabras de cada uno de los temas.
Palabras clave: Twitter, Análisis de Sentimientos, Análisis de la Opinión, método supervisado, SVM.

Abstract: In this paper is described the participation of the SINAI research group of the University of Jaén in the first edition of the workshop on Sentiment Analysis at the SEPLN congress (TASS 2012). The Workshop includes two tasks, the first one is focused in the polarity classification of a corpus of Spanish tweets, and the second one involves a topic classification. For the first task, we have chosen a supervised machine learning approach, in which we have used SVM for classifying the polarity. In the second task, we have also used SVM for the topic classification but several bags of words have been used with the goal of improving the classification performance.
Keywords: Twitter, Sentiment Analysis, Opinion Mining, Supervised Machine Learning, SVM

1 Introducción

En este artículo se presentan los experimentos y resultados obtenidos en el Taller de Análisis de Sentimientos en la SEPLN (TASS 2012) (Villena-Román et al., 2013). Concretamente se ha participado en las dos tareas propuestas: Sentiment Analysis y Trending Topic Coverage.

El Análisis de Sentimientos (AS), también conocido como Minería de Opiniones (MO) o Análisis de Opiniones (AO), se ha convertido en una prometedora disciplina de investigación que se encuadra dentro del Procesamiento del Lenguaje Natural (PLN) y la Minería de Datos. Se suele definir como el tratamiento computacional de la información subjetiva presente en cualquier tipo de documento (Pang and Lee, 2008).

La proliferación de contenidos web generados por los propios usuarios en blogs, wikis, foros o redes sociales ha motivado que empresas, investigadores y organizaciones se interesen por analizar y monitorizar toda esta información que circula por la red. Es por ello, que cada vez en más foros se presentan artículos científicos, conferencias o proyectos relacionados con el AO.

Por otra parte, el AO se ha tratado fundamentalmente sobre textos extensos como por ejemplo documentos en blogs o artículos de opinión. Sin embargo, debido al enorme éxito de las redes sociales, el interés para analizar las opiniones en textos cortos

ISSN 1135-5948

está creciendo de manera exponencial.

Aunque se trata de un área de investigación relativamente nueva, existen una gran cantidad de trabajos relacionados con AO, y más específicamente con la clasificación de la polaridad. Se pueden distinguir dos formas de tratar el problema. La primera se basa en técnicas de aprendizaje automático (Pang et al., 2002), y la segunda se fundamenta en el concepto de orientación semántica, que no necesita el entrenamiento de ningún algoritmo, pero sí debe tener en cuenta la orientación de las palabras (positiva o negativa) (Turney, 2002). En este trabajo se intentan combinar ambas aproximaciones con el fin de mejorar la precisión de los sistemas.

Además, la mayoría de las investigaciones se han centrado en textos en inglés aunque está claro que cada día más, otros idiomas como el chino o el español están más presentes en internet. Precisamente el taller que es la base de este artículo utiliza un corpus en español.

El resto del artículo se organiza como sigue: la siguiente sección incluye una pequeña revisión del estado del arte sobre AO en Twitter. A continuación, se presenta una breve introducción a la categorización de texto en Twitter. La sección 4 muestra el corpus y el proceso de preparación de los datos. En las secciones 5 y 6 se presenta la experimentación y resultados en las dos tareas abordadas.

2 Análisis de la opinión en Twitter

Cada día más, los usuarios de Internet utilizan las redes sociales para expresar sus sentimientos y sensaciones sobre cualquier tema. Un ejemplo de esto son las redes de micro-blogging, como Twitter, en la que en tiempo real los usuarios expresan sus opiniones sobre los temas más variados. En España la presencia de Twitter ha ido creciendo paulatinamente, y es a partir de 2010 cuando empresas, políticos y usuarios en general se están dando cuenta del verdadero potencial de esta red social. Sería de gran utilidad el poder determinar de forma automática la polaridad de esas opiniones permitiendo desarrollar sistemas que se encarguen de estudiar y analizar la intención de voto de los ciudadanos, la opinión de consumidores sobre algún producto o servicio concreto o el estado de ánimo de las personas.

Los *tweets* tienen características que los hacen diferentes de las opiniones y comentarios que hay en foros y páginas web. Normalmente los comentarios u opiniones que se escriben en Internet suelen ser textos más o menos extensos en los que los usuarios intentan resumir lo que piensan sobre un determinado tema, pero los *tweets* suelen estar escritos en un lenguaje informal, y su extensión está limitada a 140 caracteres. Por otra parte, muchos *tweets* no expresan opiniones sino situaciones que les ocurren a los usuarios. Por último, los *tweets* tienen que ser analizados a nivel de frase, y no a nivel de documento.

La mayor parte de la investigación publicada sobre AO en Twitter se ha realizado sobre *tweets* en inglés debido a que la explosión de popularidad de esta red social es relativamente reciente en países de habla no inglesa. Existen, por ejemplo, algunos trabajos que utilizan Twitter a modo de corpus. Petrovic et al. (2010) crean un gran corpus con 97 millones de *tweets*. Pak and Paroubek (2010) describen cómo generar de forma automática un corpus de *tweets* positivos, negativos y neutros. El corpus que crean lo utilizan para entrenar un clasificador de sentimientos. Go et al. (2009) usan técnicas de aprendizaje automático para construir un clasificador que les permita determinar la polaridad de los *tweets*. Para el etiquetado del corpus en *tweets* negativos y positivos siguen la misma estrategia que se describe en (Read, 2005).

Jansen et al. (2009) demuestran como los sitios de micro-blogging son una herramienta muy útil en marketing, e indica que los *tweets* pueden considerarse como Electronic Word Of Mouth (EWOM). Siguiendo esta línea de utilizar Twitter como otra herramienta más en marketing, en (Asur and Huberman, 2010) se utiliza un corpus de *tweets* sobre un conjunto de películas que se estrenaron a finales de 2009 y principios de 2010, para demostrar la correlación existente entre la cantidad de *tweets* y su polaridad sobre una determinada película, con la recaudación en taquilla que ha obtenido en las dos primeras semanas desde su estreno. Bollen et al. (2011) investigan la posible correlación del estado de ánimo que se manifiesta en Twitter con la variación de los

mercados de valores.

El estudio de la tendencia de la opinión política también ha sido un tema que ha atraído a la comunidad científica. O'Connor et al. (2010) analizan opiniones políticas y sobre productos comerciales y comparan los resultados usando Twitter y encuestas. Diakopoulos and Shamma (2010) clasifican la polaridad de los *tweets* durante el debate presidencial en los Estados Unidos del año 2008.

3 Categorización de texto en Twitter

La categorización de textos es una tarea que consiste en la clasificación de los textos en distintas clases predefinidas. Algunas publicaciones han estudiado este problema sobre *tweets*, por ejemplo Sriram et al. (2010) proponen una aproximación que categoriza *tweets* dependiendo del texto que contienen en un conjunto predefinido de clases genéricas como noticias, eventos, opiniones, tratos o mensajes privados. Por su parte Garcia et al. (2010) realizan una comparación entre dos redes sociales (Blippr y Twitter) clasificando en las categorías Movies, Books, Music, Apps y Games. Los resultados para ambas redes sociales en cada una de las categorías son muy parecidos. Por último, distintas técnicas de reconocimiento de entidades (NER) son aplicadas en el trabajo de (Jung, 2011) para categorizar el texto en redes sociales.

4 Preparación de los datos

La organización del taller proporcionó dos conjuntos de datos, uno de entrenamiento y otro para el test. En total más de 70.000 tweets en español sobre unas 200 personas conocidas, de diversos ámbitos como política, economía, comunicación o cultura. En (Villena-Román et al., 2013) se puede encontrar una descripción detallada del corpus propuesto.

Antes de aplicar los datos al clasificador de la polaridad, y posteriormente al clasificador de categorías, se le ha aplicado al conjunto de *tweets* un proceso de limpieza, con la intención de reducir la mayor cantidad de ruido posible. Dicho proceso está formado por:

1. Eliminación de caracteres que no sean letras del alfabeto español o números. La eliminación de los signos de exclamación se produce posteriormente a su tratamiento en aquellos experimentos en los que se ha estudiado su influencia.
2. Se ha llevado a cabo una normalización de las expresiones de risa, de manera que todas ellas se encuentren representadas por una misma expresión.
3. También se han normalizado las palabras que tienen letras repetidas. El proceso ha consistido en reducir a dos repeticiones toda letra que estuviera repetida más de tres ocasiones, de manera que se considera distinta a la palabra original, pero de la misma forma independientemente del número de repeticiones.

Una vez realizado este *preprocesado*, los datos ya están preparados para su aplicación al clasificador tanto de polaridad como de temas.

5 Tarea 1: Clasificación de la polaridad

La primera tarea consiste en el desarrollo de un sistema de clasificación de *tweets* en español en cinco niveles de polaridad: NONE, N+, N, NEU, P, P+, siendo NONE la etiqueta utilizada para aquellos tweets que no tienen sentimientos, no son subjetivos, y NEU para los neutros. En (Villena-Román et al., 2013) se describen con mayor detalle cada una de las categorías de opinión.

Para resolver el problema planteado se decidió seguir una estrategia basada en aprendizaje automático supervisado. Primeramente se llevó a cabo un proceso evaluación del clasificador. El algoritmo de clasificación elegido fue SVM (*Support Vector Machines*) (Vapnik, 1995), y más concretamente la implementación *SVMLight*[1]. La elección de SVM se ha fundamentado en los buenos resultados que suele ofrecer en los trabajos de AO, pudiéndose consultar muchos de ellos en (Pang and Lee, 2008). Además, SVM también ha sido utilizado con éxito por nuestro equipo en varios trabajos de AO (Martínez-Cámara et al., 2011a), (Martínez-Cámara et al., 2011b).

Una vez que se han limpiado los datos tal

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

y como se ha descrito en la sección 4 , se diseñaron un conjunto de 33 experimentos, que van desde el caso base en el que solamente se *tokenizan* los *tweets*, hasta configuraciones a los que se añaden número de palabras positivas, negativas que aparecen en los *tweets*. Las características que se han utilizado para evaluar la configuración del clasificador han sido:

1. ***Unigramas***: Cada *tweet* se tokeniza, y se utiliza la métrica TF para representar cada *unigrama*. TF se refiera a la frecuencia relativa de cada *unigrama* en el *tweet*. La elección de TF, y no por ejemplo TF-IDF, se basa en trabajos previos en el taller en el que siempre los mejores resultados se han obtenido con TF.
2. **Emoticonos**: Se añade como característica el número de emoticonos positivos o negativos que aparecen en el *tweet*. Para ello se ha utilizado una bolsa de emoticonos positivos y negativos, más concretamente los que aparecen en la Tabla 1.
3. **Palabras positivas y negativas**: En algunas de las evaluaciones que se han llevado a cabo se han incluido como características el número de palabras positivas y negativas que aparecen en los *tweets*. Esto se ha realizado siguiendo un enfoque basado en bolsa de palabras. En español no se ha encontrado ninguna bolsa de palabras positivas y negativas, por lo que se decidió traducir automáticamente la presentada en (Hu and Liu, 2004), y que se puede descargar desde la web[2] de uno de sus autores.
4. **Intensidad**: Aprovechando la bolsa de palabras, se decidió añadir como característica adicional el número de palabras positivas y negativas con caracteres repetidos, de manera que se pudiera modelizar la intensidad de la opinión o emoción que el autor quiere expresar.

   En otro conjunto de experimentos las palabras indicadoras de opinión y que contaban con caracteres repetidos no se les da ningún tratamiento especial, mientras que otro conjunto de evaluaciones del método al contador de palabras positivas y negativas se le añadía una unidad más en el caso de que el término tuviera letras repetidas. La intensidad puede ser expresada mediante signos de exclamación, por lo que en otro conjunto de experimentos se aumentaba con una unidad adicional el contador de palabras positivas o negativas en el caso de fueran acompañadas por un signo de exclamación.

   También se ha experimentado con el efecto de la presencia de partículas negativas en el *tweet*. La presencia de una partícula negativa delante de alguna palabra indicadora de opinión, hacía que su aportación se sumara al contador opuesto a la categoría que pertenece, es decir, que si una palabra positiva va acompañada de un elemento negativo, su aportación se registra en el contador de palabras negativas, en lugar del de positivas.

Emo. Positivos	:) :) :-) ;) ;-)=) ^_^ :-D :D :d =D C: Xd XD xD (x (= ^^ ^o^ 'u' n_n *-* *O* *o* *_*
Emo. Negativos	:-(:(:((: (D: Dx 'n' :\ /:):-/ :' ='[:_(/T_T TOT ;_;

Tabla 1 Emoticonos positivos y negativos

En las tareas de aprendizaje automático relacionadas con procesamiento de texto se suele utilizar *stopper* y *stemmer* para reducir el número de características léxicas. Para los experimentos que se han llevado a cabo para la participación en el taller, solo se ha probado la eficacia de la aplicación de *stemmer*, ya que en investigaciones previas se ha llegado a la conclusión que para la tarea de Análisis de la Opinión en Twitter, el uso de *stopper* es contraproducente.

El procedimiento de evaluación del clasificador elegido ha sido el de *K-Cross-Validation* con un valor de *K*=10.

La evaluación de la combinación de todas las características anteriores originó 33 experimentos, de los cuales se escogieron cuatro para presentarlos al taller.

Los resultados que se obtuvieron en el proceso de evaluación del clasificador se pueden ver en la Figura 1. De todas esas configuraciones del clasificador se seleccionaron los cuatro con mayor valor de

[2] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

F1. Las características de esas configuraciones son:

1. ***EXP 2***: En este caso cada *tweet* se representa como un conjunto de *unigramas*, cuya importancia es indicada por su valor de TF. Antes de su clasificación se le aplica un proceso de *stemmer*.
2. ***EXP 6***: Lo mismo que el anterior pero además se normalizan las direcciones web, de manera que cada url se sustituye por URL. En este caso también se normalizan las menciones, por lo que cada expresión de la forma *@nombre_usuario*, se sustituye por MENTION.
3. ***EXP 4***: Igual que en el experimento pol-sinai-2-5l, pero en esta ocasión solo se normalizan las menciones.
4. ***EXP 19***: En este experimento, además de las características léxicas incluidas en las tres anteriores configuraciones, se incluyen como características el número de emoticonos positivos y negativos, el número de palabras positivas y negativas, y en el caso de que alguna de esos términos tengan caracteres repetidos se considera dicha palabra como doble.

Los mejores resultados se han obtenido con las configuraciones del sistema más simples. Únicamente el experimento EXP 19 es el que incluye características que indican una cierta información semántica, como es el número de emoticonos y palabras positivas y negativas.

El que no haya funcionado el enfoque basado en bolsa de palabras positivas y negativas, como se pensaba en un principio, se cree que es debido a la calidad de la traducción del listado de términos original, y a las lógicas diferencias entre expresiones que en inglés pueden ser indicativas de una opinión, y en español no. Esto indica, que tiene que haber una mejora de los recursos lingüísticos en español relacionados con el AO, y mientras tanto seguir investigando en el uso de otros recursos de mayor calidad en inglés.

A continuación se presentan los resultados obtenidos con los datos de *test*, al mismo tiempo que se asocia el nombre del experimento con el nombre asignado por los organizadores del taller. La configuración *EXP-19* (*pol-sinai-4-5l*) fue la que mejores resultados obtuvo (Tabla 2).

Experimento	Precisión
P	0,94%
P+	60,99%
NEU	0,38%
N	33,61%
N+	32,00%
NONE	71.52%
Total	**54,68%**

Tabla 2: Resultados del experimento *pol-sinai-4-5l*

Los mayores errores se encuentran entre clases cercanas, como son NONE y NEU, y entre P y P+. El error entre las clases NONE y NEU es muy probable que sea debido a la diferencia tan grande entre el número de *tweets* NEU y NONE, por lo que seguramente el clasificador ha sobre-entrenado la clase NONE. El error en las clases positivas es muy probable que esté debido al conjunto de palabras positivas utilizadas, así como a la asignación de una doble importancia a aquellas palabras positivas con letras repetidas.

En la Tabla 3 se muestran los resultados para la configuración correspondiente al experimento *EXP 6* (*pol-sinai-2-5l*). En este caso, al no haber ninguna característica semántica, el sobre-ajuste sobre la clase NONE es mucho más evidente.

La Tabla 4 contiene los resultados del experimento *EXP 2* (*pol-sinai-1-5l*). En este caso existe un mayor sobre-ajuste sobre la clase NONE. Además, al tratarse de una configuración en la que solo se tiene en cuenta las características léxicas, se puede decir que las únicas clases con un vocabulario más determinante son P y P+, ya

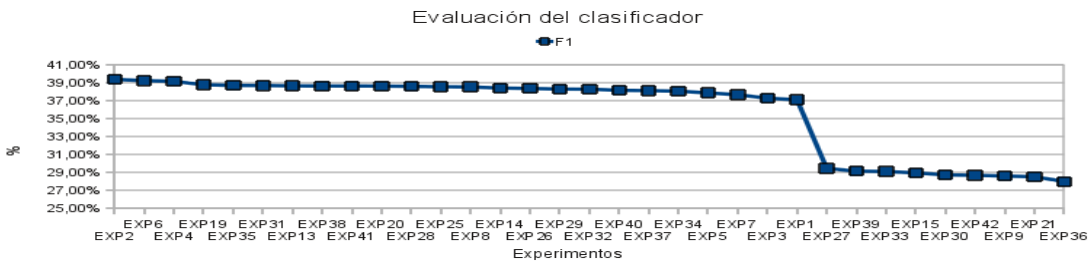

Figura 1: Valores F1 obtenidos durante el proceso de evaluación del clasificador

que son las únicas en las que unos pocos *tweets* se han clasificado correctamente.

En la Tabla 5 se muestran los resultados de la configuración del experimento *EXP 4* (*pol-sinai-3-5l*). Este experimento tiene un comportamiento similar al experimento *pol-sinai-1-5l*, pero en esta ocasión sí se consigue clasificar algún *tweet* de las clases N y N+. Esto puede ser debido a la no normalización de las direcciones web, que haya favorecido a esas clases, y haya perjudicado algo la clasificación de la clase NONE.

Experimento	Precisión
P	0,27%
P+	4,69%
NEU	0%
N	0,12%
N+	0,37%
NONE	96,48%
Total	**35,65%**

Tabla 3: Resultados del experimento *pol-sinai-2-5l*

Experimento	Precisión
P	0,87%
P+	0,81%
NEU	0%
N	0%
N+	0%
NONE	99,31%
Total	**35,28%**

Tabla 4: Resultados del experimento *pol-sinai-1-5l*

6 Tarea 2: Categorización de temas

En esta tarea se evalúa el rendimiento de un clasificador que identifique correctamente el *topic* de un *tweet*, y en función de ese *topic* analice la polaridad de dicho *tweet*. La evaluación se realiza conforme a las mismas métricas definidas en la tarea anterior (precisión, *recall* y F1).

Experimento	Precisión
P	0,47%
P+	0,58%
NEU	0%
N	0,01%
N+	0,04%
NONE	98,65%
Total	**34,97%**

Tabla 5: Resultados del experimento *pol-sinai-3-5l*

Para esta tarea se ha desarrollado un sistema de clasificación de temas que tomando como base las categorías de los *tweets* de entrenamiento, categorice correctamente un nuevo *tweet*. Las categorías identificadas son: cine, deportes, economía, entretenimiento, fútbol, literatura, música, otros, política y tecnología.

El sistema de clasificación utilizado para los experimentos oficiales es un sistema de aprendizaje automático basado en SVM, que toma diversas características para el entrenamiento. Como características a procesar se han utilizado los *unigramas* de los *tweets*, procesados y sin procesar. Se realizaron experimentos previos con otros sistemas de aprendizaje automático, siendo SVM el que obtuvo los mejores resultados. El procesado de los *tweets* es el mismo descrito en la tarea 1.

Además, se han generado dos bolsas de palabras relevantes para cada categoría para mejorar el rendimiento del sistema. La primera bolsa de palabras se ha obtenido a partir de Google AdWordsKeyWordTool[3], que permite introducir un término y devuelve las n ideas directamente relacionadas. La segunda bolsa de palabras se ha obtenido a partir de los *hashtags* de los *tweets* de entrenamiento, tomando para cada categoría los *hashtags* que únicamente aparecen en los *tweets* de dicha categoría.

En función del procesado o no de los *tweets* y las bolsas de palabras, se realizaron diversos experimentos con el conjunto de *tweets* de entrenamiento y evaluación cruzada (*10-fold cross validation*). Según los resultados previos obtenidos se presentaron los siguientes experimentos oficiales:

- **Top-sinai-1** (caso base). Los *tweets* no se han procesado ni se utilizan bolsas de palabras.
- **Top-sinai-2**. Los *tweets* se han procesado pero no se aplica ni *stopper* ni *stemmer*. No se utilizan bolsas de palabras.
- **Top-sinai-3**. Los *tweets* se han procesado y se aplica *stopper* y *stemmer*. No se utilizan bolsas de palabras.
- **Top-sinai-4**. Los *tweets* se han procesado y se aplica *stopper* y

[3] Disponible en https://adwords.google.com/o/KeywordTool

stemmer. Como bolsas de palabras se utilizan los *hashtags*, y se añaden a cada *tweet* en entrenamiento, dependiendo de su topic.

- **Top-sinai-5**. Los *tweets* se han procesado y se aplica *stopper* y *stemmer*. Como bolsas de palabras se utilizan los *hashtags* y las palabras de Adwords, y se añaden a cada *tweet* de entrenamiento, dependiendo de su *topic*.

Los resultados obtenidos con estos experimentos se muestran en la Tabla 6.

Analizando las categorías etiquetadas automáticamente por nuestro sistema observamos que la mayoría de los *tweets* de evaluación han sido etiquetados en las categorías "otros", "política" y "entretenimiento". Por este motivo se han generado estadísticas de etiquetado de la colección de entrenamiento, obteniendo los datos que se pueden ver en la Tabla 7.

Run id	Precisión
top-sinai-1	32.34%
top-sinai-2	34.76%
top-sinai-3	34.06%
top-sinai-4	37.79%
top-sinai-5	39.37%

Tabla 6: Resultados oficiales de la tarea 2

Observando estos resultados no es difícil concluir que la distribución de categorías está muy desbalanceada para un sistema de aprendizaje automático al uso, sin contar con información adicional de la categoría a la hora de entrenar, ya que, por ejemplo, con un subconjunto de entrenamiento del 1,11% (literatura) resultará imposible que un *tweet* de evaluación lo clasifique en dicha categoría.

Se está realizando un análisis más profundo de los resultados con el fin de obtener más conclusiones y trabajo a realizar para mejorar el sistema, aunque es muy probable que la mejora del clasificador pase por entrenar diferentes categorías generales y específicas con otro material externo.

7 Conclusiones y trabajo futuro

La experimentación que se ha presentado pone de manifiesto que la inclusión de características semánticas al conjunto de características ayuda al proceso de la clasificación de la polaridad. Por otro lado, también es muy importante la limpieza y normalización de los datos, como demuestran los resultados *pol-sinai-2-5l, pol-sinai-1-5l y pol-sinai-3-5l*.

Categoría	# tweets	%
cine	183	2,53%
deportes	101	1,40%
economía	525	7,27%
entretenimiento	1.209	16,75%
fútbol	225	3,12%
literatura	80	1,11%
música	411	5,69%
política	2.715	37,61%
otros	1.625	22,51%
tecnología	145	2,01%
TOTAL:	**7.219**	**100%**

Tabla 7: Nº de *tweets* por categoría

En cuanto a la tarea de categorización de temas, la incorporación al proceso de información de cada categoría mejora considerablemente los resultados.

Actualmente se está trabajando en el uso de diversos recursos semánticos, de listas de palabras, de aprovechamiento de la información del contexto y del tratamiento de la negación para mejorar la clasificación de la polaridad. Para mejorar la categorización, se está apostando por mejorar el enfoque basado en bolsa de palabras combinándolo con un sistema supervisado cuyo modelo se fuera actualizando periódicamente.

Agradecimientos

Esta investigación ha sido subvencionada parcialmente por el Fondo Europeo de Desarrollo Regional (FEDER), a través del proyecto TEXT-COOL 2.0 (TIN2009-13391-C04-02) y el proyecto ATTOS (TIN2012-38536-C03-0) por el gobierno español, y por la Comisión Europea bajo el Séptimo programa Marco (FP7 - 2007-2013) a través del proyecto FIRST (FP7-287607).

Bibliografía

Asur, Sitaram, Huberman, Bernardo A. (2010). Predicting the Future with Social Media. 2010 IEEE/WIC/ACM International Conf. on Web Intelligence and Intelligent Agent Technology, 1, pp.492-499.

Bollen, J. Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 21, pp. 1-8.

Diakopoulos, N. A. and D. A. Shamma. (2010). Characterizing debate performance via aggregated twitter sentiment. CHI ’10: Proc. of the 28th International Conf. on Human Factors in Computing Systems. New York, NY, USA. ACM. pp 1195–1198.

Garcia, S. O'Mahony, M.P., Smyth, B. Towards tagging and categorization for micro-blog. 21st National Conf. on Artificial Intelligence and Cognitive Science (AICS 2010), Galway, Ireland.

Go, A., R. Bhayani, and L. Huang. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford Digital Library Technologies Project

Hu, Minqing, Liu, Bing. (2004). Mining and Summarizing Customer Reviews. Proc. of the ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (KDD-2004). Seattle, Washington, USA.

Jansen, B., M. Zhang, K. Sobel, and A. Chowdury (2009). Twitter power:tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology.

Jung, J.J. Towards Named Entity Recognition Method for Microtexts in Online Social Networks: A Case Study of Twitter, International Conf. on Advances in Social Networks Analysis and Mining, pp.563-564

Martínez-Cámara E., Martín-Valdivia M. T., Perea-Ortega, J. M., Ureña-López, L. A. (2011b). Técnicas de clasificación de opinions aplicadas a un corpus en español. Procesamiento de Lenguaje Natural. 47, pp. 163-170.

Martínez-Cámara E., Martín-Valdivia M. T., Ureña-López, L. A. (2011a). Opinion classification techniques applied to a Spanish corpus. Natural Language Processing and Information Systems. Springer, pp 169-176.

O’Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith (2010). From Tweets to polls: Linking text sentiment to public opinion time series. International AAAI Conf. on Weblogs and Social Media, Washington, D.C.

Pak, A., P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proc. of the Seventh Conf. on International Language Resources and Evaluation (LREC’10), (ELRA), Valletta, Malta, pp. 19–21.

Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundation and Trends in Inf. Retrieval 2(1-2) 1-135

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL. pp. 79–86.

Petrovic, S., Osborne, M., Lavrenko, V. (2010). The Edinburgh Twitter corpus. SocialMedia Workshop: Computational Linguistics in a World of Social Media, pp. 25–26.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proc. of the ACL Student Research Workshop, pp. 43–48.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. 2010. Short text classification in Twitter to improve information filtering. In Proc. of the 33rd International ACM SIGIR Conf. on Research and development in Inf. Retrieval (SIGIR '10). ACM

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proc. of the 40th Annual Meeting on ACL. Morristown, NJ, USA. pp. 417–424.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Villena-Román, J. Lana-Serrano, S. González-Cristóbal, J.C. Martínez-Cámara, E. (2013). TASS – Workshop on Sentiment Analysis at SEPLN. Procesamiento de Lenguaje Natural, 50.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 61-68 recibido 30-11-12 revisado 07-02-13 aceptado 19-02-13

Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque

Transformación de las oraciones compuestas utilizando árboles de dependencias para la Simplificación Automática de Textos en Euskera

María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios
IXA NLP Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1 48014 Donostia
maxux.aranzabe@ehu.es, a.diazdeilarraza@ehu.es, igonzalez010@ikasle.ehu.es

Resumen: En este artículo se presenta uno de los módulos que forma parte del sistema de simplificación automática de textos escritos en euskera que se está implementando. Concretamente, se describe el módulo donde se lleva a cabo la transformación de las oraciones compuestas en oraciones simples. Esta transformación se realiza mediante las herramientas de alta precisión y cobertura general desarrolladas para el tratamiento automático del euskera. Además de adaptar y enriquecer el identificador de oraciones se ha implementado un algoritmo basado en árboles de dependencias sintácticas cuyo objetivo es dividir las oraciones complejas en oraciones más simples.
Palabras clave: Simplificación automática de textos, división de oraciones, euskera, identificación de las oraciones compuestas y simples

Abstract: In this paper we present a module of the Text Simplification architecture that we are implementing. Exactly, we describe the module that carries out the task of splitting sentences into clauses. This module is based on general-coverage tools. We have adapted the clause identifier in this module and we have added a algorithm based on dependency-trees to split the sentences. This way, we get simple sentences.
Keywords: Text Simplification, sentence splitting, Basque, clause boundary identification

1 Introduction

Automatic Text Simplification (TS) is a Natural Language Processing (NLP) task that aims the transformation of difficult texts to get a equivalent simple text. This may involve simplifyng syntactic phenomena, performing operations like sentence spliting, changing passive to active voice, inverting the order of the clauses, changing discourse marker by a simpler and/or more frequent one. As a result, this new text should easier to understand for humans and/or easier to process by NLP advanced applications and it should keep the meaning of original text, or at least information loss should be avoid.

TS systems and architectures have been proposed for languages like English (Siddharthan, 2006), Brazilian Portuguese (Candido et al., 2009), Swedish (Rybing, Smith, and Silvervarg, 2010), Japanese (Inui et al., 2003), Arabic (Al-Subaihin and Al-Khalifa, 2011), Spanish (Saggion et al., 2011), and French (Seretan, 2012). As method, dependency trees have been used in TS systems like (Zhu, Bernhard, and Gurevych, 2010) and (Siddharthan, 2011) among others.

The target audiences of the TS systems have been people with disabilities (Carroll et al., 1999), illiterate (Candido et al., 2009) or people who learn foreign languages (Petersen and Ostendorf, 2007) (Burstein, 2009) among others. There are TS system for NLP advanced applications such us machine translation (Poornima et al., 2011), Q&A systems (Bernhard et al., 2012), information extraction system (Jonnalagadda and Gonzalez, 2010), and so on.

One of the operations in TS is sentence splitting. In fact, it is a compulsory need to find precise splitting points in order to continue the next operations in the TS task. In this study we analyse two linguistic diverse structures in Basque like relative clauses and adverbial temporal clauses in other to evaluate how accurate our tools are. Besides, we implement and algorithm to create simple

ISSN 1135-5948

sentences out of a complex one. Although we get simple sentences, the simplification process is not achieved: complementisers and suffixes should be removed in other to get grammatically correct sentences.

This paper is strucutured as follows: In section 2 we describe the phenomena we have treated in this paper, namely relative clauses (subsection 2.1) and temporal adverbial clauses (subsection 2.2). In section 3 we describe the simplification process we follow together with our system architecture. In section 4 we explain how we transform the trees. After that in section 5 we present the evaluation. The conclusion and future work are presented in section 6.

2 Treated Phenomena

In order to make a deep analysis of the clause boundary identifier implemented in the splitting module we explain the two phenomena we have focused on: relative clauses and adverbial temporal clauses. We selected relative clauses since they are attached to a noun and on the other hand, adverbial temporal clauses have been chosen because they show varied structures.

The corpus that has been used for this task has been EPEC (*Euskararen Prozesamendurako Erreferentzia Corpusa*-Reference Corpus for the Processing of Basque). EPEC corpus contains 300,000 words written in Standard Basque and it is tagged at morphological and syntactical levels (dependency-trees) (Aduriz et al., 2006a). At semantic level the most frequent nouns have been tagged with their corresponding synset in EusWordNet and EusSemcor (Agirre et al., 2006). Besides, the instances of the most frequent verbs have been tagged with their thematic roles in (Aldezabal et al., 2010). At the pragmatic level, discourse markers (Iruskieta, Díaz de Ilarraza, and Lersundi, 2011) and coreference (Soraluze et al., 2012) are also tagged.

We will see in next sections examples illustrating the treated phenomena. We will only show the relevant morphological information in the glosses.

2.1 Relative clauses

Basque uses gapping as strategy for relativisation, which is marked as PRO[1]. Basque relative clause can be built with finite verbs (1) using the complementiser (comp) *-(e)n* and with non finite verbs (2), attaching to the participle the suffixes *(-ta/da, -ik, -i)* + *-ko* (rel). Let us see some examples where the relative clause is marked between brackets in the examples.

(1) *Horixe zen (magoak eta nik*
That was magician and I
genuen) sekretua.
had-COMP secret.

'That was the secret the magician and me shared.'

(2) *(Bildutako diruarekin,)*
CollectREL money-SOZ,
Afganistanerako hegazkin-txartela
Afghanistan-ALL plane-ticket
erosi zitzaion Pepitari.
buy aux Pepita-DAT

'With the collected money, a plane-ticket to Afghanistan was bought to Pepita.'

The location of finite relative clauses and non finite verb relative clauses within the sentence is at the left side of the antecedent. The subordinate verb is at the end of the relative sentence.

2.2 Adverbial temporal clauses

Adverbial temporal clauses are adjuncts that specify chronological ordering (anteriority, posteriority, simultaneity, delimitation, impendency and duration) having the reference of a main verb/clause. Temporal clauses constitute a heterogeneous group, not only semantically but syntactically too. They can be built with finite verbs and non finite verbs. In both cases free elements can be added.

Finite verb temporal clauses are headed by complementisers and suffixes are attached to verb (V) like *zu#-(e)n*V.COMP *#an*-INE in example (3). In some cases like (4) a free element (*bitartean*) is added after the verb with the complementiser. Let us see these examples, where the temporal clause is marked between brackets.

[1]Phonetically null but syntactically active element

(3) *(Jontxu ikusi zuenean,)* *laster*
Jontxu see aux-COMP.INE, soon
ezagutu zuen.
recognise aux
'When s/he saw Jontxu, s/he recognised him soon.'

(4) *(Indarrean egon den*
force-INE be aux-COMP
bitartean) ez du mugapenik
meanwhile not aux delimitation
izan
be
'While it has been in force, it had no delimitation.'

Non finite verb temporal clauses are formed on the basis of the verbal noun (VN) or participle. After that suffixes are added like the inessive (INE) in *itzultze*-VN#*an*-INE from (5) example. Free elements like *ostean* in (6) can be added after the verb.

(5) *(Etxera itzultzean,)*
Home-ALL come_back-INE,
Annikak makinaz pasatzen
Annika-ERG machine-INS pass
zuen testua.
aux text-ABS
'At coming back home, Annika used to type the text'.

(6) *(Maistrak agindutakoa egin*
Teacher-ERG order-REL.ABS do
ostean,) arratsalde osoa zeukaten
after, afternoon whole had
jolasteko (...)
play-FINAL CLAUSE
'After having done what the teacher ordered, they had all the afternoon to play.'

Contrary to relative clauses, the subordinate verb does not need to be always in the last position, so we can find arguments or adjuncts after it. This canonical word order alteration is difficult too for a rule based chunker, above all if there are more than one element after the verb and no punctuation marks, that could help us by giving a clue.

3 Simplification process and system architecture

In this section we present the simplification process we follow and the architecture of the system (see figure 1) we are implementing to perform the simplification process.

The simplification process illustrates the operations that should be done and the steps we follow in order to produce simple sentences out of long sentences. Before this process is initiated, the readability of the text is analysed. This task is performed by *Idazlanen Autoebaloaziorako Sistema (IAS)*[2] module (Castro-Castro et al., 2008), a system already developed by our group for the auto-evaluation of essays, which discriminates the texts that should continue the process.

Having as input a complex text, following operations are performed:

1. **Splitting:** Make as many new sentences as clauses out of the original. This operation is performed by *Mugak* (Arrieta, 2010).
2. **Reconstruction:** Two operations take place in the split sentences:
   (a) Removing no longer needed morphological features like complementisers and suffixes. Being Basque an agglutinative language we have to remove parts of words and not a whole word.
   (b) Adding new elements like adverbs or paraphrases. The goal is to maintain the meaning. In other words, the features that have been deleted should be replaced by new words. This is included in DAR (Deletion and Addition Rules) module.
3. **Reordering:** Reorder the elements in the new sentences, and ordering the sentences in the text. The set of these rules is included in ReordR (Reordering Rules) module.
4. **Adequation and Correction:** Correct the possible grammar and spelling mistakes, and fix punctuation and capitalisation. The spell checker for Basque Xuxen (Agirre et al., 1992) will carry put this operation.

[2]System for auto-evaluation of essays

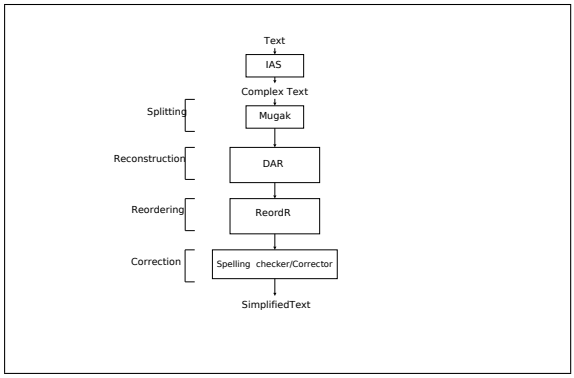

Figure 1: The architecture of system

The work presented in this paper is framed in the splitting operation and at the same time it guides the sentences to the reconstruction operation.

4 Transformation of complex sentences

Our splitting module is based on two stages: first, we apply a grammar that tags the splitting point, that is, the clause boundary is marked, and secondly, we apply an algorithm to make dependency-trees of the clauses out of the original sentence.

4.1 Splitting Point Tagging

The task of splitting point tagging is made by *Mugak* following the Constraint Grammar (CG) (Karlsson et al., 1995) formalism.

Mugak works on the basis of the output produced by several tools implemented in our group: Morpho-syntactic analysis by *Morpheus* (Alegria et al., 2002), lemmatisation and syntactic function identification by *Eustagger* (Aduriz et al., 2003), multi-words items identification (Ezeiza, 2002) (Urizar, 2012) and named entity recognition by *Eihera* (Alegria et al., 2003).

Our work consists on improving the grammar in *Mugak* (Ondarra, 2003) (Aduriz et al., 2006b) by means of adding new rules and adapting older rules based on linguistic knowledge, that lead us to get better results.

In this moment there are 78 rules and 22 of them are especially written for the phenomena we are presenting in this paper. Major improvements have been made this time in the detection of clauses headed by compound verbs and the comma. We have to remark that this is an ongoing work, that is optimised by using new corpora to find new structures and above all to determine the precision in case of non canonical order sentences.

4.2 Splitting algorithm

We have implemented an algorithm to apply several heuristics defined to transform a complex sentence into simple sentences, once the splitting point has been tagged. The usage of this algorithm is to create the dependency-trees of the new sentences. To create this algorithm and to help the following reconstruction step, we have carried out an experiment with sentences in EPEC-DEP (Basque Dependency Treebank) (Aranzabe, 2008) that were sintactically deep tagged, that is PRO[3] and pro[4] elements had a tag.

Let us explain this process by means of an example. Figure 2 shows the tree of the original sentence *Bere zeregina zatituta dagoen alderdia batzea izango dela esan zuen* (S/he said that her/his mission is to unify the political party that is divided).

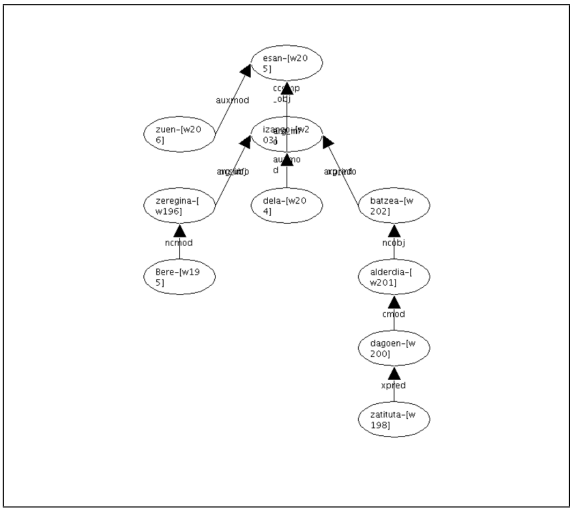

Figure 2: Original sentence: *Bere zeregina zatituta dagoen alderdia batzea izango dela esan zuen*

Having this input our algorithm works as follows:

1. The relative clause *zatituta dagoen* (that is divided) is removed out of the original sentence. This way we get two trees: the main clause *Bere zeregina alderdia batzea izango dela esan zuen.* (S/he said that her/his mission is to unify the political party.) (figure 3) and the relative

[3] see footnote 1

[4] elided arguments (pro-drop)

clause *zatituta dagoen* (that is divided) (figure 4).

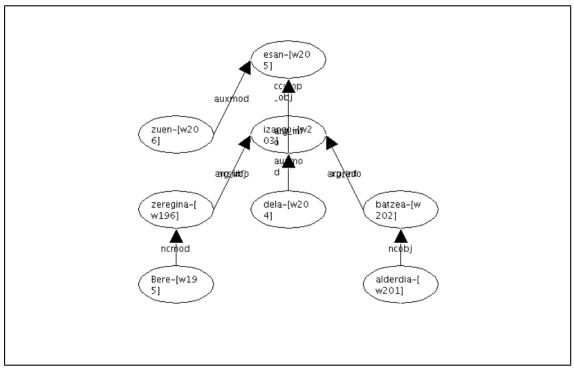

Figure 3: The main clause: *Bere zeregina alderdia batzea izango dela esan zuen*

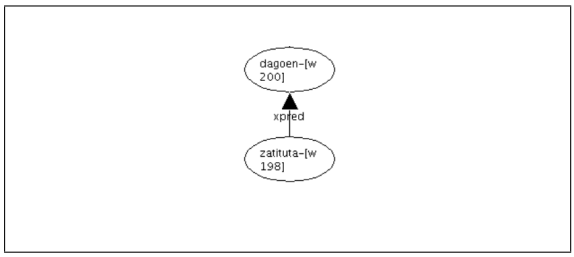

Figure 4: The relative clause: *zatituta dagoen*

2. The PRO antecedent of relative clause *alderdia* (The political party) is included in the new sentence. This way, the sentence *alderdia zatituta dagoen* is formed as shown in the tree of figure 5.

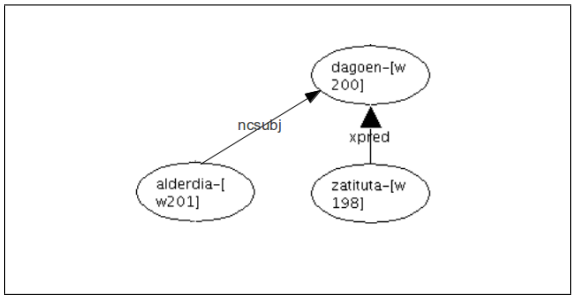

Figure 5: The new simple sentence (relative + antecedent): *alderdia zatituta dagoen*

In the case of adverbial temporal clauses, the adverbial clause is removed in the first step and an adverb will be added in the second step.

This way the reconstruction operation is over in both cases at tree level. That is simple sentences are formed, but they are not grammatically correct. The reconstruction will be over, continuing with this example, by removing the *-(e)n* complementiser of the verb.

5 Evaluation

In this section we evaluate the correctness assessing the splitting point tag and splitting the sentences.

The corpus that has been used to develop and to evaluate the grammar has been EPEC. We divided the corpus in two sets: devel and eval. We used devel for designing the rules of the grammar and eval for automatic evaluation. The latter was previously manually tagged. In table 1 we see the word and sentence number we have used for this task in the development part and the evaluation part of the corpus.

	Devel	**Eval**
Word number	61121	63766
Sentence number	5068	5211
Clause number	18301	18356

Table 1: Word, sentence and clause number in corpus

In table 2 we show the results we obtained by relative clauses and adverbial temporal clauses. The measures that we have used are precision (correctly detected clauses/detected clauses), recall (correctly detected clauses/all clauses) and F-measure (2 * precision * recall / (precision + recall)). Forth column shows the clause number of each structure.

For relative clauses, the results are high. The F-measure for the finite verbs is 0,988 and for the non finite verbs it is 0,992. By analysing the errors the chunker made we concluded that:

- We have a problem with a rule that aims a finite verb temporal clause with free elements structure that can be mixed with relative sentences.
- Another kind of error was due to errors in the PoS tagging.
- Non finite modal verbs structures were not found in the development part.

For temporal clauses, we have to divide the results in two groups: clauses without free elements and clauses with free ele-

	Precision	Recall	F-measure	Clause number
Relative finite verb clauses	0,998	0,978	0,988	547
Relative non finite verb clauses	1	0,985	0,992	335
Temporal finite verb clauses	0,955	0,964	0,960	111
Temporal non finite verb clauses	0,966	0,966	0,966	29
Temporal finite verb clauses + free element(s)	1	0,556	0,714	18
Temporal non finite verb clauses + free element(s)	0,970	0,372	0,538	86

Table 2: Evaluation results of the treated phenomena

ments. The results for the first group are quite high and similar for finite and non finite verbs. The F-measure for temporal finite verb clauses is 0,960 and for the non temporal finite verb clauses is 0,966. We analysed the errors and they are due to canonical word order alteration.

The results for the second group are, however, lower. The F-measure for the temporal finite verb clauses + free element(s) is 0,714 and for the temporal non finite verb clauses + free element(s) is 0,538. The main problem here is that the recall is very low (finite verbs 0,556 and non finite verbs 0,372). Those results are due to:

- The ambiguity of the free elements
- The richness of those structures (all of them were not found in the development part)

Anyway, apart from the problem of the ambiguity the precision we get is high (finite verbs 1 and non finite verbs 0,970).

Since our aim consists on getting accuracy (precision) it is widely achieved, so we consider that we have a basis to continue with the simplification process. This basis is extremely remarkable for relative clauses. The results of the temporal clauses are good. Nevertheless, we should keep on improving the rules, and if possible, getting more structures. It is remarkable too that recall goes down resounding when the clause has free elements, since it is difficult to cover all the possible structures with a corpus. So, defining the clause boundaries is a continuous task we have to keep on working on in order to improve our clause boundary identifier.

6 Conclusion and Future work

In this paper we have focused on the splitting module in our text simplification architecture, since we think that it is important to have a good basis to continue with the simplification process. As we have explained, this module works on two phases: clause boundary detection and splitting point tagging and building simple sentence dependency-trees out of original sentence. The first phase tagging is made by means of *Mugak* a linguistic knowledge based grammar written in the Constraint Grammar formalism and the second phase is carried out by an algorithm based on dependencies-trees as well to create so many sentences out of the clauses in the original sentences. Furthermore, this algorithm introduces the clause in the reconstruction operation.

For this task, we have deeply analysed two diverse structures, namely relative clauses and adverbial temporal clauses. We have explained their different formation and the challenge they suppose.

We have made an evaluation and concluded that we have great basis to continue with the simplification process. Moreover, the algorithm we have implemented introduces the clauses in the reconstruction step fulfilling almost the simplification process in the case of relative sentences. But, on the other hand, the improvements made here to the clause boundary identifier will serve to improve the performance of other tools which use older versions of this identifier, for example, the statistical clause boundary identifier (Arrieta, 2010).

Our next step is actually to keep on working with the syntactic simplification process. For the verb state changing, that is becoming a subordinate verb into a main verb, we plan to use finite state technology tools like FOMA (Hulden, 2009). This tool will be useful as well to implement deletion and addition rules so far defined in (Gonzalez-Dios, 2011).

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. This research was supported by the Basque Government (IT344-10), and the

Spanish Ministry of Science and Innovation (MICINN, TIN2010-20218). We thank Iñigo Lopez-Gazpio for the support in programming task.

References

Aduriz, Itziar, Izaskun Aldezabal, Iñaki Alegria, Jose Mari Arriola, Arantza Díaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing.*, pages 3–11.

Aduriz, Itziar, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. 2006a. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15.

Aduriz, Itziar, Bertol Arrieta, Jose Mari Arriola, Arantza Díaz de Ilarraza, Elixabete Izagirre, and Ainara Ondarra. 2006b. Muga Gramatikaren Optimizazioa. Technical report, UPV/EHU/LSI/TR 9-2006.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria, Mikel Iruskieta, Elixabete Izagirre, Karmele Mendizabal, and Eli Pociello. 2006. A methodology for the joint development of the Basque WordNet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*.

Agirre, Eneko, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. In *Proceedings of NAACL-ANLP'92*, pages 119–125.

Al-Subaihin, Afnan A. and Hend S. Al-Khalifa. 2011. Al-Baseet: A proposed simplification authoring tool for the Arabic language. In *International Conference on Communications and Information Technology (ICCIT)*, pages 121–125.

Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Ainara Estarrona, Kike Fernandez, and Larraitz Uria. 2010. EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua. Technical report, UPV/EHU/LSI/TR 02-2010.

Alegria, Iñaki, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6.

Alegria, Iñaki, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2003. Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI*.

Aranzabe, María Jesús. 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Ph.D. thesis, Euskal Filologia Saila (UPV/EHU).

Arrieta, Bertol. 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean*. Ph.D. thesis, Informatika Fakultatea (UPV-EHU).

Bernhard, Delphine, Louis De Viron, Véronique Moriceau, and Xavier Tannier. 2012. Question Generation for French: Collating Parsers and Paraphrasing Questions. *Dialogue and Discourse*, 3(2):43–74.

Burstein, Jill. 2009. Opportunities for Natural Language Processing Research in Education. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg.

Candido, Jr., Arnaldo, Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42. ACL.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying Text for

Language-Impaired Readers. In *9th Conference of the European Chapter of the Association for Computational Linguistics*.

Castro-Castro, Daniel, Rocío Lannes-Losada, Montse Maritxalar, Ianire Niebla, Celia Pérez-Marqués, Nancy C. Alamo-Suarez, and Aurora Pons-Porrata. 2008. A Multilingual Application for Automated Essay Scoring. In *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA*, pages 243–251. Springer New York.

Ezeiza, Nerea. 2002. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Ph.D. thesis, Informatika Fakultatea, UPV-EHU.

Gonzalez-Dios, Itziar. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of the Basque Country.

Hulden, Mans. 2009. Foma: a Finite-State Compiler and Library. In *EACL (Demos)'09*, pages 29–32.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. ACL.

Iruskieta, Mikel, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2011. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, (47).

Jonnalagadda, Siddhartha and Graciela Gonzalez. 2010. Sentence simplification aids protein-protein interaction extraction. *Arxiv preprint arXiv:1001.4273*.

Karlsson, Fred, Atro Voutilainen, Juha Heikkila, and Atro Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Ondarra, Ainara. 2003. Murriztapen Gramatikaren sintaxia. EUSMG optimizatzen. Esaldi-mugak. Master's thesis, Euskal Herriko Unibertsitatea.

Petersen, Sarah E. and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Electrical Engineering*, (SLaTE):69–72.

Poornima, C., V. Dhanalakshmi, K.M. Anand, and KP Soman. 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rybing, Jonas, Christian Smith, and Annika Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*.

Saggion, Horacio, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

Seretan, Violeta. 2012. Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Siddharthan, Advaith. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, Advaith. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11. ACL.

Soraluze, Ander, Olatz Arregi, Xabier Arregi, Klara Ceberio, and Arantza Díaz de Ilarraza. 2012. Mention Detection: First Steps in the Development of a Basque Coreference Resolution System. In *Proceedings of KONVENS 2012 (Main track: oral presentations)*, pages 128–163.

Urizar, Ruben. 2012. *Euskal lokuzioen tratamendu konputazionala.* Ph.D. thesis, UPV-EHU.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 69-76 recibido 30-11-12 revisado 11-02-13 aceptado 19-02-13

Una aproximación basada en corpus para la detección del foco geográfico en el texto

A corpus-based approach to geographical focus detection in text

Fernando S. Peregrino, David Tomás, Fernando Llopis
Universidad de Alicante
Carretera San Vicente del Raspeig s/n - 03690 Alicante (Spain)
{fsperegrino, dtomas, llopis}@dlsi.ua.es

Resumen: El foco geográfico de un documento identifica el lugar o lugares en los que se centra el contenido del texto. En este trabajo se presenta una aproximación basada en corpus para la detección del foco geográfico en el texto. Frente a otras aproximaciones que se centran en el uso de información puramente geográfica para la detección del foco, nuestra propuesta emplea toda la información textual existente en los documentos del corpus de trabajo, partiendo de la hipótesis de que la aparición de determinados personajes, eventos, fechas e incluso términos comunes, pueden resultar fundamentales para esta tarea. Para validar nuestra hipótesis, se ha realizado un estudio sobre un corpus de noticias geolocalizadas que tuvieron lugar entre los años 2008 y 2011. Esta distribución temporal nos ha permitido, además, analizar la evolución del rendimiento del clasificador y de los términos más representativos de diferentes localidades a lo largo del tiempo.
Palabras clave: Foco geográfico, recuperación de información geográfica, clasificación de textos, aprendizaje automático

Abstract: The geographical focus of a document identifies the relevant locations mentioned in text. This paper presents a corpus-based approach to detecting the geographical focus in documents. Despite other approaches focused on using solely geographical information, our proposal employs all the textual information included in the corpus under the assumption that the presence of particular names of persons, events, and even common terms can definitely help to solve this task. In order to validate our hypothesis, a study was carried out on a corpus of georeferenced news that took place between 2008 and 2011. Moreover, this temporal distribution allowed to carry out a study on the evolution of the performance of the classifier and the most representative terms for different locations over time.
Keywords: Geographical focus, geographical information retrieval, text classification, machine learning

1 Introducción

La identificación del foco geográfico de un documento consiste en determinar la principal o principales localizaciones a las que se hace referencia en el texto de entre todas las que se nombran en él (Amitay et al., 2004). Por ejemplo, en un documento que trata temas relativos a la Comunidad Valenciana, pueden aparecer frases como "*La compañía proveniente de China ha establecido en Valencia su segunda mayor fábrica, después de la de San Francisco, junto con otras empresas locales.*". Dicho documento, pese a citar localizaciones como China y San Francisco, tiene como único foco geográfico la Comunidad Valenciana, siendo la relevancia del resto de entidades meramente testimonial.

Obtener el foco geográfico de un documento es fundamental a la hora de desarrollar sistemas de *recuperación de información geográfica* (*Geographic Information Retrieval*, GIR). Estos sistemas son una especialización de los sistemas tradicionales de *recuperación de información* enfocados a la obtención de documentos relevantes para una determinada localización. Identificar el foco geográfico puede permitir a estos sistemas determinar con mayor precisión la relevancia

ISSN 1135-5948

que tiene un documento para un determinado lugar, descartando aquellos documentos donde la mención de una localización es puntual e irrelevante. Más aún, en ocasiones el foco geográfico ni siquiera se menciona explícitamente en el texto. En estos casos, la presencia de otros elementos geográficos que se nombran en él pueden ayudar a inferirlo (Amitay et al., 2004).

El presente trabajo se centra en el problema de la detección del foco geográfico en el texto. Esta tarea se ha abordado desde el punto de vista de la clasificación textual. Empleando un corpus de noticias previamente geolocalizadas y pertenecientes a 61 localidades del estado Español, hemos entrenado un sistema de clasificación automático capaz de determinar, para una nueva noticia, el principal foco de atención geográfico tratado en ella, asignando la localidad más probable de entre las 61 posibles. A diferencia de aproximaciones anteriores centradas únicamente en la información geográfica, nuestra propuesta emplea todo el contenido textual de los documentos, utilizando de esta manera un conocimiento general del mundo para la clasificación. En nuestro caso, partimos de la hipótesis de que las personas, eventos, fechas y términos comunes que caracterizan una localidad pueden resultar de gran utilidad a la hora de determinar su preponderancia en el texto.

Adicionalmente, dado que el corpus de noticias empleado en este estudio pertenece a diferentes años naturales (desde 2009 hasta 2011), hemos realizado un estudio de la evolución del rendimiento del clasificador a lo largo del tiempo, así como de la evolución de los términos más representativos de alguna de las localidades existente en el corpus.

El resto de este artículo está organizado del siguiente modo: en la Sección 2, se comentan los trabajos más destacados dentro del campo de la identificación del foco geográfico en el texto; la Sección 3 describe el corpus utilizado en nuestros experimentos; en la Sección 4 se describen los experimentos y se muestran los resultados que se han llevado a cabo; la Sección 5 resume las conclusiones de este estudio y las propuestas de trabajo futuro.

2 Trabajo relacionado

Los documentos de texto se pueden asociar frecuentemente con un determinado contexto geográfico. Para la detección de las localizaciones en el texto es común el uso de técnicas de *reconocimiento de entidades nombradas* (*Named Entity Recognition*) capaces de detectar entes de tipo geográfico. Esta detección va siempre asociada a un proceso posterior de desambiguación, a fin de concretar a cuál de las múltiples localizaciones a las que se puede asociar un mismo nombre se está refiriendo el texto (Buscaldi y Rosso, 2008). Sirvan como ejemplo los 18 Jerusalem y 63 Springfields que hay en Estados Unidos.

Este proceso de detección, sin embargo, no proporciona información sobre la verdadera relevancia que tienen las entidades geográficas nombradas en el texto. Si bien el tema de la detección y desambiguación de topónimos es un tema ampliamente tratado en la literatura científica (Leidner, 2007), son escasos los trabajos centrados en determinar el grado real de relevancia de las entidades geográficas que aparecen en el texto.

En (Martins y Silva, 2005) los autores presentan una variante del algoritmo *PageRank* para la detección del ámbito geográfico de cada documento. Para ello usan las referencias geográficas extraídas del texto y una técnica basada en ontologías, combinando ambas e infiriendo el ámbito global de cada documento a partir de la aplicación de *PageRank* al grafo generado a partir de la ontología.

En otro trabajo de los mismos autores (Anastácio, Martins, y Calado, 2009a), un conjunto de documentos obtenidos de la Web son categorizados en función de las localizaciones expresadas en ellos, utilizando *máquinas de vectores de soporte* (*Support Vectors Machines*, SVM) con diferentes vectores de características, como por ejemplo los n-gramas extraídos de las URLs de las páginas web.

El estudio presentado en (Ye et al., 2011) se centra en identificar la localización de la que se habla en un conjunto de blogs de viaje. Para ello, los autores implementan un extractor de localizaciones (Qin et al., 2010) con el fin de obtener los lugares mencionados en estos blogs. Debido a la inherente ambigüedad de los topónimos, exploran rasgos textuales en su contexto cercano (palabras alrededor del topónimo identificado) y rasgos geográficos (relaciones geográficas entre las localizaciones del texto) para llevar a cabo la clasificación por relevancia de cada una de las localizaciones identificadas.

En (Amitay et al., 2004), los autores geolocalizan contenido de la Web mediante el uso de diccionarios geográficos (*gazetteers*). Estos diccionarios son empleados para la desambiguación de topónimos, a los que se les asigna un factor de confianza. Este factor de confianza, junto con la frecuencia de aparición del topónimo en el texto y el resto de entidades geográficas que tienen lugar en él, determinan un valor final de relevancia que permite identificar el foco geográfico del texto. La novedad de esta aproximación es la capacidad de detectar el foco de un documento incluso cuando éste no aparece explícitamente en él, infiriéndolo a partir del resto de localizaciones presentes.

Finalmente, en (Anastácio, Martins, y Calado, 2009b) se pude ver un estudio comparativo sobre cuatro sistemas utilizados para la asignación del foco geográfico: *Yahoo! PlaceMaker*,[1] *Web-a-Where* (Amitay et al., 2004), *GIPSY* (Woodruff y Plaunt, 1994) y *GREASE* (Martins y Silva, 2005). La aproximación vencedora fue la llevada a cabo por el sistema *Web-a-Where*.

Por lo que respecta a aplicaciones finales de estas tecnologías, en (Clough et al., 2011) los autores muestran como enlazar los archivos oficiales del gobierno británico (*UK National Archives*) con el foco geográfico de los documentos para mejorar el acceso a los mismos.

A diferencia de la mayoría de sistemas aquí mencionados, nuestra propuesta no se centra únicamente en la información geográfica para determinar el foco, sino que emplea toda la información textual presente en un documento para su identificación. Esto nos va a permitir identificar el foco geográfico incluso en las situaciones en las que éste no aparece de forma explícita en el texto, ya que el resto de información presente en él nos puede llevar a inferir sobre qué localidad se está hablando.

3 Corpus de trabajo

Para poder llevar a cabo los experimentos propuestos en este trabajo, recopilamos un conjunto de noticias locales en español pertenecientes al periódico *20 Minutos*.[2] El corpus resultante consistió en más de 500.000 noticias (ver Tabla 1), comprendidas entre los años 2008 y 2011, pertenecientes a 61 localidades de España.[3]

Año	Noticias	Vocabulario
2.008	55.019	182.716
2.009	29.394	145.191
2.010	224.729	370.007
2.011	224.179	377.153
Total	553.321	606.229

Tabla 1: Estadísticas del corpus recopilado. Se muestra el número de noticias y el tamaño del vocabulario para cada uno de los años.

El hecho de que cada noticia tenga asociada una localidad en el corpus, nos permite considerar a dicha localidad como el foco geográfico de la noticia, ya que su contenido ha sido considerado como relevante para esa localización concreta. Este corpus de gran tamaño nos va a permitir aplicar técnicas de aprendizaje automático para construir un clasificador capaz de asignar un foco geográfico, para una nueva noticia dada, de entre las 61 localidades mencionadas anteriormente.

4 Experimentos y resultados

Para evaluar nuestra hipótesis de partida, hemos realizado diversos experimentos sobre el corpus descrito en la sección anterior. En primer lugar, hemos evaluado diferentes algoritmos de clasificación para determinar aquél que nos pudiera proporcionar un mejor rendimiento en la tarea.

El segundo experimento ha consistido en un estudio sobre la selección de características para reducir el número de dimensiones de las instancias de aprendizaje. El objetivo es incrementar el rendimiento del clasificador eliminando ruido (características innecesarias), reduciendo además el tiempo de computación necesario para llevar a cabo el proceso de entrenamiento y evaluación.

[1] http://developer.yahoo.com/geo/placemaker/

[2] http://www.20minutos.es/

[3] A Coruña, Albacete, Algeciras, Alicante, Almeria, Ávila, Badajoz, Barcelona, Bilbao, Burgos, Cáceres, Cádiz, Cartagena, Castellón de la Plana, Ceuta, Ciudad Real, Córdoba, Cuenca, Elche, Gijón, Girona, Granada, Guadalajara, Huelva, Huesca, Jaén, Jerez de la Frontera, Las Palmas de Gran Canaria, León, Lleida, Logroño, Lugo, Madrid, Málaga, Marbella, Melilla, Ourense, Oviedo, Palencia, Palma de Mallorca, Pamplona, Pontevedra, Salamanca, San Sebastián, Santa Cruz de Tenerife, Santander, Santiago de Compostela, Segovia, Sevilla, Soria, Teruel, Toledo, Valencia, Valladolid, Vigo, Vitoria, Zamora y Zaragoza.

El tercer experimento se ha enfocado a evaluar la evolución del rendimiento del clasificador al ser entrenado con muestras de un año (2008) y evaluado con muestras de años posteriores (2009 a 2011). Esto nos va a permitir ver cómo la evolución del vocabulario, un concepto muy vivo en el caso de las noticias de prensa, afecta al rendimiento del sistema con el paso del tiempo.

El último experimento está centrado en analizar la evolución del vocabulario empleado en las noticias a lo largo de los cuatro años que cubre el corpus recopilado. Para ello se han obtenido los términos más discriminatorios a la hora de identificar diferentes ciudades de nuestro corpus, comprobando cómo éstos han ido cambiando a lo largo del tiempo.

A continuación se describen en detalle estos cuatro experimentos y los resultados obtenidos.

4.1 Algoritmo de aprendizaje

Dada que nuestra hipótesis de partida es que toda la información presente en el texto, y no sólo la geográfica, puede ser relevante a la hora de determinar el foco geográfico, el conjunto de características utilizadas para la clasificación lo forman todos los unigramas obtenidos de los documentos, es decir, su vocabulario (tras su normalización a minúsculas y la eliminación de signos de puntuación). Partiendo de este conjunto de características, se han alimentado diferentes algoritmos de clasificación llevando a cabo una validación cruzada (*10-fold cross-validation*) para calcular el rendimiento del clasificador. Este experimento se llevó a cabo inicialmente sobre el corpus de 2008. Como medida de rendimiento del clasificador hemos empleado la precisión, entendida como el número de documentos correctamente clasificados del total de documentos existentes.

El conjunto de algoritmos probados lo conforman *Naïve Bayes* (implementación de Weka (Witten y Frank, 2005)), k-NN (implementación de Weka y Timbl (Daelemans y van den Bosch, 2009)) y SVM. De todos ellos, únicamente las implementaciones de SVM llevadas a cabo en *LibSVM* (Chang y Lin, 2011) y *LibLINEAR* (Fan et al., 2008) permitieron obtener resultados en un tiempo asumible.[4]

[4]Los experimentos fueron llevados a cabo en un servidor *IBM System x3400 M3 Xeon 5606* con 32Gb de RAM.

Año	Precisión
2008	72,70 %
2009	82,32 %
2010	85,32 %
2011	84,31 %

Tabla 2: Precisión obtenida con *LibLINEAR* para cada uno de los años.

Sobre el corpus de 2008 (55.019 documentos y 182.176 características de aprendizaje), *LibSVM* obtuvo una precisión de 67,05 %, frente al 72,70 % obtenido por *LibLINEAR*. Esto, unido a que el tiempo empleado por *LibSVM* en completar la tarea fue un orden de magnitud superior al empleado por *LibLINEAR*, nos hizo decantarnos por este último para completar el resto de experimentos planteados en este trabajo. El rendimiento obtenido con *LibLINEAR* para los cuatro años del corpus se puede ver en la Tabla 2.

Estos resultados revelan una rendimiento muy similar entre los años 2009, 2010 y 2011. Sin embargo, el rendimiento obtenido en 2008 resultó notablemente inferior al resto. Tras un estudio pormenorizado de la matriz de confusión obtenida para ese año, se observó que en algunas localidades cercanas, el número de noticias clasificadas de forma errónea era muy numerosa. Es el caso, por ejemplo, de Gijón y Oviedo, a tan sólo 30 Km. una de otra. El número de noticias de Oviedo correctamente clasificadas fue de 406, mientras que las incorrectamente clasificadas como pertenecientes a Gijón fue de 324. De forma similar, el número de noticias de Gijón erróneamente clasificadas como pertenecientes a Oviedo fue de 308. Analizando el corpus se observó que este fenómeno no era debido esencialmente al uso de terminología similar entre una ciudad y otra, sino a la duplicidad de noticias entre ambas. En el año 2008, el periódico *20minutos* incorporó nuevas ciudades a sus noticias locales, pero buena parte del contenido de éstas fue reutilizado de ciudades cercanas con mayor andadura en este medio. De cara a trabajos futuros, se plantea la necesidad de eliminar estas duplicidades en el corpus para favorecer el rendimiento del clasificador y la comparación equitativa con el resto de años.

4.2 Selección de características

En este experimento se evaluó la evolución del rendimiento del clasificador tras aplicar un proceso de selección de características basado en χ^2 (Yang y Pedersen, 1997). Sobre el corpus de 2008, se usó esta técnica para determinar estadísticamente cuáles eran las características que aportaban más información al proceso de aprendizaje, estableciendo diferentes umbrales de corte para eliminar aquellas que resultaran menos relevantes en el proceso (ruido). Las características seleccionadas se probaron con la mejor configuración obtenida en la sección anterior, es decir, empleando la librería *LibLINEAR*.

La Figura 1 muestra la curva de aprendizaje que se obtiene al ir reduciendo el número de características empleadas para representar las instancias del problema.

Los resultados obtenidos muestran una mejora notable en la precisión del clasificador al reducir el número de características. El rendimiento óptimo se obtiene con 3.000 características (77,32 %), proporcionando una mejora cercana al 7 % con respecto al experimento original. Además de suponer una mejora en el rendimiento del clasificador, el coste computacional del entrenamiento y evaluación se ve significativamente reducido, al pasar de más de 180.000 características para representar cada instancia a tan sólo 3.000, lo que supone una reducción de la dimensionalidad del problema de más del 98 %.

4.3 Evolución del rendimiento del clasificador

Este experimento tiene como objetivo determinar cómo evoluciona el rendimiento del clasificador cuando se entrena con el corpus de un año y se evalúa sobre años posteriores. Lo que se pretende observar es cómo afecta al rendimiento del clasificador los cambios de vocabulario inherentes a un medio como el de las noticias periodísticas, donde la terminología empleada en un periodo de tiempo viene supeditada a los temas y personajes del momento.

En este caso se entrenó sobre el corpus de 2008 con la mejor configuración de nuestro sistema (*LibLINEAR* con 3.000 características de aprendizaje), empleando posteriormente para la evaluación los corpus de 2009, 2010 y 2011. Los resultados se pueden ver en la Tabla 3.

Año	Precisión
2008	77,32 %
2009	78,76 %
2010	65,82 %
2011	63,91 %

Tabla 3: Precisión obtenida con *LibLINEAR* y 3.000 características de aprendizaje, entrenando sobre el corpus de 2008 y evaluando sobre el resto de años. El rendimiento para 2008 se obtuvo mediante validación cruzada (*10-fold cross-validation*).

Se puede observar como el rendimiento obtenido para 2008 y 2009 es muy similar (77,32 % y 78,76 % respectivamente). El hecho de que en 2009 sea ligeramente superior a 2008 puede resultar contradictorio, pero teniendo en cuenta que el sistema entrenado y evaluado sobre 2009 obtenía una precisión un 13 % mayor que la obtenida en 2008 (ver Tabla 2), este valor refleja en realidad un decremento notable en el rendimiento del clasificador con respecto al experimento original. Se puede observar además que el rendimiento decae de forma monótona conforme avanzan los años, ya que la pérdida de rendimiento es mayor en 2010 que en 2009 y en 2011 que en 2010. Esto confirma el hecho de que el vocabulario empleado en las noticias evoluciona a lo largo del tiempo, produciendo un deterioro paralelo del rendimiento del clasificador. La solución a este problema pasaría por un reentrenamiento periódico del sistema para actualizar el vocabulario empleado como características de aprendizaje.

4.4 Evolución del vocabulario

En este último experimento analizaremos la evolución a lo largo de los años 2008 y 2009 de los términos más relevantes para la clasificación en tres ciudades distintas: Madrid, Valencia y Alicante.

La Tabla 4 muestra los 10 términos más discriminatorios en el proceso de clasificación (según la ponderación realizada mediante χ^2), ordenados de mayor a menor relevancia para cada una de las tres ciudades citadas anteriormente.

Como era de esperar, una parte de los términos más representativos de estas ciudades son aquellos que hacen referencia a la propia ciudad: sus nombres, gentilicios y localidades cercanas. Sin embargo, también se observa la inclusión de determinados eventos

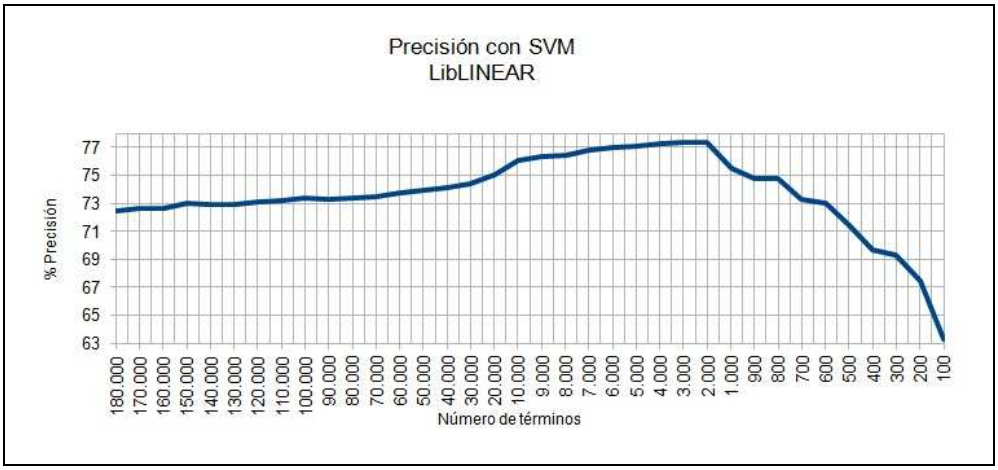

Figura 1: Precisión del clasificador en función del número de características seleccionadas mediante χ^2. Se empleó el corpus de 2008 y la librería *LibLINEAR* para la clasificación.

Madrid		**Valencia**		**Alicante**	
2008	**2009**	**2008**	**2009**	**2008**	**2009**
madrid	madrid	valencia	valencia	alicante	alicante
madrileños	madrileños	valenciana	comunitat	alicantinos	alicantinos
madrileño	madrileño	valencianos	valenciana	alicantina	alicantina
aguirre	madrileña	comunitat	valencianos	benidorm	alicantino
gallardón	aguirre	valenciano	valenciano	alicantino	orihuela
madrileña	gallardón	conselleria	conselleria	torrevieja	dénia
comunidad	comunidad	valencianas	generalitat	vicent	castedo
samur	parís	fallas	camps	dénia	benidorm
coslada	esperanza	xàtiva	consell	orihuela	torrevieja
parís	vallecas	consell	valencianas	elche	elche

Tabla 4: Términos más relevantes (ponderados mediante χ^2) para Madrid, Valencia y Alicante en los años 2008 y 2009, ordenados de mayor a menor relevancia.

(como las "fallas" en Valencia) y personajes (como "gallardón" en Madrid, "camps" en Valencia y "castedo" en Alicante) que presentan una gran relevancia para la clasificación y que van cobrando mayor o menor preponderancia dependiendo del año en el que nos encontremos.

Para profundizar un poco más en este aspecto, hemos analizado los 100 términos más relevantes según χ^2 para estas tres ciudades. La Tabla 5 muestra un resumen de las características no geográficas más singulares detectadas en este estudio, es decir, aquellos términos que resultaron más relevantes para la clasificación y que no se consideran como información geográfica (eventos, personajes, empresas, etc.). Las columnas *2008* y *2009* muestran la posición que ocupan dichos términos en le ranking de relevancia para cada uno de esos años. Un guión ("-") indica que el término no se encontraba entre los 100 más relevantes ese año.

Este análisis muestra la existencia de numerosos términos no geográficos entre los más relevantes a la hora de determinar el foco geográfico de las tres localidades presentadas. Entre estos términos nos encontramos personajes como "barberá" en Valencia, comidas como "tonyina" en Alicante y monumentos como "cibeles" en Madrid.

Además, podemos observar la evolución de la importancia de algunos términos en el tiempo (entre 2008 y 2009 en este caso). Por ejemplo, el término "alperi" (en referencia a Luis Díaz Alperi, alcalde de la ciudad de Alicante desde el año 1995 hasta el 2008), presenta una relevancia notable en 2008 (puesto 14), mientras que en 2009 deja de aparecer entre los 100 términos más relevantes (coincidiendo con el abandono del cargo en esta ciudad). Por el contrario, su sucesora en el cargo Sonia Castedo (representada por el término

Madrid			Valencia			Alicante		
Término	**2008**	**2009**	**Término**	**2008**	**2009**	**Término**	**2008**	**2009**
aguirre	4	5	comunitat	4	2	comunitat	11	85
comunidad	7	7	conselleria	6	6	alperi	14	-
samur	8	11	fallas	8	17	cicu	16	30
parís	10	8	fgv	11	29	samu	18	32
summa	12	23	maquinistas	12	-	03001	29	-
esperanza	16	9	turia	13	33	tonyina	30	-
distrito	21	22	metrovalencia	14	37	cam	39	-
cibeles	24	14	barberá	15	18	dinos	41	-
portavoz	26	59	ferrocarrils	17	48	castedo	45	7

Tabla 5: Términos no geográficos relevantes para Madrid, Valencia y Alicante, ordenados de mayor a menor importancia. Las columnas *2008* y *2009* muestran la posición que ocupan dichos términos en le ranking de relevancia determinado mediante χ^2 para esos años.

"castedo" en nuestra lista), pasa de la posición 45 en 2008 a la 7 en 2009 en términos de relevancia.

Una situación curiosa se da en el caso de "parís", que presenta una gran relevancia para Madrid tanto en 2008 (posición 10) como en 2009 (posición 8). Esta situación cobra sentido si consideramos que ambas son capitales de nación, teniendo tendencia a aparecer de forma conjunta cuando se realizan referencias a cualquier tipo de situación política o social que involucre a ambos países.

5 Conclusiones y trabajo futuro

En este artículo se ha presentado una aproximación a la detección del foco geográfico en el texto basada en aprendizaje automático, empleando características textuales generales, y no sólo geográficas, para su identificación. Para evaluar nuestra aproximación hemos experimentado con un corpus de más de 500.000 noticias locales que tuvieron lugar entre 2009 y 2011. Esto nos ha permitido realizar un análisis de la evolución temporal de la terminología empleada en los textos, así como del rendimiento del clasificador al ser evaluado con noticias cada vez más alejadas en el tiempo de aquellas que se emplearon para su entrenamiento.

El tamaño del corpus empleado en este estudio (553.321 noticias y un vocabulario de 606.229 términos para el total de los cuatro años comprendidos) hizo que muchos de los algoritmos tradicionales de clasificación (como *Naïve Bayes* y *k-NN*) fueran incapaces de completar la tarea debido al tamaño del problema. La implementación de SVM llevada a cabo en la librería *LibLINEAR* proporcionó los mejores resultados en cuanto a rendimiento y eficiencia temporal, siendo tomada como base para el la realización de nuestro estudio.

El estudio realizado sobre selección de características mediante χ^2, demostró que se puede conseguir un incremento de la precisión del sistema (cercana al 7 %) gracias a la eliminación de ruido que produce el exceso de términos irrelevantes en el proceso de clasificación. Los mejores resultados se obtuvieron estableciendo el umbral de corte en 3.000 características, lo que supone una reducción de más del 98 % del conjunto de características de aprendizaje, influyendo decisivamente en el coste computacional del proceso.

Para evaluar cómo evoluciona el rendimiento del clasificador en el tiempo, se entrenó el sistema sobre el corpus de 2008 y se evaluó sobre el resto de años. Los resultados mostraron que, efectivamente, el rendimiento del clasificador disminuye de forma monótona conforme avanza la línea del tiempo. Esto es debido a la evolución del vocabulario empleado en un medio tan vivo como el de las noticias periodísticas, que hace que de un año para otro cambien notablemente los tópicos de interés tratados en éstas, así como sus personajes y eventos.

Finalmente, por lo que respecta al análisis de los términos más relevantes para una localidad, se observó que para los casos de estudio realizados buena parte de los términos más importantes a la hora de determinar el foco geográfico de una localidad no eran de tipo geográfico. Personajes, eventos y términos comunes resultaron ser de gran relevancia para esta tarea. Esto respalda nuestra hipóte-

sis inicial de que la información general del mundo puede resultar de gran utilidad para la detección del foco geográfico en el texto.

Además de analizar estos términos relevantes con respecto a una localidad, se vio cómo evolucionaban a lo largo del tiempo. Los casos estudiados muestran cómo la relevancia que tenga, por ejemplo, un determinado personaje en el ámbito local de una ciudad, afecta definitivamente a lo importante que éste resulte a la hora de identificar el foco geográfico de la misma.

Como trabajo futuro, se plantea la inclusión de nuevas características en el proceso de aprendizaje (empleando fuentes externas de conocimiento como Wikipedia[5] y Geonames[6]), el estudio de la influencia de la granularidad (a nivel de ciudad, provincia, comunidad, etc.) en el rendimiento del clasificador y la extensión de los experimentos a corpus que cuenten con un registro diferente de lenguaje, como pueden ser los blogs de tipo turístico presentes en Internet.

Bibliografía

Amitay, Einat, Nadav Har'El, Ron Sivan, y Aya Soffer. 2004. Web-a-where: geotagging web content. En *Proceedings of the 27th annual international ACM SIGIR conference*, SIGIR '04, páginas 273–280, New York, NY, USA. ACM.

Anastácio, Ivo, Bruno Martins, y Pável Calado. 2009a. Classifying documents according to locational relevance. En *Progress in Artificial Intelligence*, volumen 5816. Springer Berlin Heidelberg, páginas 598–609.

Anastácio, Ivo, Bruno Martins, y Pável Calado. 2009b. A comparison of different approaches for assigning geographic scopes to documents. En *1st INForum-Simpósio de Informática*, páginas 285–296.

Buscaldi, Davide y Paulo Rosso. 2008. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22(3):301–313, Enero.

Chang, Chih-Chung y Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, Mayo.

[5]http://www.wikipedia.org/

[6]http://www.geonames.org/

Clough, Paul, Jiayu Tang, Mark M Hall, y Amy Warner. 2011. Linking archival data to location: a case study at the uk national archives. *ASLIB Proceedings*, 63(2/3):127–147.

Daelemans, Walter y Antal van den Bosch. 2009. *Memory-Based Language Processing*. Cambridge University Press, New York, NY, USA, 1st edición.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, y Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, Junio.

Leidner, Jochen Lothar. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. tesis, School of Informatics, University of Edinburgh.

Martins, Bruno y M. J. Silva. 2005. A graph-ranking algorithm for geo-referencing documents. En Jiawei Han y Et Al.Editor, editores, *Fifth IEEE International Conference on Data Mining ICDM05*, volumen 2002, páginas 741–744. IEEE.

Qin, Teng, Rong Xiao, Lei Fang, Xing Xie, y Lei Zhang. 2010. An efficient location extraction algorithm by leveraging web contextual information. En *Proceedings of the 18th SIGSPATIAL*, GIS '10, páginas 53–60, New York, NY, USA. ACM.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edición.

Woodruff, Allison Gyle y Christian Plaunt. 1994. Gipsy: automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655.

Yang, Yiming y Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. En *ICML '97*, páginas 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ye, Mao, Rong Xiao, Wang-Chien Lee, y Xing Xie. 2011. Location relevance classification for travelogue digests. En *WWW '11*, páginas 163–164, New York, NY, USA. ACM.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 77-84 recibido 01-12-12 revisado 14-02-13 aceptado 19-02-13

Sentiment Analysis and Topic Classification based on Binary Maximum Entropy Classifiers

Análisis de Sentimiento y Categorización de texto basado en clasificadores binarios de máxima entropía

Fernando Batista, Ricardo Ribeiro

Laboratório de Sistemas de Língua Falada (L2F) - INESC-ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal

Instituto Universitário de Lisboa (ISCTE-IUL)
Av. Forças Armadas, 1649-026 Lisboa, Portugal

{Fernando.Batista, Ricardo.Ribeiro}@inesc-id.pt

Resumen: En este trabajo se presenta una estrategia basada en clasificadores binarios de máxima entropía para el análisis de sentimiento y categorización de textos de Twitter enfocados al español. El sistema desarrollado consigue los mejores resultados para la categorización temática, y el segundo lugar para el análisis de sentimiento, en un esfuerzo de evaluación conjunta (Villena-Román et al., 2012). Se han explorado diferentes configuraciones para ambas tareas. Esto llevó a la utilización de una cascada de clasificadores binarios para el análisis de sentimiento y una estrategia de tipo uno-vs-todo para la clasificación de tema, donde los temas más probables para cada tweet fueron seleccionados.
Palabras clave: Análisis de sentimiento, categorización de texto en temas de interés, medios sociales, regresión logística, máxima entropía.

Abstract: This paper presents a strategy based on binary maximum entropy classifiers for automatic sentiment analysis and topic classification over Spanish Twitter data. The developed system achieved the best results for topic classification, and the second place for sentiment analysis in a joint evaluation effort – the TASS challenge (Villena-Román et al., 2012). Different configurations have been explored for both tasks, leading to the use of a cascade of binary classifiers for sentiment analysis and a one-vs-all strategy for topic classification, where the most probable topics for each tweet were selected.
Keywords: Sentiment analysis, topic detection, social media, logistic regression, maximum entropy.

1 Introduction

Social Networks take part in the nowadays life of a large number of people, providing revolutionary means for people to communicate and interact. Each social network targets different audiences, offering a unique range of services that people find useful in the course of their lives. Twitter offers a simple way for people to express themselves, by means of small text messages of at most 140 characters that can be freely used.

Twitter can be accessed in numerous ways, ranging from computers to mobile phones or other mobile devices. That is particularly important because accessing and producing content becomes a trivial task, therefore assuming an important part of people's lives. One relevant aspect that differentiates Twitter from other communication means is its ability to rapidly propagate such content and make it available to specific communities, selected based on their interests. Twitter data is a powerful source of information for assessing and predicting large-scale facts. For example, (O'Connor et al., 2010) capture large-scale trends on consumer confidence and political opinion in tweets, strengthening the potential of such data as a supplement for traditional polling. In what concerns stock markets, (Bollen, Mao, and Zeng, 2010) found that Twitter data can be used to significantly improve stock market predictions accuracy.

The huge amount of data, constantly being produced in a daily basis, makes it impracticable to manually process such content. For that reason, it becomes urgent to apply automatic

ISSN 1135-5948

processing strategies that can handle, and take advantage, of such amount of data. However, processing Twitter is all but an easy task, not only because of specific phenomena that can be found in the data, but also because it may require to process a continuous stream of data, and possibly to store some of the data in a way that it can be accessed in the future.

This paper tackles two well-known Natural Language Processing (NLP) tasks, commonly applied both to written and speech corpora: sentiment analysis and topic detection. The two tasks have been performed over Spanish Twitter data provided in the context of a contest proposed by "TASS – workshop on Sentiment Analysis" (Villena-Román et al., 2012), a satellite event of the SEPLN 2012 conference.

The paper is organized as follows: Section 2 overviews the related work, previously done for each task. Section 3 presents a brief description of the data. Section 4 describes the most relevant strategies that have been considered for tackling the problem. Section 5 presents and analyses a number of experiments, and reports the results for each one of the approaches. Section 6 presents some conclusions and discusses the future work.

2 Related work

Sentiment analysis and topic detection are two well-known NLP (Natural Language Processing) tasks. Sentiment analysis is often referred by other names (e.g. sentiment mining) and consists of assigning a sentiment, from a set of possible values, to a given portion of text. Topic detection consists of assigning a class (or topic) from a set of possible predefined classes to a given document. Often, these two tasks are viewed as two classification problems that, despite being characterized by their specificities, can be tackled using similar strategies. The remainder of this section overviews the related work previously done concerning these two tasks.

2.1 Sentiment analysis

Sentiment analysis can be performed at different complexity levels, where the most basic one consists just on deciding whether a portion of text contains a positive or a negative sentiment. However, it can be performed at more complex levels, like ranking the attitude into a set of more than two classes or, even further, it can be performed in a way that different complex attitude types can be determined, as well as finding the source and the target of such attitudes.

Dealing with the huge amounts of data available on Twitter demand clever strategies. One interesting idea, explored by (Go, Bhayani, and Huang, 2009) consists of using emoticons, abundantly available on tweets, to automatically label the data and then use such data to train machine learning algorithms. The paper shows that machine learning algorithms trained with such approach achieve above 80% accuracy, when classifying messages as positive or negative. A similar idea was previously explored by (Pang, Lee, and Vaithyanathan, 2002) for movie reviews, by using star ratings as polarity signals in their training data. This latter paper analyses the performance of different classifiers on movie reviews, and presents a number of techniques that were used by many authors and served as baseline for posterior studies. As an example, they have adapted a technique, introduced by (Das and Chen, 2001), for modeling the contextual effect of negation, adding the prefix `NOT_` to every word between a "negation word" and the first punctuation mark following the negation word.

Common approaches to sentiment analysis involve the use of sentiment lexicons of positive and negative words or expressions. The General Inquirer (Stone et al., 1966) was one of the first available sentiment lexicons freely available for research, which includes several categories of words, such as: positive vs. negative, strong vs. week. Two other examples include (Hu and Liu, 2004), an opinion lexicon containing about 7000 words, and the MPQA Subjectivity Cues Lexicon (Wilson, Wiebe, and Hoffmann, 2005), where words are annotated not only as positive vs. negative, but also with intensity. Finally, (Baccianella, Esuli, and Sebastiani, 2010) is another available resource that assigns sentiment scores to each synset of the wordnet.

Learning polarity lexicons is another research approach that can be specially useful for dealing with large corpora. The process starts with a seed set of words and the idea is to increasingly find words or phrases with similar polarity, in semi-supervised fashion (Turney, 2002). The final lexicon contains much more words, possibly learning domain-specific information, and therefore is more prone to be robust. The work reported by (Kim and Hovy, 2004) is another example of learning algorithm that uses WordNet synonyms and antonyms to learn polarity.

2.2 Topic Detection

Work on Topic Detection has its origins in 1996 with the Topic Detection and Tracking (TDT)

initiative sponsored by the US government (Allan, 2002). The main motivation for this initiative was the processing of the large amounts of information coming from newswire and broadcast news. The main goal was to organize the information in terms of events and stories that discussed them. The concept of topic was defined as the set of stories about a particular event. Five tasks were defined: story segmentation, first story detection, cluster detection, tracking, and story link detection. The current impact and the amount information generated by social media led to a state of affairs similar to the one that fostered the pioneer work on TDT. Social media is now the context for research tasks like topic (cluster) detection (Lee et al., 2011; Lin et al., 2012) or emerging topic (first story) detection (Kasiviswanathan et al., 2011).

In that sense, closer to our work are the approaches described by (Sriram et al., 2010) and (Lee et al., 2011), where tweets are classified into previously defined sets of generic topics. In the former, a conventional bag-of-words (BOW) strategy is compared to a specific set of features (authorship and the presence of several types of twitter-related phenomena) using a Naïve Bayes (NB) classifier to classify tweets into the following generic categories: News, Events, Opinions, Deals, and Private Messages. Findings show that authorship is a quite important feature. In the latter, two strategies, BOW and network-based classification, are explored to classify clusters of tweets into 18 general categories, like Sports, Politics, or Technology. In the BOW approach, the clusters of tweets are represented by tf-idf vectors and NB, NB Multinomial, and Support Vector Machines (SVM) classifiers are used to perform classification. The network-based classification approach is based on the links between users and C5.0 decision tree, k-Nearest Neighbor, SVM, and Logistic Regression classifiers were used. Network-based classification was shown to achieve a better performance, but being link-based, it cannot be used for all situations.

3 Data

Experiments described in this paper use Spanish Twitter data provided in the context of the TASS contest (Villena-Román et al., 2012). The provided training data consists of an XML file containing about 7200 tweets, each one labelled with sentiment polarity and the corresponding topics. We decided to consider the first 80% of the data for training our models (5755 tweets) and the remaining 20% for development (1444 tweets). The provided test data is also available in XML and contains about 60800 unlabeled tweets. The goal consists in providing automatic sentiment and topic classification for that data.

Each tweet in the labelled data is annotated in terms of polarity, using one of six possible values: *NONE*, *N*, *N+*, *NEU*, *P*, *P+* (Section 4.2 contains information about their meaning). Moreover, each annotation is also marked as *AGREEMENT* or *DISAGREEMENT*, indicating whether all the annotators performed the annotation coherently. In what concerns topic detection, each tweet was annotated with one or more topics, from a list of 10 possible topics: *política* (politics), *otros* (others), *entretenimiento* (entertainment), *economía* (economics), *música* (music), *fútbol* (football), *cine* (movies), *tecnología* (technology), *deportes* (sports), and *literatura* (literature).

It is also important to mention that, besides the tweets, an extra XML file is also available, containing information about each one of the users that authored at least one of the tweets in the data. In particular, the information includes the type of user, assuming one of three possible values – *periodista* (journalist), *famoso* (famous person), and *politico* (politician) – which may provide valuable information for these tasks.

Apart from the provided data, some experiments described in this paper also made use of Sentiment Lexicons in Spanish[1], a resource created at the University of North Texas (Perez-Rosas, Banea, and Mihalcea, 2012). From this resource, only the most robust part was used, known as *fullStrengthLexicon*, and containing 1346 words automatically labelled with sentiment polarity.

4 Approach

We have decided to consider both tasks as classification tasks, thus sharing the same method. The most successful and recent experiments cast the problem as a binary classification problem, which aims at discriminating between two possible classes. Binary classifiers are easier to develop, offer faster convergence ratios, and can be executed in parallel. The final results are then produced by combining all the different binary classifiers.

The remainder of this section describes the method and the architecture of the system when applied to each one of the tasks.

[1] http://lit.csci.unt.edu/

4.1 Maximum Entropy models

We have adopted an approach based on logistic regression classification models, which corresponds to the maximum entropy classification for independent events, firstly applied to natural language problems in (Berger, Pietra, and Pietra, 1996). This approach provides a clean way of expressing and combining different aspects of the information, and naturally implements feature selection. That is specially useful for twitter data, in which a large number of sparse features are used. A ME model estimates the conditional probability of the events given the corresponding features. Let us consider the random variable $y \in C$ that can take k different values, corresponding to the classes c_1, c_2, ... ,c_k. The ME model is given by the following equation:

$$P(c|d) = \frac{1}{Z_\lambda(F)} \times exp\left(\sum_i \lambda_{ci} f_i(c,d)\right) \quad (1)$$

determined by the requirement that $\sum_{c\in C} P(c|d)$=1. $Z_\lambda(F)$ is a normalizing term, used just to make the exponential a true probability, and is given by:

$$Z_\lambda(F) = \sum_{c'\in C} exp\left(\sum_i \lambda_{c'i} f_i(c',d)\right) \quad (2)$$

f_i are feature functions corresponding to features defined over events, and $f_i(c,d)$ is the feature defined for a class c and a given observation d. The index i indicates different features, each of which has associated weights λ_{ci}, one for each class. The ME model is estimated by finding the parameters λ_{ci} with the constraint that the expected values of the various feature functions match the averages in the training data. These parameters ensure the maximum entropy of the distribution and also maximize the conditional likelihood $\prod_i P(y^{(i)}|d^{(i)})$ of the training samples. Decoding is conducted for each sample individually and the classification is straightforward, making it interesting for on-the-fly usage. ME is a probabilistic classifier, a generalization of Boolean classification, that provides probability distributions over the classes. The single-best class corresponds to the class with the highest probability, and is given by:

$$\hat{c} = \underset{c\in C}{argmax}\, P(c|d) \quad (3)$$

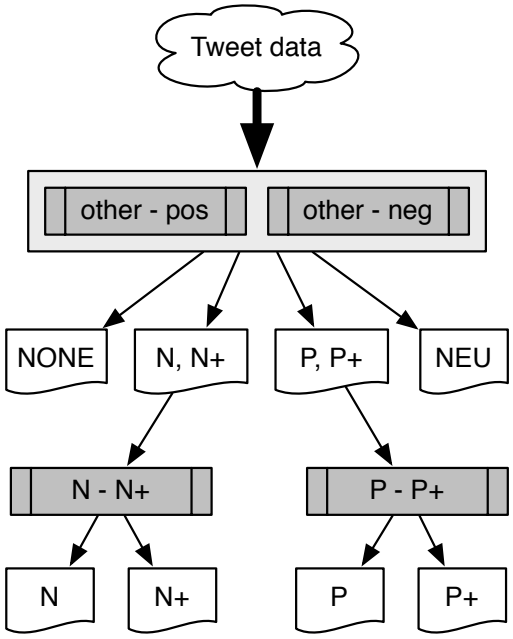

Figure 1: Approach for sentiment analysis.

The ME models used in this study are trained using the `MegaM` tool (Daumé III, 2004), which uses an efficient implementation of conjugate gradient (for binary problems).

4.2 Sentiment analysis

As previously mentioned, the sentiment classification considers 6 possible classes: *N*, *N+* → negative polarity; *P*, *P+* → positive polarity; *NEU* → contains both positive and negative sentiments; *NONE* → without polarity information. The plus sign (+) signals the sentiment intensity.

The first interesting results were achieved by combining 5 different binary classifiers, one for each class. A first classifier <`NONE, other`> was used to discriminate between NONE and all the other classes. Two other classifiers <`other, neg`> and <`other, pos`> were applied after the first classifier for detecting negative and positive sentiments, respectively. These two latest classifiers make it possible to distinguish between three classes: *Positive*, *Negative*, and *Neutral*. These three classifiers, one can now discriminate between four classes: *NONE*, *Negative*, *Positive* and *Neutral*. Finally, two other classifiers: <`N, N+`> and <`P, P+`>, allow perceiving the sentiment intensity. Only tweets annotated as *N* and *N+* were used for training the <`N, N+`> classifier, and only tweets marked as *P* or *P+* were used for training the second. That is different from the first three classifiers, which have used all the available data for training.

After some other experiments, we observed

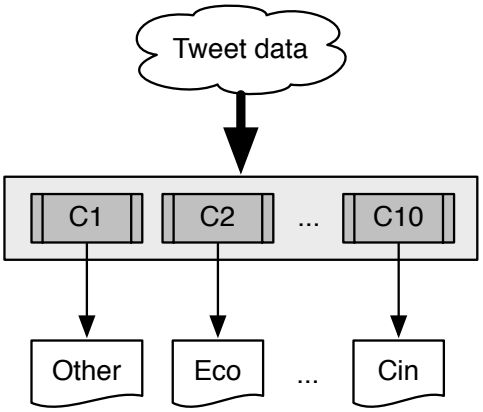

Figure 2: Approach for topic classification.

that similar results can be achieved by using the second and third classifiers to also indicate if no sentiment was present and then eliminating the need of the first classifier. The idea is that the classifiers <other,N> and <other,P> can, in fact, discriminate between four classes, by considering the class NONE whenever both return "other". Figure 1 illustrates the resulting configuration, where only four binary classifiers are used in a cascade fashion.

4.3 Topic classification

Figure 2 illustrates the classification process, where 10 distinct binary classifiers have been used, one for each topic. Each classifier selects its corresponding topic, which may lead to zero, one, or several topics. The number of selected topics have not been limited to a maximum, but when no topic is selected, the most probable topic is chosen based on the available classification probabilities.

5 Experiments

This section describes the steps taken, the features that have been used, and experiments that have been conducted using the previously described approaches.

5.1 Tweet content pre-processing

The content of each tweet was firstly tokenized using *twokenize*, a tokenization tool for English tweets[2], with some minor modifications for dealing with Spanish data instead of English.

5.2 Features

The following features, concerning the tweet text, were used for each tweet:

- Punctuation marks.
- Words occurring after the words "*nunca*" (never) or "*no*" (no) were prefixed by "NO_" until reaching some punctuation mark or until reaching the end of the tweet content (Pang, Lee, and Vaithyanathan, 2002).
- Each token starting with "http:" was converted into the token "HTTP", and it's weight as a feature was reduced.
- All tokens starting with "#" were expanded into two features: one with "#", and other without it. A lesser weight was given to the stripped version of the token.
- All tokens starting with "@" were used as feature, but the feature "@USER" was introduced as well, with a smaller weight.
- All words containing more than 3 repeating letters were also used. Whenever such words occur, two more features are produced: "LONG_WORD" with a lower weight, and the corresponding word without repetitions with a high weight (3 times the standard weight).
- All cased words were used, but the corresponding lowercase words were used as well. Uppercase words were assigned also to a higher weight, since they are often used for emphasis.

Apart from the features extracted from the text, two more features were used:

- *Username* of the author of the tweet.
- *Usertype*, corresponding to the user classification, according to `users-info.xml`.

Most of the previously described features were used both for sentiment analysis and for topic detection. Some of them were combined as bigrams for some experiments. Feature bigrams involve the following tokens: *HTTP*, words starting with # without the diacritic #, *@USER*, *LONG_WORD*, all other words converted to lowercase.

5.3 Results for sentiment analysis

Our experiments for sentiment analysis consider 6 possible classes, as described in Section 4.2. The Initial experiments achieved 52.5 Acc (Accuracy) in the development set, using all previously described features except punctuation, tweet's author name, and the *user type*. This baseline result was then further improved to 53.6 Acc [+1.1] by using the tweet's author name, and by adding the *user type* it was further improved to

[2]By Brendan O'Connor (brenocon@gmail.com)

	development	test
Unigrams only	55.2	63.4
Unigrams, Bigrams	53.8	62.2
Sentiment lexicon	54.8	63.2

Table 1: Submitted runs (Accuracy).

54.2 [+0.6]. The best results in our development set were achieved by also providing punctuation marks as features: 55.2 Acc [+1].

After establishing the feature set, bigrams and a sentiment lexicon were also tested as additional resources. Table 1 summarizes the obtained results for the development and test sets, revealing that results over the development set are consistent with results over the test set. However, we have concluded that sentiment lexicons and bigram-based features turned out not to be helpful the way they have been used. Nevertheless, differences were not statistical significant using Wilcoxon signed-rank test.

5.4 Results for topic classification

The evaluation performed in the scope of the TASS challenge assumes that the set of topics manually labeled for each tweet must be matched (Villena-Román et al., 2012). For example, if a tweet was previously marked with topic t_1 and t_2, then the system must also suggest the same set of two topics.

Differences across experiments are always subtle, because improvements in one classifier may worsen results in another classifier. In terms of feature usage, experiments revealed that adding the author's name produced slightly better results but, contrarily to what was expected, providing the *user type* as a feature did not improve results. Adding punctuation marks decreased the overall performance. The best combination of features, using unigrams, led to 43.2 Acc in the development set and 64.9 Acc in the test set.

Apart from the previous evaluation, we have also performed evaluations for each topic individually. Table 2 shows the corresponding results for the test set, sorted by SER (Slot Error Rate) (Makhoul et al., 1999) performance. The first column shows the number of correct classifications, and the other columns show the corresponding *Precision*, *Recall*, *F1-measure*, and SER, respectively.

Similarly to what has been done for sentiment analysis, we have also performed experiments that combined features as bigrams. That strategy proved to be a good solution for the test set (65.4 Acc [+0.5]), but not so good for the development set (42.5 Acc [-0.7]). However, a deeper analysis on the test set, considering each topic individually have revealed that such strategy increases the recall but decreases the precision, leading to lower F1-measure [-0.5%].

Topic	**Cor**	**Prec**	**Rec**	**F1**	**SER**
política	26830	0.89	0.87	0.88	0.237
literatura	45	0.48	0.44	0.46	0.239
música	1345	0.90	0.47	0.62	0.268
deportes	70	0.52	0.63	0.57	0.271
tecnología	205	0.71	0.63	0.67	0.274
cine	418	0.70	0.44	0.54	0.287
fútbol	444	0.54	0.65	0.59	0.299
entreten.	5055	0.93	0.48	0.63	0.357
economía	2212	0.87	0.47	0.61	0.379
otros	18039	0.64	0.91	0.75	0.442
Total	54663	0.79	0.77	0.78	0.442

Table 2: Separate results per topic.

Figure 3 shows the confusion matrix for topic detection (each topic is represented by its first letter, except for "entertainment" which is represented by ET). As expected the highest values appear in the diagonal of the matrix. However, it is possible to observe that topic "others" is frequently assigned to tweets classified in the reference with more than one topic, being "others" one of them. It is also possible to observe that "others" is also incorrectly predicted for tweets classified in the reference as "movies", "economics", "entertainment", and "music", something that it is not very surprising. Also expected are the misclassifications of "economics" as "politics". For the construction of the confusion matrix, we have used evaluation criterion of the TASS challenge, but it also possible to perceive that for classification with more than one topic, in general, our approach correctly predicts at least one of the topics (usually by predicting only one of the reference topics).

6 Conclusions

The paper describes a shared classification approach that has been applied to automatic sentiment analysis and topic classification over Spanish Twitter data. The strategy, based on binary maximum entropy classifiers, is easy to develop, offer fast convergence ratios, can be executed in parallel, and is language independent, except for the detection of the negation. A cascade of binary classifiers was used for discriminating between six possible sentiment classes, and a one-vs-all strategy was used for topic classification, where the most probable topics for each tweet were

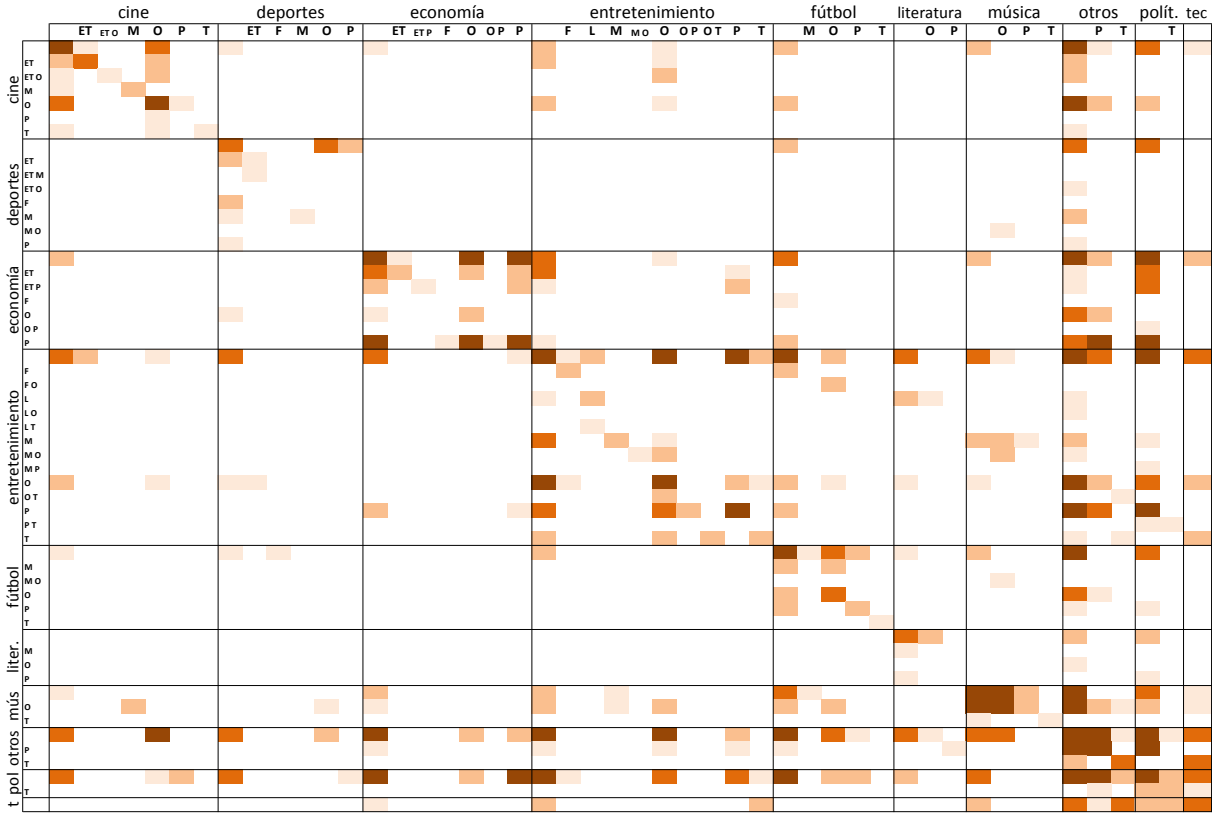

Figure 3: Topics confusion matrix (higher values darker; lower values lighter).

selected. The developed system achieved the best results for topic classification (+5.2 Acc, with statistical significance using the Wilcoxon signed-rank test: $W = 9425$, $p < 0.001$), and the second place for sentiment analysis (-1.9 Acc, without statistical significance, also using the Wilcoxon signed-rank test) in a joint evaluation effort (Villena-Román et al., 2012). In what concerns sentiment analysis, our experiments have shown that knowledge about the author and punctuation marks contribute to improved results. However, using bigram-based features and sentiment lexicons did not show a positive contribution with our setup. In what concerns the topic classification, the author type did not show a strong contribution, contrarily to what was expected.

Future experiments will make use of the remainder information available. The sentiment polarity type (AGREEMENT, DISAGREEMENT), together with other information about the user (e.g. number of tweets, followers, and following), will probably have impact on the results. Another possible direction is to automatically learn lexicons from the data and use them as an additional source of information.

Acknowledgements

This work was partially supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under project PEst-OE/EEI/LA0021/2011, and by DCTI - ISCTE-IUL – Lisbon University Institute.

References

Allan, James, editor. 2002. *TOPIC DETECTION AND TRACKING: Event-based Information Organization*. Kluwer Academic Publishers.

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

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

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.

Daumé III, Hal. 2004. Notes on CG and LM-BFGS optimization of logistic regression. http://hal3.name/megam/.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford University.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the* 10*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177. ACM.

Kasiviswanathan, S. P., P. Melville, A. Banerjee, and V. Sindhwani. 2011. Emerging Topic Detection using Dictionary Learning. In *CIKM'11: Proceedings of the 20th ACM international conference on Information and Knowledge Management*, pages 745–754. ACM.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*. ACL.

Lee, K., D. Palsetia, R. Narayanan, M. Patwary, A. Agrawal, and A. Choudhary. 2011. Twitter trending topic classification. In *International Conference on Data Mining Workshops (ICDMW)*, pages 251–258. IEEE.

Lin, C., Y. He, R. Everson, and S. Rüger. 2012. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions On Knowledge And Data Engineering*, 24(6):1134–1145.

Makhoul, J., F. Kubala, R. Schwartz, and R. Weischedel. 1999. Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM'10)*.

Pang, Bo, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Perez-Rosas, V., C. Banea, and R. Mihalcea. 2012. Learning Sentiment Lexicons in Spanish. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC'12)*. ELRA.

Sriram, Bharath, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short Text Classification in Twitter to Improve Information Filtering. In *SIGIR'10: Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 841–842. ACM.

Stone, P., D. Dunphy, M. Smith, and D. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Turney, Peter 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*, pages 417–424. ACL.

Villena-Román, J., J. García-Morera, C. Moreno-Garcia, L. Ferrer-Ureña, S. Lana-Serrano, J. C. González-Cristobal, A. Westerski, E. Martínez-Cámara, M. A. García-Cumbreras, M. T. Martín-Valdivia, and Ureña-López L. A. 2012. TASS-Workshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50.

Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354. ACL.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 85-92 recibido 01-12-12 revisado 14-02-13 aceptado 19-02-13

Sistema cross-lingüe de acceso inteligente a la información de casos clínicos mediante dispositivos móviles

Cross-lingual intelligent information access system from clinical cases using mobile devices

María Lorena Prieto, Fernando Aparicio, Manuel de Buenaga, Diego Gachet, Mari Cruz Gaya
Grupo de Sistemas Inteligentes, DSIAC, Escuela Politécnica
Universidad Europea de Madrid
C/ Tajo, s/n. Villaviciosa de Odón.
{marialorena.prieto,fernando.aparicio,buenaga,gachet,mcruz@uem.es}

Resumen: A lo largo de la última década se está produciendo un crecimiento vertiginoso tanto del desarrollo de nuevas tecnologías móviles inteligentes (dispositivos Smartphone y Tablet) como de su uso (a través de gran cantidad de aplicaciones). Por otro lado, en el ámbito biomédico cada vez existe un número mayor de recursos, en diferentes formatos, que pueden ser explotados haciendo uso de Sistemas Inteligentes de Acceso a la Información y técnicas para la recuperación y la extracción de información. En este artículo se presenta el desarrollo de una interfaz de acceso móvil que, haciendo uso de diferentes fuentes de conocimiento locales (diccionarios y ontologías previamente pre-procesadas), técnicas de procesamiento de lenguaje natural y fuentes de conocimiento remotas (con las que se realiza la anotación de entidades en el texto introducido en el sistema a través de servicios Web), permite la extracción cruzada (cross-lingue) de conceptos médicos en Inglés y en Español, a partir de un texto médico en inglés o en español (e.g. un caso clínico). El usuario de la aplicación móvil puede introducir un texto médico o una imagen del mismo, obteniendo como resultado un conjunto de entidades médicas relevantes. Sobre las entidades médicas reconocidas, extraídas y mostradas a través de la interfaz, el usuario puede obtener más información de las mismas, obtener más información de otros conceptos relacionados con los extraídos originalmente y localizar publicaciones científicas procedentes de MEDLINE/PubMed.
Palabras clave: Smartphone, Tablet, dispositivos móviles, extracción de conceptos médicos, reconocimiento de entidades nombradas, CLIR.

Abstract: Over the last decade there has been a rapid growth of both the development of new smart mobile devices (Smartphone and Tablet) and their use (through many applications). Furthermore, in the biomedical field there are a greater number of resources in different formats, which can be exploited by using Intelligent Information Access Systems and techniques for information retrieval and extraction. This paper presents the development of a mobile interface access that, using different local knowledge sources (dictionaries and ontologies previously preprocessed), techniques of natural language processing and remote knowledge sources (which performs the annotation of entities in text inputted into the system via Web services), allows the cross-lingual extraction of medical concepts in English and Spanish, from a medical text in English or Spanish (e.g. a clinical case). The mobile application user can enter a medical text or a picture of it, resulting in a set of relevant medical entities. On recognized medical entities, extracted and displayed through the interface, the user can get more information on them, get more information from other concepts related to originally extracted and search for scientific publications from MEDLINE/PubMed.
Keywords: Smartphone, Tablet, mobile devices, medical concept extraction, named entity recognition, CLIR.

ISSN 1135-5948

1 Introducción

El auge en el uso de dispositivos móviles, así como el desarrollo de aplicaciones para estos, ha sido especialmente significativo en la última década. Esto se debe a la aparición de dispositivos móviles inteligentes como son el Smartphone y el Tablet, cuyo número de usuarios se ha incrementado provocando que el de usuarios de teléfonos básicos y de gama media descienda.

En el dominio biomédico se ha experimentado un igual crecimiento del interés en este tipo de aplicaciones, orientándose a satisfacer necesidades de los diferentes perfiles de usuarios básicos en él: principalmente, profesionales, estudiantes de medicina y pacientes. Una investigación reciente relacionada, sobre artículos en PubMed (en abril de 2011) vinculados con aplicaciones enfocadas al sector sanitario (Mosa, Yoo, y Sheets, 2012), estudiaba de forma similar a como señalábamos anteriormente su diferente orientación a profesionales, estudiantes de medicina y pacientes. De forma representativa, en ella se seleccionaron 55 artículos, de un total de 2.894 encontrados, en los que se hacía referencia a 83 aplicaciones: 57 para profesionales, 11 para estudiantes y 15 para pacientes.

La recuperación de información cross-lingue (CLIR) está relacionada con la recuperación de recursos en idiomas distintos a los empleados por los usuarios (Goker y Davies, 2009), existiendo iniciativas ampliamente conocidas que afrontan el problema proponiendo diferentes competiciones (e.g. CLEF[1], TREC[2], NTCIR[3] o FIRE[4]).

Desde un punto de vista meramente estructural, hacer uso de ontologías para almacenar el conocimiento en los sistemas de extracción de información es ventajoso, ya que facilita tanto la actualización de este conocimiento (al separarlo de otros componentes), como la portabilidad a otros dominios (Karkaletsis et al. 2011). Uno de los temas que han comenzado a investigarse recientemente estudia cómo las ontologías pueden ayudar a cubrir necesidades de CLIR, generándose nuevos sistemas basados en ontologías multilingües (e.g. Gracia et al. 2012; Embley et al. 2011). Por otro lado, a pesar de que la creación de ontologías multilingües a partir de monolingües puede ser altamente costoso, diferentes equipos de investigadores están trabajando para solventar estas dificultades desde diferentes puntos de vista (e.g. Choi et al. 2011; Sellami et al. 2011).

Los Sistemas de Acceso Inteligente a la Información (Armano et al. 2010; Martín-Valdivia et al. 2009) pueden aprovechar el gran desarrollo de recursos, monolingües y multilingües, que se está produciendo en el dominio biomédico, para integrar información desde diferentes fuentes heterogéneas, haciendo uso de las técnicas de procesamiento de lenguaje natural ya existentes. Estos sistemas, además, pueden tener aplicaciones muy directas para los usuarios finales, dando resultados en tiempos razonables desde servicios online.

La aplicación móvil presentada en este artículo[5], permite al usuario procesar un texto médico (e.g. un caso clínico) o bien insertándolo directamente en la aplicación (a través del navegador de su Smartphone o Tablet) o bien haciendo una foto al texto, para posteriormente acceder a un sistema cross-lingue de extracción de conceptos y acceso inteligente a la información (CLEiM, disponible en [6]). Este sistema es la evolución de una versión monolingüe presentada en otros trabajos (Aparicio et al., 2011; Aparicio et al. 2012). En concreto, el sistema ha sido evaluado en el contexto del EEES, dando soporte a una actividad de aprendizaje con estudiantes de medicina basada en un caso clínico, obteniéndose: *(i)* la percepción de los usuarios al utilizar el sistema; *(ii)* una comparativa de los resultados de una prueba objetiva, basada en el caso, de un conjunto de estudiantes utilizando todas las fuentes de Internet frente a otro conjunto utilizando el sistema monolingüe.

En la sección 2 se presenta un análisis básico sobre la rápida evolución tecnológica y de uso que se está produciendo en cuanto a las tecnologías móviles, particularizándolo para el dominio biomédico y su utilización en educación. En la sección 3 se describen la nueva funcionalidad móvil del sistema. En la sección 4 se exponen casos de uso para el acceso desde Smartphone y Tablet que ofrece la interfaz presentada. Por último, en la sección 5

[1] www.clef-initiative.eu
[2] trec.nist.gov
[3] research.nii.ac.jp/ntcir/index-en.html
[4] www.isical.ac.in/~clia

[5] orion.esp.uem.es/Mobile
[6] cleim.sourceforge.net

se mencionan algunas conclusiones y líneas de trabajo futuras.

2 Uso de dispositivos móviles y aplicaciones médicas

En los últimos años se ha observado una evolución en el uso de los dispositivos móviles desde la aparición de los teléfonos inteligentes o Smartphone, de forma tal que el número de usuarios de dispositivos móviles básicos o de gama media ha ido descendiendo, mientras que el de Smartphone y Tablet se ha incrementado.

En la Figura 1 presentamos datos representativos de esta evolución, para el periodo reciente de Septiembre de 2010 a Febrero de 2012. Los datos de la figura han sido tomados de estudios realizados trimestralmente por Empirica Influentials & Research[7], comScore[8], IAB Spain[9] y AIMC10 (Asociación para la Investigación de Medios de Comunicación). En la figura puede apreciarse el descenso de los usuarios de móvil básico o sin móvil, así como los de gama media, y un continuo crecimiento de usuarios de Smartphone y Tablet.

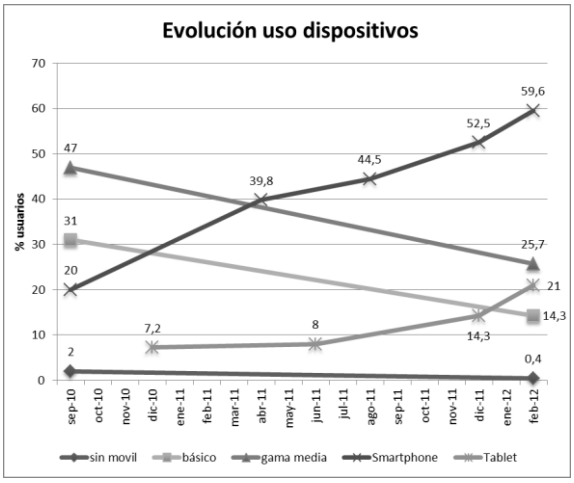

Figura 1: Evolución del uso de dispositivos móviles

Centrándonos ahora en los usuarios del sector sanitario, también el uso de estos dispositivos cada vez está más presente entre ellos. Es de gran importancia que los profesionales de la medicina dispongan de una buena comunicación y colaboración constante entre ellos, por ello se comenzó por emplear dispositivos "busca", de forma que mejoraran estos aspectos. Este fue sustituido por teléfonos móviles y dispositivos PDA en los 90 (Burdette, Herchline y Oehler, 2008), con lo que agilizaban la comunicación y, además, mejoraban la organización y el acceso a la información (Buenaga et al., 2008). En la actualidad tenemos las funcionalidades del busca, teléfono móvil y PDA en un único dispositivo: el Smartphone, por lo que se ha extendido su uso entre el personal sanitario (Wu et al., 2010).

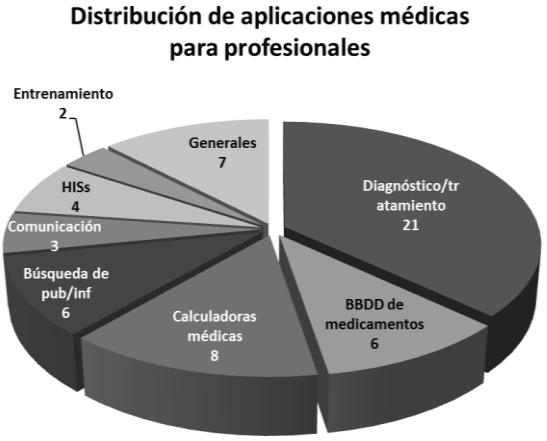

Figura 2: Distribución de aplicaciones médicas para profesionales

En torno a la utilización de Smartphone, en 2009 se realizó un estudio sobre médicos de EEUU[11], dando como resultado un 64% de usuarios, un 34% más que en 2001. Este estudio también predecía que en 2012 se incrementaría hasta el 81% de usuarios médicos. Además, en los últimos años se han implantado en el sistema hospitalario diversas tecnologías como: Sistemas de Información Hospitalaria (HISs), Sistemas Clínicos de Apoyo de Decisión (CDSSs), de Almacenamiento de Imágenes y Sistemas de Comunicación (PACSs), Sistemas de Información de Laboratorio (LISs), etc.; recursos de publicaciones como PubMed y UpToDate; aplicaciones clínicas como calculadoras médicas, bases de datos de medicamentos, y aplicaciones de diagnóstico de

[7] marketingyconsumo.com/estudio-sobre-el-uso-de-smartphones-en-espana.html

[8] www.comscore.com

[9] www.iabspain.net/noticias/investigacion/iab-spain-lanza-su-iii-estudio-sobre-mobile-marketing

[10] www.aimc.es/-Navegantes-en-la-Red-.html

[11] manhattanresearch.com/News-and-Events/Press-Releases/physician-smartphones-2012

enfermedades; y nuevos sistemas de comunicación (Junglas, Abraham y Ives, 2009). La implantación de estas tecnologías y el incremento del uso de Smartphone por profesionales sanitarios ha provocado que se hayan desarrollado diversas aplicaciones con el fin de mejorar estos sistemas (Mosa, Yoo, y Sheets, 2012). En la Figura 2 se puede apreciar que las más utilizadas por profesionales son las de diagnóstico de enfermedades, referencias de medicamentos y calculadoras médicas.

Hay estudios que valoran el uso de dispositivos móviles y ordenadores portátiles en la educación en medicina (Briscoe et al., 2006; Grasso, Yen, y Mintz, 2006). E investigaciones sobre e-Learning han demostrado la eficacia de estas tecnologías (Cook, 2009).

En (Davies et al., 2012) se realizó un estudio para conocer como, estudiantes de medicina, usaban dispositivos móviles para mejorar su aprendizaje. El 78% de los encuestados tenía un ordenador personal (302 estudiantes de un total de 387). El 38% ya disponía de un dispositivo móvil. Las impresiones fueron que utilizar estos dispositivos en la educación médica sería de utilidad dado el acceso y movilidad casi instantáneos. El 98% indicó que le gustaría que la iniciativa continuara en sus Escuelas. El Smartphone es adecuado para el e-Learning dado que su uso se puede distribuir a través de las redes telefónicas (i.e. alternativa a las clásicas redes de ordenadores en países en desarrollo).

3 Arquitectura del sistema

3.1 Componentes del sistema

La aplicación móvil que hemos desarrollado[12], permite al usuario procesar un texto médico (principalmente casos clínicos, y orientado a su uso por estudiantes de medicina) utilizando un sistema cross-lingüe de extracción de conceptos y acceso inteligente a la información (CLEiM, disponible en [13]). Este sistema es la evolución de una versión monolingüe presentada en otros trabajos (Aparicio et al., 2011; Aparicio et al. 2012).

El sistema tiene una arquitectura diseñada de forma que facilita la integración de nuevos componentes (Aparicio et al., 2011). Los módulos principales de esta arquitectura son los siguientes:

Módulo de acceso: módulo en el que se encuentran las diferentes herramientas con las que acceder al sistema, mediante el protocolo HTTP, ya sea a través de la interfaz o de servicios Web. Además, comunica los módulos de procesamiento del lenguaje natural y de búsqueda, explicados a continuación.

Módulo de procesamiento de lenguaje natural: módulo que, empleando librerías de GATE[14] y servicios de anotación del NCBO[15], procesa el texto introducido extrayendo los conceptos.

Módulo de recuperación de información: módulo que gestiona los conceptos en tiempo de ejecución en el módulo anterior, para optimizar la ejecución online.

Módulo de búsqueda de la información: módulo que suministra las fuentes en las que se realiza la búsqueda de información de los conceptos extraídos tras procesar el texto. Hasta la fecha, las fuentes empleadas son Freebase, MedlinePlus y PubMed.

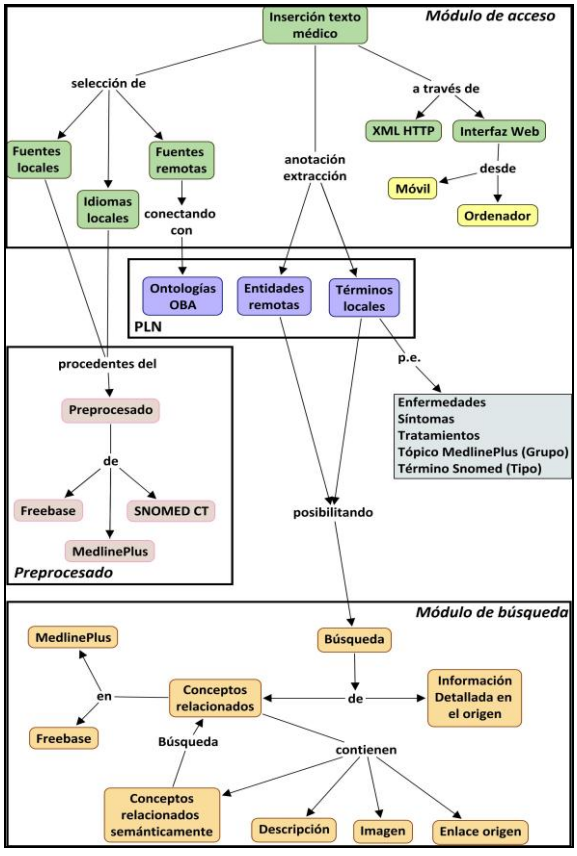

Figura 3: Relaciones entre los elementos principales de la arquitectura del sistema y funcionalidades

[12] orion.esp.uem.es/Mobile
[13] cleim.sourceforge.net
[14] gate.ac.uk
[15] www.bioontology.org

En la adaptación se ha mantenido esta arquitectura, incorporando al módulo de acceso un componente nuevo, consistente en páginas jsp y servlets, que generan las distintas funciones de la interfaz accesibles desde dispositivos móviles. Además, este componente integra un reconocimiento óptico de caracteres (OCR), explicado en detalle en el apartado 3.3.

La interfaz de la aplicación web ha sido desarrollada utilizando el framework orientado a dispositivos móviles JQuery Mobile[16]. Para el acceso a la cámara del dispositivo se usa la extensión HTML Media Capture[17] de HTML5, realizándose el procesado de la imagen y el texto en el servidor. En la Figura 3 se detallan las funcionalidades del sistema una vez realizada la adaptación.

3.2 Análisis semántico

El sistema tiene como objetivo principal analizar casos clínicos, extrayendo los conceptos médicos relevantes y mostrando información relacionada a través de ontologías. Un caso clínico, según (Real Academia Nacional de Medicina, 2012), es la descripción ordenada tanto de los acontecimientos que le ocurren a un paciente en el curso de una enfermedad, como de los datos complementarios proporcionados por los procedimientos diagnósticos, el curso del razonamiento clínico, la conclusión diagnóstica, el tratamiento empleado y la evolución del enfermo. El uso de casos clínicos es muy común en la educación en medicina, dado que permite demostrar la aplicación de conceptos teóricos, por lo que el desarrollo de una herramienta que permita un análisis de los mismos, supone una gran ayuda para el proceso de aprendizaje. La principal fuente de casos clínicos empleada en nuestras evaluaciones del sistema y en evaluaciones con usuarios está disponible en la escuela de medicina de la Universidad de Pittsburgh[18].

En los últimos años se han desarrollado diversas ontologías biomédicas, como GO (Gene Ontology), UMLS (Unified Medical Language System) o SNOMED.CT (Systematized Nomenclature of Medicine – Clinical Terms) (Muñoz et al., 2012). Además, el NCBO está trabajando en herramientas y servicios para el uso con estas ontologías.

En este sistema se han empleado, principalmente, dos fuentes de información: Medlineplus como fuente con contenido más formal, y Freebase como fuente con contenido menos formal. Para la información de publicaciones científicas se ha empleado PubMed, dado que es una de las fuentes más utilizadas en la actualidad.

El incluir Freebase como fuente a pesar de disponer de un contenido menos formal que Medlineplus se debe a que es una base de datos colaborativa donde la información almacenada se encuentra estructurada, agrupada por dominios (entre ellos el de medicina) e identificada por tipos, y propiedades, como pueden ser, en el caso de la medicina, “disease”, “symptoms”, “treatments”, “risk factors”, etc. (Aparicio et al., 2011). Además, la calidad de esta información es alta, y está disponible en varios idiomas. Se ha utilizado Freebase para la obtención de listas de conceptos que son pre-procesados antes de ser utilizados por el componente Gazetteer del sistema ANNIE incluido en GATE.

Otro aspecto que se ha tenido en cuenta en el desarrollo de este sistema es la recuperación de información cross-lingüe (CLIR), relacionada con la recuperación de información en diferentes idiomas. Dado que las ontologías permiten el almacenamiento de datos en diferentes idiomas, conectando los conceptos mediante meta-información, las convierte en una herramienta muy útil para desarrollar sistemas cross-lingüe, como sucede con Freebase y Medlineplus.

Respecto al procesamiento del texto de los casos clínicos en Español e Inglés, un elemento relevante en nuestro sistema es la forma de obtención de la terminología en Español e Inglés con que se ha construido el analizador basado en GATE. Por un lado, la información es extraída de Freebase realizando una consulta MQL a través de una clase Java, cuya respuesta está en formato JSON y, cada tipo (enfermedad, síntoma o tratamiento), es solicitado tanto en Inglés como en Español, disponible en el sistema. Por otro lado, la información de Medlineplus está disponible en ambos idiomas. Nuestro componente Java procesa el XML que devuelve al realizar una petición HTTP, en el que se indica el idioma de la información.

[16] jquerymobile.com
[17] dev.w3.org/2009/dap/camera
[18] path.upmc.edu/cases.html

3.3 Procesamiento de imágenes

En el sistema hemos incorporado también la funcionalidad de procesar el texto a través de imágenes. El usuario puede hacer una fotografía o captura de pantalla del texto y directamente extraer y navegar por los conceptos clave.

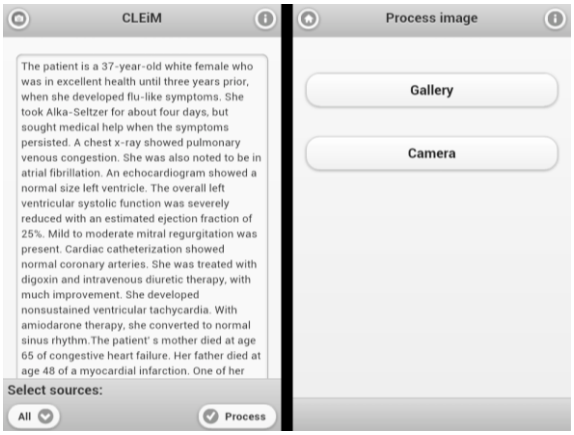

Figura 4: Páginas para enviar el texto al sistema.

Existen diversos sistemas OCR: páginas web en las que un usuario sube una imagen y se muestra el texto contenido en ella, servicios API para utilizar el OCR desde una aplicación propia o aplicaciones de escritorio. El software integrado en nuestro sistema es el motor tesseract-ocr de Google (distribuido bajo la licencia de Apache). El procesamiento es rápido (1 segundo para imágenes a color y 0.82 para imágenes en escala de grises) y tiene una tasa de acierto del 61% para imágenes de color y del 70% para imágenes en escala de grises, según se probó en (Patel, Patel, y Patel, 2012). Además, da soporte a diversos idiomas, aunque en nuestro caso solo precisemos de los paquetes de inglés y español.

4 Métodos de acceso y caso de uso

El sistema ha sido pensado para el uso en estudiantes de medicina. Concretamente para extraer información de casos clínicos, muy empleados en la educación de esta disciplina.

Para valorar su utilidad en este escenario de aplicación, se realizó una evaluación del sistema con 60 estudiantes de segundo grado en medicina, realizando un cuestionario de preguntas de tipo test, relacionadas con un caso clínico seleccionado previamente, con el que se midió cuantitativamente los conocimientos adquiridos por los estudiantes tras realizar la prueba. Los resultados de esta evaluación se pueden encontrar detallados en (Aparicio et al., 2012). Esta prueba se realizó con una versión del sistema CLEiM monolingüe, en inglés y con una interfaz inicial orientada a ordenadores de sobremesa.

La versión actual del sistema, realiza la recuperación de información cross-lingüe a través de textos o imágenes en español e inglés.

La aplicación permite escribir o pegar un texto directamente del portapapeles, seleccionar la fuente de procesamiento (Freebase y/o Medline Plus para esta versión) y enviar dicho texto al servidor para realizar la extracción de conceptos médicos. También dispone de la opción de seleccionar una imagen de la galería o acceder a la cámara del dispositivo para tomar dicha imagen en el momento (ver Figura 4).

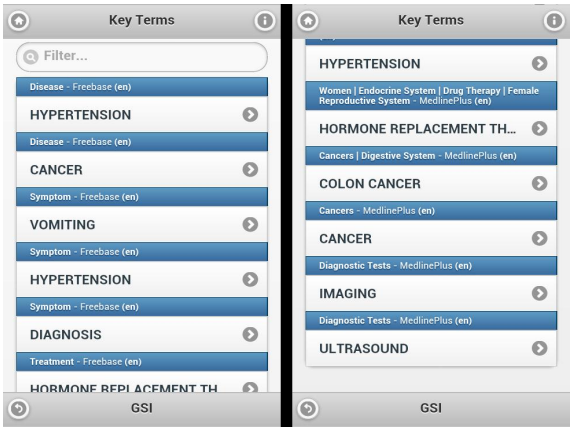

Figura 5: Página de listado de conceptos extraídos del texto.

Una vez procesado el texto en el servidor, se muestra el listado de conceptos extraídos (ver Figura 5), indicando la fuente de procedencia de la información y el tipo (síntoma, tratamiento o enfermedad, en el caso de Freebase).

Si es seleccionado un concepto cuya fuente es Freebase (que en el momento actual, no dispone de interfaz móvil) se muestra un listado de conceptos (enfermedades, tratamientos y síntomas) relacionados con el concepto seleccionado, incluyendo una imagen, y pudiendo acceder a información detallada de cada uno (ver Figura 6).

La información del concepto varía en función de si se trata de una enfermedad, un síntoma o un tratamiento. Si el concepto seleccionado procede de la fuente MedlinePlus, el listado de conceptos no incluye imagen y la

información detallada es distinta, enlazándose también un listado de publicaciones relacionadas y extraídas de PubMed, como se observa en la Figura 7.

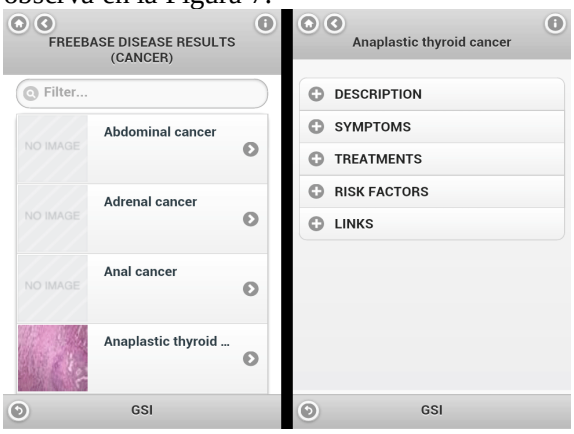

Figura 6: Listado de conceptos de Freebase e información detallada de una enfermedad.

5 *Conclusiones y trabajo futuro*

El aprovechamiento de, por un lado, la gran cantidad de información biomédica disponible y, por otro lado, la mejora de los algoritmos y herramientas para la extracción de información a partir de textos escritos en lenguaje natural, puede dar lugar a aplicaciones que tengan una utilidad para usuarios finales en diferentes contextos, tal y como puede ser el educativo o los servicios de salud para médicos y pacientes. Además, dado que esta proliferación también se está dando en otros dominios de conocimiento, el uso de estructuras para el almacenamiento de los datos como las ontologías, puede facilitar la extensión de estas aplicaciones a otros campos y usuarios.

En este artículo se ha presentado una aplicación que permite el acceso a diferentes recursos médicos, a partir de un texto médico introducido desde un dispositivo móvil, incorporando además un OCR para facilitar la introducción del texto al usuario a través de la cámara. La obtención de estos conceptos está basada tanto en terminologías preprocesadas a partir de diccionarios existentes y recursos ontológicos, así como de servicios Web que proporcionan anotaciones en tiempo real.

Para trabajos futuros, estamos trabajando fundamentalmente en las siguientes direcciones: se están elaborando nuevas evaluaciones con usuarios que hagan uso de sus dispositivos móviles; se está estudiando la aplicación de nuevas técnicas de procesado del texto y la incorporación de nuevos recursos, manteniendo la capacidad del sistema para dar respuestas en línea; por último, se está valorando la incorporación de elementos de procesado en los sistemas móviles, así como el desarrollo de aplicaciones instalables que posibiliten la distribución de los componentes entre el sistema móvil y el servidor.

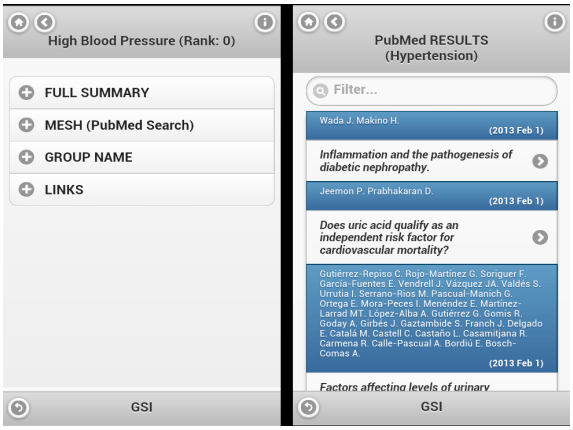

Figura 7: Páginas de información de Medline Plus y de publicaciones de PubMed

Referencias bibliográficas

Aparicio, F., Buenaga, M., Rubio, M., Hernando, M. A., Gachet, D., Puertas, E., y Giráldez, I. (2011). TMT: A tool to guide users in finding information on clinical texts. *Procesamiento de Lenguaje Natural*, *46*(0), 27-34.

Aparicio, F., Buenaga, M., Rubio, M., y Hernando, A. (2012). An intelligent information access system assisting a case based learning methodology evaluated in higher education with medical students. *Computers & Education*, *58*(4), 1282-1295.

Armano, G., Gemmis, M., Semerano, G., y Vargiu, E. (2010). *Intelligent Information Access*. Springer.

Briscoe, G. W., Fore Arcand, L. G., Lin, T., Johnson, J., Rai, A., y Kollins, K. (2006). Students' and residents' perceptions regarding technology in medical training. *Academic psychiatry: the journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, *30*(6), 470-479.

Buenaga, M., Gachet, D., Maña, M. J., Villa, M. y Mata, J. (2008), Clustering and Summarizing Medical Documents to Improve Mobile Retrieval, *ACM-SIGIR 2008 Workshop on Mobile Information Retrieval*, *4(2)*, 54-57.

Burdette, S. D., Herchline, T. E., y Oehler, R. (2008). Practicing Medicine in a Technological Age: Using Smartphones in Clinical Practice. *Clinical Infectious Diseases*, *47*(1), 117-122.

Choi, I., Rho, S., Jeong, Y.-S., y Kim, M. (2011). Relation Extraction from Documents for the Automatic Construction of Ontologies. En J. J. Park, L. T. Yang, & C. Lee (eds.), *Future Information Technology* (pp. 21-28). Springer Berlin Heidelberg.

Cook, D. A. (2009). The failure of e-learning research to inform educational practice, and what we can do about it. *Medical teacher*, *31*(2), 158-162.

Davies, B., Rafique, J., Vincent, T., Fairclough, J., Packer, M., Vincent, R., y Haq, I. (2012). Mobile Medical Education (MoMEd) - how mobile information resources contribute to learning for undergraduate clinical students - a mixed methods study. *BMC Medical Education*, *12*(1), 1.

Embley, D. W., Liddle, S. W., Lonsdale, D. W., y Tijerino, Y. (2011). Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search. En M. Jeusfeld, L. Delcambre, & T.-W. Ling (eds.), *Conceptual Modeling – ER 2011* (pp. 147-160). Springer Berlin Heidelberg.

Goker, A., y Davies, J. (2009). *Information Retrieval: Searching in the 21st Century*. John Wiley & Sons.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., y McCrae, J. (2012). Challenges for the multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *11*(0), 63-71.

Junglas, I., Abraham, C., y Ives, B. (2009). Mobile technology at the frontlines of patient care: Understanding fit and human drives in utilization decisions and performance. *Decis. Support Syst.*, *46*(3), 634–647.

Karkaletsis, V., Fragkou, P., Petasis, G., y Iosif, E. (2011). Ontology Based Information Extraction from Text. En G. Paliouras, C. D. Spyropoulos, & G. Tsatsaronis (eds.), *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution* (Vol. 6050, pp. 89-109). Berlin, Heidelberg: Springer Berlin Heidelberg.

Martín-Valdivia, M. T., Montejo-Ráez, A., Díaz-Galiano, M. C., Perea Ortega, J. M., y Ureña-López, L. A. (2009). Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization System. In V. Prince, & M. Roche (Eds.), *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration* (pp. 38-57). Hershey, PA: Medical Information Science Reference.

Mosa, A. S. M., Yoo, I., y Sheets, L. (2012). A Systematic Review of Healthcare Applications for Smartphones. *BMC Medical Informatics and Decision Making*, *12*(1), 67.

Muñoz, R., Aparicio, F., y Buenaga, M. (2012). Sistema de Acceso a la Información basado en conceptos utilizando Freebase en Español-Inglés sobre el dominio Médico y Turístico. *Procesamiento de Lenguaje Natural*, *49*(0), 29-38.

Patel, C., Patel, A., y Patel, D. (2012). Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications*, 55(10), 50–56.

Real Academia Nacional de Medicina. (2012). Diccionario de Términos Médicos. Madrid: Panamericana.

Sellami, Z., Camps, V., Aussenac-Gilles, N., y Rougemaille, S. (2011). Ontology Co-construction with an Adaptive Multi-Agent System: Principles and Case-Study. En A. Fred, J. L. G. Dietz, K. Liu, & J. Filipe (eds.), *Knowledge Discovery, Knowlege Engineering and Knowledge Management* (pp. 237-248). Springer Berlin Heidelberg.

Wu, R. C., Morra, D., Quan, S., Lai, S., Zanjani, S., Abrams, H., y Rossos, P. G. (2010). The use of smartphones for clinical communication on internal medicine wards. *Journal of Hospital Medicine*, *5*(9), 553–559.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 93-100 recibido 17-12-12 revisado 07-02-13 aceptado 19-02-13

Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish

Análisis de Sentimiento basado en lexicones de mensajes de Twitter en español

Antonio Moreno-Ortiz, Chantal Pérez Hernández
Facultad de Filosofía y Letras
Universidad de Málaga
{amo, mph}@uma.es

Resumen: Los enfoques al análisis de sentimiento basados en lexicones difieren de los más usuales enfoques basados en aprendizaje de máquina en que se basan exclusivamente en recursos que almacenan la polaridad de las unidades léxicas, que podrán así ser identificadas en los textos y asignárseles una etiqueta de polaridad mediante la cual se realiza un cálculo que arroja una puntuación global del texto analizado. Estos sistemas han demostrado un rendimiento similar a los sistemas estadísticos, con la ventaja de no requerir un conjunto de datos de entrenamiento. Sin embargo, pueden no resultar ser óptimos cuando los textos de análisis son extremadamente cortos, tales como los generados en algunas redes sociales, como Twitter. En este trabajo llevamos a cabo tal evaluación de rendimiento con la herramienta Sentitext, un sistema de análisis de sentimiento del español.
Palabras clave: análisis de sentimiento basado en lexicones, analítica de texto, textos cortos, Twitter, evaluación de rendimiento.

Abstract: Lexicon-Based approaches to Sentiment Analysis (SA) differ from the more common machine-learning based approaches in that the former rely solely on previously generated lexical resources that store polarity information for lexical items, which are then identified in the texts, assigned a polarity tag, and finally weighed, to come up with an overall score for the text. Such SA systems have been proved to perform on par with supervised, statistical systems, with the added benefit of not requiring a training set. However, it remains to be seen whether such lexically-motivated systems can cope equally well with extremely short texts, as generated on social networking sites, such as Twitter. In this paper we perform such an evaluation using Sentitext, a lexicon-based SA tool for Spanish.
Keywords: lexicon-based sentiment analysis, text analytics, short texts, Twitter, performance evaluation.

1 Introduction[1]

1.1 Approaches to Sentiment Analysis

Within the field of sentiment analysis it has become a commonplace assertion that successful results depend to a large extent on developing systems that have been specifically developed for a particular subject domain. This view is no doubt determined by the methodological approach that most such systems employ, i.e., supervised, statistical machine learning techniques. Such approaches have indeed proven to be quite successful in the past (Pang and Lee, 2004; Pang and Lee, 2005). In fact, machine learning techniques, in any of their flavors, have proven extremely useful, not only in the field of sentiment analysis, but in most text mining and information retrieval applications, as well as a wide range of data-intensive computational tasks. However, their obvious disadvantage in terms of functionality is their limited applicability to subject domains other than the one they were designed for. Although interesting research has been done aimed at extending domain applicability (Aue and Gamon 2005), such efforts have shown limited success. An important variable for these approaches is the amount of labeled text available for training the classifier, although they perform well in terms of recall even with relatively small training sets (Andreevskaia and

[1] This work is funded by the Spanish Ministry of Science and Innovation. LingMotif Project FFI2011-25893.

Bergler, 2007).

On the other hand, a growing number of initiatives in the area have explored the possibilities of employing unsupervised lexicon-based approaches. These rely on dictionaries where lexical items have been assigned either *polarity* or a *valence*[2], which has been extracted either automatically from other dictionaries, or, more uncommonly, manually. The works by Hatzivassiloglou and McKewon (1997) and Turney (2002) are perhaps classical examples of such an approach. The most salient work in this category is Taboada et al. (2011), whose dictionaries were created manually and use an adaptation of Polanyi and Zaenen's (2006) concept of Contextual Valence Shifters to produce a system for measuring the semantic orientation of texts, which they call SO-CAL(culator). This is exactly the approach we used in our Sentitext system for Spanish (Moreno-Ortiz et al., 2010).

Combining both methods (machine learning and lexicon-based techniques) has been explored by Kennedy and Inkpen (2006), who also employed contextual valence shifters, although they limited their study to one particular subject domain (the traditional movie reviews), using a "traditional" sentiment lexicon (the General Inquirer), which resulted in the "term-counting" (in their own words) approach.The degree of success of knowledge-based approaches varies depending on a number of variables, of which the most relevant is no doubt the quality and coverage of the lexical resources employed, since the actual algorithms employed to weigh positive against negative segments are in fact quite simple.

Another important variable concerning sentiment analysis is the degree of accuracy that the system aims to achieve. Most work on the field has focused on the *Thumbs up or thumbs down* approach, i.e., coming up with a positive or negative rating. Turney's (2002) work, from which the name derives, is no doubt the most representative. A further step involves an attempt to compute not just a binary classification of documents, but a numerical rating on a scale. The rating inference problem was first posed by Pang and Lee (2005), and the approach is usually referred to as *seeing stars* in reference to this work.

1.2 Sentiment Analysis for Spanish

Work within the field of Sentiment Analysis for Spanish is, by far, scarcer than for English.

Cruz et al. (2008) developed a document classification system for Spanish similar to Turney (2002), i.e. unsupervised, though they also tested a supervised classifier that yielded better results. In both cases, they used a corpus of movie reviews taken from the Spanish Muchocine website. Boldrini et al. (2009) carried out a preliminary study in which they used machine learning techniques to mine opinions in blogs. They created a corpus for Spanish using their Emotiblog system, and discussed the difficulties they encountered while annotating it. Balahur et al. (2009) also presented a method of emotion classification for Spanish, this time using a database of culturally dependent emotion triggers.

Finally, Brooke et al. (2009) adapted a lexicon-based sentiment analysis system for English (Taboada et al., 2006, 2011) to Spanish by automatically translating the core lexicons and adapting other resources in various ways. They also provide an interesting evaluation that compares the performance of both the original (English) and translated (Spanish) systems using both machine learning methods (specifically, SVM) and their own lexicon-based semantic orientation calculation algorithm, the above mentioned SO-CAL. They found that their own weighting algorithm, which is based on the same premises as our system (see below), achieved better accuracy for both languages, but the accuracy for Spanish was well below that for English.

Our system, Sentitext (Moreno-Ortiz et al., 2010, 2011), is very similar to Brooke et al.'s in design: it is also lexicon-based and it makes use of a similar calculation method for semantic orientation. It differs in that the lexical knowledge has been acquired semi-automatically and then fully manually revised from the ground up over a long period of time, with a strong commitment to both coverage and quality. It makes no use of user-provided, explicit ratings that supervised systems typically rely on for the training process, and it produces an index of semantic orientation based on weighing positive against negative text segments, which is then transformed into a ten-

[2] Although the terms *polarity* and *valence* are sometimes used interchangeably in the literature, especially by those authors developing binary text classifiers, we restrict the usage of the former to non-graded, binary assignment, and the latter is used to refer to an *n*-point semantic orientation scale.

point scale and a five-star rating system.

2 Sentiment Analysis with Sentitext

Sentitext is a web-based, client-server application written in C++ (main code) and Python (server). The only third-party component in the system is Freeling (Atserias et al., 2006, Padró, 2011), a powerful, accurate, multi-language NLP suite of tools, which we use for basic morphosyntactic analysis. Currently, only one client application is available, developed in Adobe Flex, which takes an input text and returns the results of the analysis in several numerical and graphical ways, including visual representations of the text segments that were identified as sentiment-laden[3]. Lexical information is stored in a relational database (MySQL).

Being a linguistically-motivated sentiment analysis system, special attention is paid to the representation and management of the lexical resources. The underlying design principle is to isolate lexical knowledge from processing as much as possible, so that the processors can use the data directly from the database. The idea behind this design is that all lexical sources can be edited at any time by any member of the team, which is facilitated by a PHP interface specifically developed to this end (GDB). This kind of flexibility would not be possible with the monolithic design typical of proof-of-concept systems.

2.1 Lexical resources

Sentitext relies on three major sources: the individual words dictionary (*words*), the multiword expressions dictionary (*mwords*), and the context rules set (*crules*), which is our implementation of Contextual Valence Shifters. The individual words dictionary currently contains over 9,400 items, all of which are labeled for valence. The acquisition process for this dictionary was inspired by the bootstrapping method recurrently found in the literature (e.g., Riloff and Wiebe, 2003, Gamon and Aue, 2005). Lexical items in both dictionaries in our database were assigned one of the following valences: -2, -1, 0, 1, 2. A more detailed description of these resources can be found in (Moreno-Ortiz et al., 2010).

The most similar sentiment analysis system to ours (Taboada et al., 2011) uses a scale from –5 to 5, which makes sense for a number of graded sets of near synonyms such as those given as examples by the authors (p. 273). In our opinion, however, as more values are allowed, it becomes increasingly difficult to decide on a specific one while maintaining a reasonable degree of objectivity and agreement among different (human) acquirers, especially when there is no obvious graded set of related words, which is very often the case.

There are two ways in which the original valence of a word or phrase can be modified by the immediately surrounding context: the valence can change in degree (intensification or downtoning), or it may be inverted altogether. Negation is the simplest case of valence inversion. The idea of Contextual Valence Shifters (CVS) was first introduced by Polanyi and Zaenen (2006), and implemented for English by Andreevskaia and Bergler (2007) in their CLaC System, and by Taboada et al. (2011) in their Semantic Orientation CALculator (SO-CAL). To our knowledge, apart from Brooke et al.'s (2009) adaptation of the SO-CAL system, to the best of our knowledge, Sentitext is the only sentiment analysis system to implement CVS for Spanish natively. Our context rules account both for changes of degree and inversion, and are stored in a database table which is loaded dynamically at runtime.

2.2 Global Sentiment Value

Sentitext provides results as a number of metrics in the form of an XML file which is then used to generate the reports and graphical representations of the data. The crucial bit of information is the Global Sentiment Value (GSV), a numerical score (on a 0-10 scale) for the sentiment of the input text. Other data include the total number of words, total number of lexical words (i.e., content, non-grammatical words), number of neutral words, etc.

To arrive at the global value, a number of scores are computed beforehand, the most important of which is what we call Affect Intensity, which modulates the GSV to reflect the percentage of sentiment-conveying words the text contains.

Before we explain how this score is obtained, it is worth stressing the fact that we do not count words (whether positive, negative, or neutral), but *text segments* that correspond to lexical units (i.e., meaning units from a lexicological perspective).

[3] The application can be accessed and tested online at http://tecnolengua.uma.es/sentitext

As we mentioned before, items in our dictionaries are marked for valence with values in the range -2 to 2. Intensification context rules can add up to three marks, for maximum score of 5 (negative or positive) for any given segment.

The simplest way of computing a global value for sentiment would be to add negative values on the one hand and positive values on the other, and then establishing it by simple subtraction. However, as others have noted (e.g., Taboada et al. 2011), things are fairly more complicated than that. Our Affect Intensity measure is an attempt to capture the impact that different proportions of sentiment-carrying segments have in a text. We define Affect Intensity simply as the percentage of sentiment-carrying segments. Affect Intensity is not used directly in computing the global value for the text, however, an intermediate step consists of adjusting the upper and lower limits (initially -5 and 5). The Adjusted Limit equals the initial limit unless the Affect Intensity is greater than 25 (i.e., over 25% of the text's lexical items are sentiment-carrying. Obviously, using this figure is arbitrary, and has been arrived at simply by trial and error. The Adjusted Limit is obtained by dividing the Affect Intensity by 5 (since there are 5 possible negative and positive valence values).

A further variable needs some explaining. Our approach to computing the GSV is similar to Polanyi and Zaenen's (2006) original method, in which equal weight is given to positive and negative segments, but it differs in that we place more weight on extreme values. This is motivated by the fact that it is relatively uncommon to come across such values (e.g. "extremely wonderful"), so when they do appear, it is a clear marker of positive sentiment. Other implementations of Contextual Valence Shifters (Taboada et al. 2011) have put more weight only on negative segments when modified by valence shifters (up to 50% more weight), operating under the so-called "positive bias" assumption (Kennedy and Inkpen 2006), i.e., negative words and expressions appear more rarely than positive ones, and therefore have a stronger cognitive impact, which should be reflected in the final sentiment score.

In our implementation, equal weight is placed to positive and negative values. However, we do not simply assign more weight to both extremes of the scale (-5 and 5), we place more weight on each increasingly toward both ends of the scale.

The resulting method for obtaining the Global Sentiment Value for a text is defined as:

$$GSV = \frac{(\sum_{i=1}^{5} 2.5 \cdot i \cdot N_i + \sum_{i=1}^{5} 2.5 \cdot i \cdot P_i) \cdot uB}{5 \cdot (LS - NS)} \quad (1)$$

where N_i is the number of each of the negative valences found, and P_i is the equivalent for positive values. The sum of both sets is then multiplied by the Affect Intensity (uB). LS is the number of lexical segments and NS is the number of neutral ones. Although not expressed in the equation, the number of possible scale points (5) needs to be added to the resulting score, which, as mentioned before, is on a 0-10 scale.

3 Task description

The evaluation experiment described in this paper was performed as conceived for the TASS Workshop on Sentiment Analysis, a satellite event of the SEPLN 2012 Conference. See Villena-Román et al., (2013) for a detailed description of the tasks involved.

4 Analysis of results

Although it might seem obvious, it is worth stressing that lexicon-based systems rely heavily on the availability of a certain number of words on which to apply the weighing operations. As described in section 2.2 above, Sentitext basically computes its GSV index by weighing the number and valences of polarity words and phrases against the number lexical segments found in the text. Although it does include threshold control (the Affect Intensity index discussed in 2.2 above) for varying text lengths, such threshold was designed to be applied to larger texts, considering "short" the average length of a media article or blog entry.

However, Twitter, with its 140 character limit, involves a radically different concept of "short". The average number of lexical segments per tweet, i.e., individual words and identified multiword expressions, that we obtained in our analysis of the test set was 14.1, whereas the average number of polarity-conveying segments was 5.5. This is a very high ratio indeed, implying that social networking sites are commonly used for expressing sentiments and opinions. This is in accord with what many scholars have found when analyzing SNS content (e.g., Siemens,

2011). Sentitext's Affect Intensity, i.e., the control threshold, is established at 25%, which, in our experience, is rarely reached except for extremely short texts with a high emotional load. These data are summarized in Table 1.

	N	%	AVG/tweet
Lexical	857,727	100	14.1
Polarity	337,238	39,32	5.5
Neutral	520,489	60,68	8.6

Table 1: Polarity of text segments

It is therefore not surprising that our analysis of this Twitter test set throws an average Affect Intensity of 19.22, which is extremely high, especially if we bear in mind that 38.4% of the tweets have an Affect Intensity of 0, that is, they are neutral. As for the tweets classification task itself, we show and discuss the results in the following section, where we also offer figures of a more typical evaluation scenario in which texts are categorized as negative, neutral, or positive, i.e., there is no intensification for polarity categories and no distinction between the NEU and NONE categories (both are considered as neutral).

4.1 Three levels + NONE test

Table 2 below summarizes the hit rate for each of the categories, as well as overall.

	N	Hits	H %	Misses	M %
N	15,840	8,848	55.86	6,992	44,14
NEU	1,302	647	49.69	655	50.31
P	22,231	13,284	59.75	8,947	40.25
NONE	21,411	84	0.39	21,327	99.61
Total	60,784	21,327	37.61	37,921	62.39

Table 2: Hit rate for 3L+N test

Results are above average for polarity categories, but not so much for neutral and especially for the NONE category, with just a 0.39% hit ratio. The reason for this is that we decided to classify tweets as belonging to this category exclusively when they were essentially void of content, for example, those that contained just a URL. Clearly this is not what was meant, but we have to say that even after analyzing the correct results, the difference between NEU and NONE is still not clear.

The first conclusion that can be drawn, as far as actual performance is concerned, is that Sentitext has an excessive tendency to assign middle-of-the-scale ratings, both when the correct assignment is negative and positive. Since our tool does not classify, but simply assign a rating on a scale, the actual classification implied deciding on the scale boundaries for each of the categories. Table 3 below shows the boundaries we selected for this test.

Category	GSV Range
P+	GSV>8
P	GSV<=8
NEU	GSV<=5.4
N	GSV<=4
N+	GSV<=2
NONE	No content

Table 3: GSV ranges used for classification

An obvious way in which we could have optimized these ranges and obtained better results would have been contrasting the training set results with ours. This would have also softened the impact caused by the NEU-NONE issue.

Table 4 below offers performance results in terms of the usual metrics for classifiers.

	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.497	0.023	0.043
P	0.598	0.688	0.639
NONE	0.004	1	0.008

Table 4: Evaluation metrics for 3L+N test

As expected, these figures are extremely low for the NEU and NONE categories. The high recall rate for the NONE category is due to the fact that we only classified 84 tweets under this category, all of which were correct. Of course the harmonizing F-measure is very low anyway.

Even the metrics for the negative and positive are relatively low in comparison with previous tests (e.g., Moreno-Ortiz et al., 2011). We believe this may be due to the short length of the texts, and it is something will seek to improve in the future.

4.1.1 Unofficial 3L(-N) test

Since we do not think adding a NONE category serves any practical purpose, we decided to perform the same test removing the NONE category, in order to obtain more useful conclusions as to performance in a real-world scenario, and also to measure the precise impact

that the NONE issue had on the overall performance.

Table 5 below offers the hit rate for each category, which are obviously unchanged from the 3L+N test, but show important differences with the NEU one.

	N	Hits N	H %	Misses	M %
N	15,840	8,848	55.86	6,992	44.14
NEU	22,713	15,709	69.16	7,004	30.84
P	22,231	13,284	59.75	8,947	40.25
Total	60,784	37,841	62.25	22,943	37.75

Table 5: Hit rates for 3L(-N) test

By removing the NONE category, the overall hit rate rises from 37.61% to 62.25%, a difference of 24.64%, as a consequence of dramatic improvement of the hit rate for the NEU category.

In terms of precision and recall, we obtain a proportional improvement for the NEU category, as shown in Table 6 below.

	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.692	0.548	0.611
P	0.598	0.688	0.639

Table 6: Evaluation metrics for 3L(-N) test

4.2 Five levels + NONE test

Again, the influence of the NEU-NONE issue is strong, as can be clearly seen both in the figures in Figure 2, where it is also noticeable the better performance for negative cases than positive ones. Table 7 summarizes the hit rate for each of the categories and overall.

	N	Hits	H %	Misses	M %
N+	4,552	955	20.98	3,597	79.02
N	11,281	5,075	44.99	6,206	55.01
NEU	1,300	647	49.77	653	50.23
P	1,483	760	51.25	723	48.75
P+	20,741	2,643	12.74	18,098	87.26
NONE	21,409	84	0.39	21,325	99.61
Total	60,766	10,164	16.73	5,0602	83.27

Table 7: Hit rate for 5L+N test

The overall hit rate is in this case extremely low, nearly half as in the 3L+N test (16.73%). This is the result of the above-mentioned tendency that Sentitext displays toward middle-of-the-scale values, since most misses (apart from the one caused by NONE) come from classifying N+ as N and P+ as P. It is quite apparent that our current implementation of the GSV index needs a revision in order to make extreme values at both ends of the scale more easily attainable.

Of course, another way in which we could overcome this issue would be simply to use different ranges in our classification scale (see Table 4 above). This quick-and-dirty approach may be worth trying, at least as an interim solution whenever classification is required.

The standard evaluation metrics are provided in Table 8 below.

	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.498	0.023	0.043
P	0.512	0.047	0.085
P+	0.127	0.885	0.223
NONE	0.004	1.000	0.008

Table 8: Evaluation metrics for 5L+N test

The figures clearly reflect the low performance, which falls below 50% in all cases, and especially at both extremes (N+ and P+). It is interesting, though, how recall for the P+ category is particularly high in relation to precision, even proportionally to that of N+. This is because our analyzer only assigned 2,643 cases to this category, which in fact had 21,409 cases, a surprising figure that contrasts with the 4,552 cases for N+. Table 9 below summarizes the official results and provides percentages for each category.

	N	%	3L%
N+	4,552	7.49	26.05
N	11,281	18.56	
NEU	1,300	2.14	
P	1,483	2.44	36.58
P+	20,741	34.13	
NONE	21,409	34.23	34.23
Total	60,766	100	100

Table 9: Official results for each category

Any number of conclusions can be drawn from these numbers.

4.2.1 Unofficial 5L(-N) test

As we did before, we show the hypothetical results that we would obtain if the NONE

category were to be removed. Tables 10 shows the hit rate in this scenario.

	N	Hits	H %	Misses	M %
N+	4,552	955	20.98	3,597	79.02
N	11,281	5,075	44.99	6,206	55.01
NEU	22,709	15,709	69.18	7,000	30.82
P	1,483	760	51.25	723	48.75
P+	20,741	2,643	12.74	18,098	87.26
Total	60,766	25,142	41.38	35,624	58.62

Table 10: Hit rate for 5L(-N)

And finally, precision and recall figures:

	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.692	0.548	0.612
P	0.512	0.047	0.085
P+	0.127	0.885	0.223

Table 11: Evaluation metrics for 5L(-N)

As before, precision and recall are the same for all categories except NEU, which rises significantly in precision, and extremely in recall. Hit rate improves in the same proportion as in the 3L(-N) test.

5 Conclusions

Performing this test has been extremely useful to identify weaknesses in our current implementation of Sentitext's Global Sentiment Value. On the one hand, this test confirms our initial impressions after carrying out some informal tests with Twitter messages, that GSV is strongly affected by the number of lexical units available in the text (or the lack of them, rather). On the other hand, we have also confirmed Sentitext's tendency to assign middle-of-the-scale ratings, or at least avoid extreme values, which is reflected on its poor performance for the N+ and P+ classes, most of which were assigned to the more neutral N and P classes. This happens despite the fact that our GSV calculation places more weight on extreme values. Conversely, we found a relatively high proportion of polarized lexical segments found (high Affect Intensity). This is something that could not be inferred from the results of machine learning classifier. Even with a high proportion of neutral messages, these numbers clearly support the claims of many social media analysts that social networking sites, are used mainly to circulate news and express emotions about them. But our data also indicate that they tend to avoid strong language to convey their opinions, relying more on mild expression, implicature, and shared knowledge.

The third important conclusion is that differentiating between neutral and no polarity may not be the best decision, since it is not clear what the difference is. In fact, after checking the official assignment of these tags to the test set, it seems to us completely random. Therefore, it is very difficult to obtain good results in these two categories. Furthermore, there really is no need whatsoever to make this distinction from a practical perspective.

6 References

Andreevskaia, A., and Bergler, S. (2007). CLaC and CLaC-NB: knowledge-based and corpus-based approaches to sentiment tagging. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 117–120). Prague, Czech Republic: Association for Computational Linguistics.

Atserias, J., Casas, B., Cornelles, E., González, M., Padró, L., and Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation*. Presented at the LREC 2006, Genoa, Italy: ELRA.

Aue, A., and Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. Presented at the Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.

Balahur, A., Kozareva, Z., and Montoyo, A. (2009). Determining the Polarity and Source of Opinions Expressed in Political Debates. *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09 (pp. 468–480). Berlin, Heidelberg: Springer-Verlag.

Boldrini, E., Balahur, A., Martínez-Barco, P., and Montoyo, A. (2009). EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. *Proceedings of The 2009 International Conference on Data Mining*

(pp. 491–497). Presented at the DMIN 2009, Las Vegas, USA: CSREA Press.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. *Proceedings of RANLP 2009, Recent Advances in Natural Language Processing*. Presented at the RANLP 2009, Borovets, Bulgaria.

Cruz, F., Troyano, J. A., Enriquez, F., and Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41), 73–80.

Gamon, M., and Aue, A. (2005). Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms (pp. 57–64). Ann Arbor, Michigan: Association for Computational Linguistics.

Hatzivassiloglou, V., and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174–181). Madrid, Spain: Association for Computational Linguistics.

Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110–125.

Moreno-Ortiz, A., Pineda Castillo, F., and Hidalgo García, R. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento de Lenguaje Natural*, *45*, 31–39.

Moreno-Ortiz, A., Pérez-Hernández, C., and Hidalgo-García, R. (2011). Domain-neutral, Linguistically-motivated Sentiment Analysis: a performance evaluation. *Actas del XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (pp. 847–856).

Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica*, *3*(2), 13–20.

Pang, B., and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 271). Barcelona, Spain: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=1218990anddl=GUIDEandcoll=GUIDEandCFID=80308782andCFTOKEN=73139236

Pang, B., and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL 2005* (pp. 115–124). Presented at the ACL.

Polanyi, L., and Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Dordrecht, The Netherlands: Springer.

Riloff, E., and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03 (pp. 105–112). Stroudsburg, PA, USA: Association for Computational Linguistics.

Siemens, G. (2011, July 30). Losing interest in social media: there is no there there. *Elearnspace*. Retrieved from http://www.elearnspace.org/blog/2011/07/30/losing-interest-in-social-media-there-is-no-there-there/

Taboada, M., Brooks, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 417–424). Presented at the ACL 2002, Philadelphia, USA.

Villena-Román, J., García-Morera, J., Moreno-Garcia, C., Ferrer-Ureña, L., Lana-Serrano, S., González-Cristobal, J. C., Westerski, A., Martínez-Cámara, E., García-Cumbreras, M. A., Martín-Valdivia, M. T., Ureña-López L. A. 2012. TASS-Workshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 103-105 recibido 21-11-12 revisado 15-01-13 aceptado 19-02-13

On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism *

Detección de texto reutilizado y plagio monolingüe y translingüe

Alberto Barrón Cedeño
DSIC Universitat Politècnica de València

TALP Research Center
Universitat Politècnica de Catalunya
albarron@[lsi.upc.edu | gmail.com]

Resumen: Tesis de doctorado en ciencias de la computación (con mención europea del doctorado) escrita por Alberto Barrón Cedeño bajo la supervisión del Dr. Paolo Rosso en la Universitat Politècnica de València. El autor fue examinado en Valencia en julio de 2012 por un jurado compuesto por los siguientes doctores: Paul Clough (*University of Sheffield*), Benno Stein (*Bauhaus-Universität Weimar*), Ricardo Baeza-Yates (*Yahoo! Research*), Fabio Crestani (*Università della Svizzera italiana*) y José Miguel Benedí (Universitat Politècnica de Valencia). La mención europea fue obtenida tras una estancia de 4 meses en la *Information School* de la *University of Sheffield* (Reino Unido) bajo la supervisión del Dr. Paul Clough.
Palabras clave: recuperación de información translingüe, plagio traducido, texto reutilizado, plagio parafrástico, plurilingüismo en Wikipedia

Abstract: Ph.D. thesis (European doctorate mention) in Computer Science written by Alberto Barrón Cedeño under the advice of Dr. Paolo Rosso at the Universitat Politècnica de València. The author was examined in Valencia in July 2012 by a jury composed of the following doctors: Paul Clough (University of Sheffield), Benno Stein (Bauhaus-Universität Weimar), Ricardo Baeza-Yates (Yahoo! Research), Fabio Crestani (Università della Svizzera italiana), and José Miguel Benedí (Universitat Politècnica de Valencia). The European mention was received after a 4 months internship at the Information School of the University of Sheffield (UK) under the advice of Dr. Paul Clough.
Keywords: cross-language information retrieval, re-used text, cross-language plagiarism, paraphrase plagiarism, Wikipedia multilingualism

1. Introduction

Automatic text re-use detection is the task of determining whether a text has been produced by considering another as its source. Plagiarism, the unacknowledged re-use of text, has gained the greatest notoriety. Favoured by the easy access to information through electronic media, plagiarism has raised in recent years, requesting for the attention of experts in text analysis.

Automatic text re-use detection takes advantage of NLP and IR technology to compare thousands of documents —looking for the potential source of a presumably case of re-use. Machine translation technology can be used in order to uncover cases of cross-language re-use. By exploiting such technology, thousands of exhaustive comparisons are possible, also across languages, something impossible to manually achieve.

In this dissertation we pay special attention to three aspects of text re-use:

1. Cross-language text re-use: we propose a cross-language similarity assessment model that represents one of the best

* This PhD research was supported by the National Council of Science and Technology of Mexico (CONACyT) through the 192021/302009 scholarship. The Ministry of Education of Spain supported my internship in the University of Sheffield through the TME2009-00456 grant. The investigation was carried out in the framework of the MICINN project Text-Enterprise 2.0 (TIN2009-13391-C04-03).

ISSN 1135-5948

options when looking for exact translations.

2. Paraphrase text re-use: we investigate what types of paraphrasing are more frequently applied when plagiarising and how they difficult plagiarism detection; something never done before.

3. Mono- and cross-language re-use within and from Wikipedia: the encyclopedia is explored as a multi-authoring framework, where texts are re-used within versions of an article and across languages.

2. Thesis Overview

The dissertation consists of 9 chapters, describing our efforts to approach the main difficulties of automatic text re-use detection. The contents are described following.

Chapters 2 and 3 are an overall introduction of the covered topics. Chapter 2 offers an overview of text re-use, with special emphasis on plagiarism. Our contribution comes in the form of the survey we held in different Mexican universities; aiming to assess how often students plagiarise across languages and their attitudes respect to paraphrase plagiarism (factors never analysed before). Chapter 3 introduces the IR and NLP concepts used through the rest of the thesis.

Chapter 4 describes corpora for (automatic) analysis of text re-use and plagiarism available up to date. Our participation in the construction of three corpora —coderivatives, CL!TR, and to a smaller extent PAN-PC— are cutting edge contributions discussed in this chapter. Evaluation metrics are also discussed: some are well known in IR and related areas, whereas others were recently proposed —and specially designed— for evaluating text re-use detection.

Chapter 5 defines the two main approaches to re-use detection: intrinsic and external. Our contributions to external (monolingual) detection are discussed. Our main contribution is a model for retrieving those related documents to the suspicious one, hence reducing the load when performing the actual plagiarism detection process. Such a problem is often neglected in the plagiarism detection literature, that *assumes* that either the step is not necessary or it is already solved; an absolutely false idea.

Chapter 6 describes our model for cross-language detection (this is one of the least approached problems of re-use detection!): CL-ASA. CL-ASA is compared to state-of-the-art models over different sub-tasks of the detection process. A variety of languages is considered to analyse the strengths and weaknesses of the different models.

Chapter 7 discusses the international competitions we ran during three years: the PAN International Competition on Plagiarism Detection. We also experiment with our detection models on the generated test-beds and discuss the obtained results.

Chapter 8 analyses plagiarism from the point of view of paraphrasing, providing a bridge between the two disciplines: plagiarism detection and paraphrase analysis. Our findings on the use of paraphrasing when plagiarising represent useful insights to take into account when developing the next generation of plagiarism detection systems.

In Chapter 9 we analyse monolingual coderivation among revisions of Wikipedia articles and cross-language text re-use from Wikipedia. Related to the latter issue, we offer a preliminary discussion on the PAN competition we organised at FIRE on cross-language text re-use: PAN Cross-Language !ndian Text Re-Use; where the potentially re-used documents were written in Hindi and the potential source documents were written in English.

3. Thesis Contributions

The main contributions of this research are described below.

Detection of text re-use across languages. We explored a range of cross-language information retrieval techniques. We observed that (*i*) a simple model based on characterising texts by short character n-grams (CL-CNG) was worth considering when dealing with common-alphabet languages (and different alphabets, after transliteration), and particularly if they have some influence (Barrón-Cedeño et al., 2010; Potthast et al., 2011); (*ii*) the model cross-language explicit semantic analysis (CL-ESA), based on large comparable corpora such as Wikipedia, performs well when looking for related documents across languages (Potthast et al., 2011). We proposed a model —cross-language alignment-based similarity analysis, CL-ASA—, based on translation probabilities and length distributions between texts (Barrón-Cedeño et al., 2008; Pinto et al., 2009). Our empirical results showed that

CL-ASA is competitive when looking for reused texts, regardless if they were manually or automatically translated. CL-ASA performs better than CL-ESA and CL-CNG, identified as two of the most appealing models for cross-language similarity assessment, when dealing with translations at document and fragment level (Potthast et al., 2011).

Creation of standard collections of documents for the study and development of plagiarism detection. We helped in the creation of two "sister" corpora with simulated cases of re-use and plagiarism. The PAN-PC series look at composing a realistic IR challenge: it includes thousands of documents, with thousands of plagiarism cases (both manually and automatically generated) (Potthast et al., 2010). The CL!TR corpus looks at composing a realistic cross-language challenge: it contains a few thousand documents, with hundreds of re-use cases (manually generated across distant languages) (Barrón-Cedeño et al., 2011). These corpora (particularly the PAN-PC series) have become a reference in the development of models for plagiarism detection, filling an important gap.[1]

Analysis of paraphrase plagiarism and its detection. The vast majority of models for text re-use detection are designed to uncover "cut and paste" cases, as they consider surface information only. These models are unsuccessful when facing paraphrase plagiarism. For the first time, we analysed the paraphrase phenomena applied when text is plagiarised (Barrón-Cedeño et al., 2013 (to appear)). Our seminal study showed that lexical substitutions are the paraphrase mechanisms used the most. Moreover, the paraphrasing tends to be used to generate a simplified version of the re-used text. A model intended to succeed in detecting paraphrase re-use requires robust text pre-processing and characterisations: the expansion (or contraction) of related vocabulary, the normalisation of formatting and word forms, and the inclusion of mechanisms that model the expected length of a re-used fragment given its source.

[1]The PAN-PC corpora, created in the framework of the PAN International Competition on Plagiarism Detection, are available at `http://www.uni-weimar.de/cms/medien/webis/research/corpora.html`. The CL!TR corpus, created in the framework of the PAN Cross-Language !ndian Text Reuse challenge, is available at `http://memex2.dsic.upv.es/workshops/2011/clitr/`

References

Barrón-Cedeño, Alberto, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In Huang and Jurafsky (Huang and Jurafsky, 2010).

Barrón-Cedeño, Alberto, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, and Mark Stevenson. 2011. PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In FIRE, editor, *FIRE 2011 Working Notes. Third Workshop of the Forum for Information Retrieval Evaluation.*

Barrón-Cedeño, Alberto, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, volume 377, pages 9–13, Patras, Greece. CEUR-WS.org. http://ceur-ws.org/Vol-377.

Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013 (to appear). Plagiarism meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics.*

Huang, Chu-Ren and Dan Jurafsky, editors. 2010. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August. COLING 2010 Organizing Committee.

Pinto, David, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. 2009. A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60.

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, 45(1):1–18.

Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 997–1005.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 107-109 recibido 27-11-12 revisado 01-02-13 aceptado 19-02-13

Linguistic-based Patterns for Figurative Language Processing: The Case of Humor Recognition and Irony Detection*

Patrones lingüísticos para el procesamiento del lenguaje figurado: el caso de reconocimiento de humor y detección de ironía

Antonio Reyes Pérez
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de Vera s/n, 46022. Valencia, Spain

Instituto Superior de Intérpretes y Traductores
Laboratorio de Tecnologías Lingüísticas
Río Rhin 40, 06500. Mexico City, Mexico
antonioreyes@isit.edu.mx

Resumen: Tesis doctoral en Informática realizada por Antonio Reyes Pérez y dirigida por el doctor Paolo Rosso (Universitat Politècnica de València). La lectura de la tesis fue realizada en la ciudad de Valencia (España) el día 2 de julio de 2012 ante un tribunal compuesto por los doctores: Antónia Martí Antonín (Universitat de Barcelona), Walter Daelemans (University of Antwerp), Richard Anthony (Tony) Veale (University College Dublin), Carlo Strapparava (Fondazione Bruno Kessler FBK -IRST) y José Antonio Troyano Jiménez (Universidad de Sevilla). La calificación obtenida fue de *Apto* con la mención *Cum Laude*.
Palabras clave: Reconocimiento de humor, detección de ironía, lenguaje figurado.

Abstract: Ph. D. thesis in Computer Science written by Antonio Reyes Pérez under the supervision of Dr. Paolo Rosso (Universitat Politècnica de València). The thesis defense was done in Valencia (Spain) on July 2nd, 2012. The doctoral committee was integrated by the following doctors: Antónia Martí Antonín (University of Barcelona), Walter Daelemans (University of Antwerp), Richard Anthony (Tony) Veale (University College Dublin), Carlo Strapparava (Fondazione Bruno Kessler FBK -IRST), and José Antonio Troyano Jiménez (University of Sevilla). The obtained grade was *Cum Laude*.
Keywords: Humor recognition, irony detection, figurative language.

1. Introduction

This investigation aimed to show how two specific domains of figurative language: humor and irony, could be automatically handled by means of considering linguistic-based patterns. We were especially focused on discussing how underlying knowledge, which relies on shallow and deep linguistic layers, could represent relevant information to automatically identify figurative uses of language. In particular, and contrary to most researches on figurative language, we focused on identifying figurative uses of language in social media. This means that our findings do not rely on analyzing prototypical jokes or literary examples of irony; rather, we tried to find patterns in social media texts of informal register whose intrinsic characteristics are quite different to the characteristics described in the specialized literature. For instance, a joke which exploits phonetic devices to produce a funny effect, or a tweet in which irony is self-contained in the situation. In this context, we proposed a set of features which work together as a system: no single feature was particularly humorous or ironic, but all together provided a useful linguistic inventory for detecting humor and irony at textual level.

2. Objective

Figurative language is in some way inherent to discourse, whatever the type of text. In this respect, the problem of automatically detecting figurative language cuts through

* Thesis funded by the National Council for Science and Technology (CONACyT - Mexico); as well as partially supported by the Text-Enterprise 2.0 project (TIN2009-13391-C04-03).

ISSN 1135-5948

every aspect of language, from pronunciation to lexical choice, syntactic structure, semantics and conceptualization. As such, it is unrealistic to seek a computational silver bullet for figurative language, and a general solution will not be found in any single technique or algorithm. Rather, we tried to identify specific aspects and forms of figurative language that were susceptible to be computationally analyzed, and from these individual treatments attempt to synthesize a gradually broader solution. In this context, our objective was to deeply analyze two figurative devices: humor and irony, in order to detect textual patterns to be applied in their automatic processing, especially, in their automatic identification at textual level. In order to achieve such objective, several conceptual and practical issues were addressed throughout the thesis:

I. Literal and figurative language are windows to cognitive processes that are linguistically verbalized: the meaning cannot be derived only from lexicon.

II. Specialized literature defines humor and irony in fine-grained terms. Such granularity cannot be directly mapped from theory to praxis: need of representing the core of both devices the less abstract as possible in order to describe deeper and more general attributes of both phenomena; rather than only *ad hoc* cases.

III. Overlapping is quite common in figurative language. Indeed, irony is a common mechanism to produce a humorous effect, and *vice versa*: there are not formal linguistic boundaries to accurately separate both figurative devices.

IV. Humor and irony are typical devices in which both literal and non-literal meaning might be simultaneously active: there are not linguistic marks to denote where the figurative meaning is starting.

V. There are no available data to assess hypothesis or models: need of building objective corpus.

3. Thesis Overview

The thesis is conceptually organized as follows: In Chapter 2 we described the linguistic background as well as the theoretical issues regarding literal language and figurative language. We emphasized the importance of considering language as a dynamic system, rather than a static one. Both humor and irony were conceptually described and discussed in detail. In Chapter 3 we introduced the related work concerning figurative language processing. First, the framework in which the thesis is developed was described. Then, the challenges that any computational treatment of figurative language faces was outlined.

In Chapter 4 we described, both conceptually and pragmatically, our humor recognition model. Hypotheses, patterns, experiments, and results were presented. Moreover, evaluation data sets were introduced. Finally, we discussed model's implications. In Chapter 5, in turn, we presented our irony detection model. First, operational bases, as well as aims, were outlined. Then, experiments and results were explained. Like in the previous chapter, all the evaluation data sets were introduced. Lastly, results and further implications were discussed.

In Chapter 6 we described how both models were assessed in terms of their applicability in tasks related to information retrieval, sentiment analysis, and trend discovery. Such evaluations were intended to represent real scenarios concerning figurative language processing beyond the data sets employed in Chapters 4 and 5. Finally, in Chapter 7 we outlined the main conclusions of the thesis, as well as its contributions and lines for future work.

4. Contributions

In this thesis we approached two tasks in which the automatic processing of figurative language has been involved: humor recognition and irony detection. Each task was undertaken independently by means of a linguistic pattern representation. In this respect, two models of figurative language were proposed:

- HRM (Humor Recognition Model);
- IDM (Irony Detection Model).

Both models go beyond surface elements to extract different types of patterns from a text: from lexicon to pragmatics. Since our target was focused on representing figurative language concerning social media texts, each

model was evaluated by considering non-prototypical texts that are laden with social meaning. Such texts were automatically collected by chiefly taking advantage of user-generated tags. The data sets are freely available for research purposes.

Two goals were highlighted while evaluating the models: representativeness and relevance. The former was intended to consider the appropriateness or representativeness of different patterns to humor recognition and irony detection, respectively; whereas the latter was focused on considering the empirical performance of each model on a text classification task.

Below are summarized our major findings:

I. By representing humor and irony in terms of their conceptual use rather than only of their theoretical description, our models seem to efficiently capture the core of the most salient attributes of each figurative device.

II. Our figurative language representation is given by analyzing the linguistic system as an integral structure which depends on grammatical rules as well as on cognitive, experiential, and social contexts, which altogether, represent the meaning of what is communicated.

III. We provided a methodology to automatically identify figurative uses of language in order to foster figurative language processing beyond the tasks described in the document.

IV. By analyzing non prototypical examples of humor and irony, we provided a pair of models which were supported by general patterns used by people to effectively communicate figurative intents.

V. With our approach (that is focused on taking advantage of user-generated tags), we have reduced the constraints facing corpus-based research. For instance, the subjectivity of determining figurativity at textual level is reduced by collecting examples that are intentionally labeled with a descriptor (user-generated tag) whose goal is to focus people's posts on particular topics.

VI. By making freely available our data sets we are collaborating to the spread of researches related to figurative language, as well as palliating the lack of resources for figurative language processing.

VII. Figurative language is a widespread phenomenon in web content. In this respect, the empirical insights described in the document showed how our models provide fine-grained knowledge concerning their applicability in tasks as diverse as information retrieval, sentiment analysis, trend discovery, or online reputation.

5. Publications (Impact Factor)

Reyes A., P. Rosso, D. Buscaldi. 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74:1–12. DOI: 10.1016/j.datak.2012.02.005.

Reyes A., P. Rosso. 2012. Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. *Journal on Decision Support Systems*, 53(4):754–760. DOI: 10.1016/j.dss.2012.05.027.

Reyes A., P. Rosso, T. Veale. 2012. A Multidimensional Approach For Detecting Irony in Twitter. *Language Resources and Evaluation*, 47(1). DOI: 10.1007/s10579-012-9196-x.

Reyes A., P. Rosso. Forthcoming. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation *Knowledge and Information Systems.*

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 111-114 recibido 29-11-12 revisado 08-02-13 aceptado 19-02-13

Nominalizaciones deverbales: denotación y estructura argumental

Deverbal nominalizations: denotation and argument structure

Aina Peris
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
aina.peris@ub.edu

Resumen: Tesis doctoral en Lingüística Computacional realizada por Aina Peris en la Universitat de Barcelona (UB) bajo la dirección de la Dra. Mariona Taulé (UB) y el Dr. Horacio Rodríguez (Universitat Politècnica de Catalunya). El acto de defensa de la tesis tuvo lugar el viernes 11 de mayo de 2012 ante el tribunal formado por los doctores Piek Vossen (Vrije Universiteit of Amsterdam), Lidia Moreno (Universitat Politècnica de Valencia) y Mª Antònia Martí (UB). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad con mención europea.
Palabras clave: Nominalizaciones deverbales, desambigüación automática, etiquetador de roles semánticos

Abstract: Ph.D. Thesis in Computational Linguistics, written by Aina Peris at the University of Barcelona (UB), under the supervision of Dr. Mariona Taulé (UB) and Dr. Horacio Rodríguez (Technical University of Catalonia). The author was examined on friday, 11$^{\text{th}}$ of May 2011, by a commitee formed by the doctors Piek Vossen (Vrije Universiteit of Amsterdam), Lidia Moreno (Technical University of Valencia) and Mª Antònia Martí (UB). The grade obtained was Excellent *Cum Laude* unanimously (with European mention).
Keywords: Deverbal nominalizations, automatic disambiguation, semantic role labeling

1 Introducción

Las nominalizaciones deverbales del español son construcciones lingüísticas que se caracterizan por presentar propiedades propias de los sustantivos pero al mismo tiempo por heredar la estructura argumental de los verbos de los que derivan. Esta dualidad les confiere un notable interés lingüístico porque pueden denotar tanto un estado o el resultado de la acción denotada por el verbo base correspondiente, y también pueden denotar la misma acción o evento que expresa el verbo base, y por tanto, ser paráfrasis de cláusulas oracionales. Por otra parte, son sustantivos que tienen capacidad argumental, es decir, seleccionan argumentos y, en este sentido, es relevante observar los patrones de realización sintáctico-semántica de los argumentos de las nominalizaciones, ya que suponen una manera alternativa de expresar el significado contenido en una oración.

Por lo tanto, dado que las nominalizaciones deverbales pueden expresar el mismo contenido semántico que los predicados verbales y que son construcciones bastante frecuentes en el lenguaje escrito, nos parecía necesario estudiarlas desde el punto de vista de la Lingüística Computacional, contribuyendo, así, a los trabajos que hasta ahora han ido un paso más allá de los verbos en la representación semántica de los textos. Sin embargo, estos trabajos se centran básicamente en las nominalizaciones deverbales del inglés, por lo que también creímos necesario emprender este estudio en español. Veamos ejemplos del tipo de fenómeno con el que tratamos:

(1) [La **construcción** hotelera] ha sido derribada tras la sentencia judicial que así lo ordenaba.

(2) La reflexión fue necesaria para [la posterior **construcción** de la democracia].

En el ejemplo 1 la nominalización *construcción* hace referencia al edificio resultado de las acción del verbo mientras que en el ejemplo 2 se refiere a la acción o evento de

ISSN 1135-5948

construir. En ambos ejemplos, además, las nominalizaciones tienen complementos del nombre (CN) que indican el objeto construido. Por lo tanto, ambos CNs pueden ser asociados a la posición argumental de paciente (arg1-pat).

Además del intrínseco valor lingüístico que tiene el estudio de estas construcciones, también desde un punto de vista del Procesamiento del Lenguaje Natural (PLN) resulta interesante disponer de herramientas y recursos que traten y representen las nominalizaciones deverbales del español, tanto en lo que se refiere a la denotación como a la estructura argumental. Tareas como la resolución de la correferencia o la detección de paráfrasis pueden beneficiarse de una herramienta o un recurso que trate el tipo denotativo de las nominalizaciones, y aplicaciones de extracción de información o sistemas de etiquetado semántico, pueden aprovechar herramientas y recursos que representen la estructura argumental de las nominalizaciones.

2 Organización de la tesis

Esta tesis se estructura en cuatro partes: los antecedentes en el estudio de las nominalizaciones deverbales, la estructura argumental, la denotación y los recursos derivados que las representan. La primera parte introduce el concepto de nominalización deverbal, la importancia de su estudio (Capítulo 1) y ofrece una panorámica de los trabajos realizados, tanto desde el punto de vista lingüístico como computacional (Capítulo 2). La segunda parte centra su atención en la estructura argumental de las nominalizaciones deverbales, tanto el estudio empírico realizado sobre este aspecto (Capítulo 3) como el sistema automático desarrollado (RHN) para la anotación de dicha información en el corpus (Capítulo 4). La tercera parte trata la distinción denotativa entre evento y resultado, tanto el estudio empírico realizado sobre este aspecto (Capítulo 5), como el sistema de clasificación automático desarrollado (ADN) para la anotación de dicha información en el corpus (Capítulo 6) y los experimentos desarrollados con este clasificador (Capítulo 7). En la cuarta parte se describen los recursos lingüísticos derivados de esta investigación, el corpus AnCora-Es enriquecido con la anotación de las nominalizaciones deverbales (Capítulo 8) y el léxico derivado AnCora-Nom (Capítulo 9). Finalmente, en el Capítulo 10 se recogen las conclusiones globales de este trabajo, las aportaciones del mismo y las líneas de trabajo futuro.

3 Contribuciones

Las contribuciones de esta tesis se resumen a continuación:

- Conjunto de criterios lingüísticos que permiten establecer una distinción entre nominalizaciones eventivas y nominalizaciones resultativas del español. Estos criterios se han obtenido a partir del estudio empírico sobre un subconjunto de 100.000 palabras del corpus AnCora-Es, que nos permitió establecer qué criterios de la bibliografía eran válidos para el español y detectar también una serie de criterios nuevos que ayudan a distinguir entre estas dos lecturas denotativas.
- Estudio lingüístico de la estructura argumental de las nominalizaciones deverbales, es decir, de los distintos patrones de realización sintáctica de los argumentos de estos predicados. A partir de las observaciones iniciales del estudio empírico y su implementación en las reglas de proyección de RHN, hemos obtenido nuevas e interesantes observaciones lingüísticas.
- Construcción del ADN-Classifier, un sistema de clasificación automática de nominalizaciones deverbales según su denotación.
- Implementación de RHN, conjunto de reglas heurísticas que tienen en cuenta la información del léxico AnCora-Verb y a partir de las cuales se ha anotado automáticamente la estructura argumental de las nominalizaciones deverbales del corpus AnCora-Es.
- Enriquecimiento del corpus AnCora-Es con la validación manual de los procesos automáticos de anotación (denotación y estructura argumental) de las nominalizaciones deverbales.
- Creación de AnCora-Nom, un léxico de 1.655 nominalizaciones deverbales en español.

Estas contribuciones se clasifican en tres grandes grupos que detallamos en las siguientes subsecciones: 1) caracterización lingüística de las nominalizaciones deverbales (denotación y estructura argumental); 2) herra-

mientas computacionales para tratar estos dos aspectos de las nominalizaciones deverbales automáticamente, y 3) creación de recursos lingüísticos que representan estas construcciones lingüísticas.

3.1 Caracterización Lingüística

En relación a la distinción denotativa entre evento y resultado de las nominalizaciones deverbales, se han definido una serie de criterios que permiten identificar una de las dos lecturas (Peris y Taulé, 2009). Se analizó si los criterios establecidos en la bibliografía para el inglés eran válidos para el español. Entre los criterios evaluados, los más relevantes para el español son: 1) la clase semántica del verbo del que deriva la nominalización; 2) su capacidad de pluralización; 3) los tipos de determinantes; 4) la preposición que introduce al complemento agentivo; y 5) la presencia obligatoria de un argumento interno (arg1). Estos rasgos se han representado como atributos en las entradas léxicas nominales del léxico AnCora-Nom. Además, el estudio lingüístico llevado a cabo nos permitió encontrar criterios nuevos para la identificación, especialmente, de las nominalizaciones eventivas (puesto que con los criterios de la bibliografía no eran todas identificables): los selectores y el criterio de la paráfrasis. Los selectores pueden ser de dos tipos: (i) selectores externos, elementos que desde fuera del SN indican la denotación de la nominalización (la preposición *durante* por ejemplo); y (ii) selectores internos, prefijos de la nominalización que indican un tipo concreto de denotación (el prefijo *re-* reiterativo se aplica a acciones, por lo tanto las nominalizaciones que lo emplean son eventivas). En cuanto al criterio de la paráfrasis, si un SN cuyo núcleo es una nominalización y puede parafrasearse por una oración con el verbo base, se considera que es una nominalización eventiva.

Respecto a la estructura argumental de las nominalizaciones deverbales, se realizó un estudio lingüístico basado en corpus que permitió definir una serie de patrones de realización sintáctico-semántica que luego se implementaron en la herramienta de etiquetado semántico RHN. A partir del análisis de errores de esta herramienta, hemos podido establecer algunas características de la estructura argumental de las nominalizaciones deverbales. En primer lugar, la hipótesis de trabajo inicial de que las nominalizaciones deverbales heredan la estructura argumental del verbo base correspondiente se confirma ya que RHN consigue un F1 del 77 % y se basa principalmente en la información contenida en el léxico AnCora-Verb. En segundo lugar, se muestra que el orden de los constituyentes de los SNs de núcleo deverbal es más libre que el de los complementos verbales, y que hasta cierto punto depende del contexto. En tercer lugar, cabe destacar que los argumentos de las nominalizaciones están marcados por un alto grado de opcionalidad. Esto afecta especialmente al arg0, que no aparece realizado en numerosas ocasiones. Finalmente, detallamos las características argumentales de los constituyentes que pueden ser complementos de las nominalizaciones deverbales: los SAs no relacionales, los Sadv y las oraciones subordinadas no son argumentos en un SN de núcleo deverbal. Respecto a los SNs complementos de nominalizaciones deverbales, se puede establecer que aquellos anotados como una entidad con nombre locativa o temporal reciben la etiqueta de adjunto locativo (argM-loc) o temporal (argM-tmp. Respecto a los SPs, aquellos introducidos por una preposición especifica como *durante, tras, para* etc., se corrobora que dichas preposiciones apuntan a una determinada etiqueta argumental. También se ha comprobado que las preposiciones regidas de los complementos de régimen verbal no siempre se mantienen en el dominio nominal. En cuanto a los SAs relacionales, encontramos un 45 % que no eran argumentales. Parece confirmarse que que los adjetivos relaciones están sometidos al fenómeno de la co-ocurrencia léxica, es decir, que se anotan como argumentales o no argumentales dependiendo del nombre al que complementen. Los determinantes posesivos, por su parte, se interpretan mayoritariamente como el argumento correspondiente al sujeto verbal.

3.2 Sistemas automáticos

A continuación describimos las dos herramientas computacionales desarrolladas en esta tesis: el sistema RHN y el clasificador ADN.

El sistema de RHN está formado por 107 reglas heurísticas, cuyo objetivo es ligar un constituyente del SN del núcleo deverbal con un argumento y papel temático usando el léxico AnCora-Verb, el corpus AnCora-Es y una lista predefinida de adjetivos relacionales. Estas reglas se organizan en un forma-

to de lista de decisión y se aplican a un SN constituido por una nominalización (N) y un contexto que puede ser de uno, dos o tres constituyentes. Cada regla satisface una condición, una combinación lógica de predicados sobre N o sobre el contexto, y así, se asigna una etiqueta semántica. Hay dos tipos de reglas: (i) catorce reglas generales basadas en la información lingüística de AnCora-Es, y (ii) noventa y tres reglas específicas que también tienen en cuenta la información contenida en el léxico AnCora-Verb. RHN logra un 77 % de F1 (Peris y Taulé, 2011b).

El clasificador ADN clasifica automáticamente las nominalizaciones deverbales del español según su denotación sea de tipo eventivo, resultativo o subespecificado, o formen parte en construcciones lexicalizadas. Se desarrollaron una serie de experimentos para poner a prueba los diferentes modelos de clasificación de ADN y en diferentes escenarios y se han obtenido buenos resultados. Los modelos basados en rasgos del léxico AnCora-Nom superan a los modelos basados en rasgos del corpus. De la misma manera que los modelos que trabajan a nivel de sentido superan a los que trabajan a nivel de lema. ADN logra una mayor precisión en la detección de nominalizaciones resultativas que eventivas.(Peris, Taulé, y Rodríguez, 2009; Peris et al., 2010; Peris, Taulé, y Rodríguez, 2012)

3.3 Recursos Léxicos

Esta tesis ha dado lugar a dos nuevos recursos: se ha enriquecido la anotación del corpus AnCora-Es (Peris, Taulé, y Rodríguez, 2010) con la anotación de 23.431 ocurrencias de nominalizaciones deverbales con su denotación y su estructura argumental y se ha creado el léxico AnCora-Nom (Peris y Taulé, 2011a), con 1.655 entradas léxicas de nominalizaciones deverbales en español.

El enriquecimiento del corpus AnCora-Es se ha llevado a cabo en dos etapas: 1) se realizaron dos procesos automáticos de manera independiente, uno para la anotación de la denotación (con ADN) y otra para la estructura argumental (con RHN) y 2) se validaron manualmente estos dos tipos de informaciones. El corpus AnCora-Es es el único corpus del español anotado con este tipo de información.

El léxico AnCora-Nom, por su parte, fue creado automáticamente a partir de la información contenida en el corpus AnCora-Es. Incluye todos los lemas de las nominalizaciones del corpus con sus denotaciones y sus posibilidades de combinatoria de la estructura argumental.

Bibliografía

Peris, Aina y Mariona Taulé. 2009. Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. En *Proceedings of the 1st International Conference on Corpus Linguistics*, páginas 596–611, Murcia, España.

Peris, Aina y Mariona Taulé. 2011a. AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural.*, 46:11–19.

Peris, Aina y Mariona Taulé. 2011b. Annotating the argument structure of deverbal nominalizations in Spanish. doi: 10.1007/s10579-011-9172-x. *Language Resources and Evaluation.*

Peris, Aina, Mariona Taulé, Gemma Boleda, y Horacio Rodríguez. 2010. ADN-Classifier: Automatically Assigning Denotation Types to Nominalizations. En *Proceedings of the Language Resources and Evaluation Conference*, páginas 1422–1428, Valleta, Malta.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2009. Hacia un sistema de clasificación automática de sustantivos deverbales. *Procesamiento del Lenguaje Natural.*, 43:23–31.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2010. Semantic Annotation of Deverbal Nominalizations in the Spanish AnCora Corpus. En *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, páginas 187–198, Tartu, Estonia.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2012. Empirical methods for the study of denotation in nominalizations in Spanish. *Computational Linguistics*, 38(4):827–865.

Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 115-118 recibido 01-12-12 revisado 21-01-13 aceptado 19-02-13

Multilingual Acquisition of Large Scale Knowledge Resources

Adquisición multilingüe de bases de conocimiento de gran escala

Montse Cuadros
Vicomtech-IK4
Mikeltegi Paselekua, 57
2009 Donostia-San Sebastián
mcuadros@vicomtech.org

Resumen: Tesis doctoral en Informática realizada por Montse Cuadros y dirigida por Dr. Lluís Padró y Dr. German Rigau. La defensa de la tesis fue en la facultad de Informática de la Universitat Politècnica de Catalunya el día 22 de noviembre de 2011. El tribunal estuvo formado por Dr. Horacio Rodríguez (Universitat Politècnica de Catalunya), Prof. Dr. Piek Vossen (Vrije Universiteit Amsterdam), Dra. Arantza Díaz de Ilarraza (Universidad del Pais Vasco), Dra. Irene Castellón (Universitat de Barcelona) y Dr. Roberto Navigli (Sapienza University of Rome), que le otorgaron la nota de Sobresaliente Cum Laude.
Palabras clave: Adquisición de conocimiento, adquisición de léxico, evaluación de recursos, WordNet, desambiguación de acepciones

Abstract: Ph. D. thesis in Computer Science written by Montse Cuadros under the supervision of Dr. Lluís Padró and Dr. German Rigau. The thesis defense was done on 22th November 2011 at the Computer Science Faculty of the Universitat Politècnica de Catalunya. The Doctoral Examination Committee was composed by Dr. Horacio Rodríguez (Universitat Politecnica de Catalunya), Prof. Dr. Piek Vossen (Vrije Universiteit Amsterdam), Dra. Arantza Diaz de Ilarraza (Universidad del Pais Vasco), Dra. Irene Castellón (Universitat de Barcelona) and Dr. Roberto Navigli (Sapienza University of Rome). The thesis was graded Cum Laude.
Keywords: Knowledge acquisition, lexical acquisition, resource evaluation, WordNet, word sense disambiguation

1 Introduction

The use of large-scale semantic resources, such as WordNet, has become a usual, often necessary, practice for most current NLP systems. Princeton WordNet (WN) is by far the most widely-used semantic resource in NLP.

However, even manually, the construction of large-scale semantic repositories for broad-coverage NLP is not a trivial task. It is quite difficult to acquire and consistently integrate large amounts of knowledge into an existing resource. The construction of large and rich knowledge bases for broad-coverage semantic processing takes a great deal of expensive manual effort and incurs large development costs. It involves large research groups for long periods of development. For instance, hundreds of person-years have been invested in the development of wordnets for various languages. In the case of the English WordNet, in more than ten years of manual construction (from 1995 to 2006, that is, from version 1.5 to 3.0), WordNet grew from 103,445 to 235,402 semantic relations[1], which represents a growth of around one thousand new relations per month. However, in 2008, the Princeton group released a new resource with 458,825 word forms from the WordNet definitions manually linked to its appropriate WordNet sense[2].

Furthermore, the relevant knowledge changes across domains and cultures and it has to be steadily kept up to date. New knowledge emerges day by day everywhere and has to be combined with the existing knowledge. For these reasons, knowledge acquisition is still a highly active area of research since the existing knowledge repositories do not

[1] Symmetric relations are counted only once.
[2] http://wordnet.princeton.edu/glosstag.shtml

ISSN 1135-5948

seem to be rich enough to support advanced concept-based NLP applications directly. It seems that such applications require more detailed general-purpose (and also domain-specific) semantic knowledge, which have to be built by automatic means to keep development cost and time inside affordable limits. Obviously, this fact has severely hampered the state-of-the-art of advanced NLP applications.

Thus, the automatic acquisition of the necessary knowledge from available resources, such as naturally occurring text, is one of the most challenging tasks in NLP since it requires some *knowledge understanding* capabilities, which is our final goal. This vicious circle is known as the *acquisition bottleneck*. The intrinsic cycling nature of the problem also suggests a cycling approach for solving it, with incremental iterations of *acquisition–identification* stages. Ideally, the process would start with a *minimal* knowledge base and the relevant resources containing the *implicit* knowledge to be acquired. Then, the automatic acquisition process might produce new content that should be *identified* with respect to the existing knowledge base. This identification process is necessary in order to facilitate the integration of the new knowledge into the existing one, to form a comprehensive and computationally useful knowledge base. Arguably, although these sub-tasks are undeniably difficult, combining them might simplify both.

Figure 1 shows the senses of *party* in WordNet 3.0[3]. From left to right the figure shows the senses, the total number of explicit semantic relations encoded for each synset, the new semantic relations gathered from the semantically tagged WordNet definitions[4] and the gloss. Consider the subtle distinctions among some of them. The first three senses are groups of people and the fourth refers to an entertaining event. Obviously, these senses are defining different aspects of related concepts. This is a major drawback when trying to acquire specific knowledge for each sense.

Hopefully, the semantic relations encoded for each sense can help its proper characterization. For instance, Figure 2 shows some of

Sense	#rel.	#gloss	Gloss
party^1_n	36	114	an organization to gain political power
party^2_n	18	27	a group of people gathered together for pleasure
party^3_n	9	41	a band of people associated temporarily in some activity
party^4_n	13	38	an occasion on which people can assemble for social interaction and entertainment
party^5_n	3	87	a person involved in legal proceedings:

Figura 1: Number of relations for party$_n$ in WordNet 3.0

the related concepts encoded in WordNet[5]. Additionally, this table also presents some of the relations captured by KnowNet (KN), a very large lexical knowledge base which has been derived during our research.

Sense	relation	Sense
party^1_n	hypernym	organization^1_n, organisation^1_n
	member-holonym	political_system^1_n, form_of_government^1_n
	hyponym	American_Labour_Party^1_n
	rgloss	machine^5_n, political_machine^1_n
	related-to	**election**^1_n, **political**^1_a,**vote**^1_v, **elect**^1_v
party^2_n	hypernym	social_gathering^1_n, social_affair^1_n
	hyponym	shindig^1_n, shindy^1_n
	hyponym	dinner^1_n, dinner_party^1_n
	hyponym	wedding^3_n, wedding_party^1_n
	related-form	party^1_v
	rgloss	carouse^1_n carousal^1_n bender^2_n toot^2_n, booze-up^1_n
	related-to	**invitation**^1_n,**ceremonial**^1_a,**cocktail**^1_n, **farewell**^2_n
party^3_n	hypernym	set^5_n, circle^2_n, band^1_n, lot^1_n
	hyponym	rescue_party^1_n
	rgloss	fairly^2_r, fair^2_r evenhandedly^1_r
	related-to	**camp**^4_n,**landing**^2_n, **stretcher**^3_n, **Olympiad**^2_n
party^4_n	hypernym	affair^3_n, occasion^2_n, social_occasion^1_n, function^6_n, social_function^1_n
	hyponym	birthday_party1^1_n
	hyponym	cocktail_party1^1_n
	related-form	party^1_v
	rgloss	party_game^1_n
	related-to	**nuptials**^1_n,**prom**^1_n,**reception**^2_n, **gift**^1_n
party^5_n	hypernym	person^1_n, individual^1_n, someone^1_n, somebody^2_n, mortal^1_n, soul^2_n
	hyponym	assignee^1_n
	domain	law^1_n, jurisprudence^2_n
	rgloss	submission^5_n
	related-to	**accountancy**^1_n, **appearance**^3_n, **attendance**^1_n, **court**^1_n

Figura 2: Some relations for party$_n$ in WordNet 3.0 and KnowNet(in bold)

1.1 Research goals

The main goal of the research presented in this thesis is to devise new methods and tools

[3] word$^{num}_{pos}$, where pos is the part-of-speech (n for nouns, v for verbs, a for adjectives and r for adverbs)

[4] That is, the number of glosses that include that particular sense annotated in its definition

[5] *rgloss* stands for reverse gloss. That is, the corresponding sense of party appears in its gloss. These relations are gathered from the manually sense-disambiguated glosses

for creating automatically new semantic relations between WordNet senses. That is, to accurately increase by automatic means the knowledge represented in WordNet.

In particular, our research requires the construction of new methods and tools for:

1. Acquiring relevant words from general or domain corpora for an specific WordNet word sense.
2. Identifying the *implicit* word senses of the acquired relevant words with respect to an *existing* knowledge base (in particular, WordNet).
3. Evaluating empirically the quality of the resulting *new* semantic relations in a controlled multilingual evaluation framework.

2 Thesis overview

The thesis is organised in seven chapters:

- **Chapter 1: Introduction**
  This chapter presents an overview of the thesis. It revises the motivation and presents the main contributions of the thesis to the state-of-the-art.
- **Chapter 2: State of the Art**
  This chapter reviews the state of the art. It revises the use of wide-coverage *semantic resources* in different NLP tasks. Furthermore, it presents the main methodologies, approaches and techniques used for *building large-scale knowledge resources* in general, manually and automatically. Finally, it overviews the existing *evaluation frameworks* used in the research field to assess the quality of the acquired knowledge.
- **Chapter 3: Knowledge Acquisition Method**
  This chapter describes the knowledge acquisition architecture developed in this research.
- **Chapter 4: Acquisition of topic signatures**
  This chapter reviews the different methods applied to acquire automatically topic signatures as well as the methodology for evaluating their quality.
- **Chapter 5: KnowNet**
  This chapter depicts the KnowNet building process and its grounding Word Sense Disambiguation algorithm, used to obtain word-sense relations from topic signatures acquired from general corpora.
- **Chapter 6: deepKnownet**
  This chapter explores a new method for building KnowNets, named deepKnowNets. Basically, instead of a Word Sense Disambiguation algorithm, the method exploits a graph-based similarity measure to rerank the topic signatures.
- **Chapter 7: Concluding remarks and future directions**
  This chapter draws the main conclusions of this thesis and outlines some further steps to follow.

3 Main Contributions

The knowledge acquisition bottleneck problem is particularly acute for open domain (and also domain specific) semantic processing. However, we acquired by fully automatic means highly connected knowledge bases, increasing the total number of semantic relations from less than one million (the current number of available relations in WordNet) to millions of new and accurate semantic relations between WordNet senses. The different versions of KnowNet seem to be a major step towards the autonomous acquisition of knowledge from text, since they are several times larger than the available knowledge resources which encode relations between WordNet senses, and the knowledge they contain outperforms any other resources when they are empirically evaluated in a common framework.

Firstly, in order to acquire relevant semantic relations from large text collections corresponding to general or particular domains, we apply several methodologies and settings to automatically acquire **topic signatures** (TS) (Cuadros, Padró, and Rigau, 2005; Cuadros, Padró, and Rigau, 2006). Originally, topic signatures were used to describe a set of words related to the same topic or domain, but in our case, the topic is a WordNet sense[6]. Thus, topic signatures are sets of words related to that particular WordNet sense. We

[6]The name of Topic Signature, instead of the more appropriate *concept signature*, *word sense signature* or *synset signature*, is maintained for consistency with the literature

use in this research the original topic signatures acquired from the web [7] together with new sets of automatically acquired topic signatures which result in new acquisition methods, new tools and different resources, including different types of corpora. All these topic signatures are compared in a common framework together with existing knowledge bases. **ExRetriever**[8] (Cuadros et al., 2004; Cuadros et al., 2005) is used for the automatic acquisition of examples of particular WordNet word senses. It is a tool to automatically extract a subcorpus of text examples from a large corpus (for instance, BNC, SemCor or the Web).

Secondly, in order to identify the *implicit* semantic relations encoded by a Topic Signature (the sets of words related to a particular WordNet sense) with respect an *existing* knowledge base (in this case, also WordNet), we apply a graph-based Word Sense Disambiguation (WSD) algorithm, SSI-Dijkstra (Cuadros and Rigau, 2008b), also developed in the framework of this thesis. SSI-Dijkstra is based on the Structural Semantic Interconnections (SSI) algorithm. The method uses a robust and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to semantically related words associated to a particular WordNet sense. In that way, the method identifies a particular WordNet sense for each word in the Topic Signature, converting the original list of concept-to-word relations into a list of concept-to-concept relations.

Thirdly, a variant of SSI-Dijkstra has been applied in a task to integrate a very large domain thesaurus with millions of Species into WordNet (Toral et al., 2010; Cuadros et al., 2010). The process disambiguate every taxonomy of species in several languages.

Finally, the full list of new concept-to-concept relations between WordNet senses forms new knowledge bases, which we call **KnowNet**[9] (Cuadros and Rigau, 2008b) and **deepKnowNet**. Different sets of new KnowNets are empirically evaluated in different evaluation frameworks (Cuadros and Rigau, 2008b; Cuadros and Rigau, 2008c; Cuadros and Rigau, 2008a; Cuadros and Rigau, 2008d; Agirre et al., 2010).

[7] http://ixa.si.ehu.es/Ixa/resources/sensecorpus

[8] http://www.lsi.upc.edu/ nlp/meaning/downloads.html

[9] http://adimen.si.ehu.es/web/KnowNet

References

Agirre, Eneko, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring Knowledge Bases for Similarity. In *Proceedings of LREC 2010. ISBN: 2-9517408-6-7. Pages 373–377.*

Cuadros, Montse, Jordi Atserias, Mauro Castillo, and German Rigau. 2004. Automatic Acquisition of Sense Examples Using Exretriever. In *Proceedings of IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*, pages 97–104, November.

Cuadros, Montse, Jordi Atserias, Mauro Castillo, and German Rigau. 2005. The MEANING approach for automatic acquisition of sense examples. *MEANING Workshop*, February.

Cuadros, Montse, Egoitz Laparra, German Rigau, Piek Vossen, and Wauter Bosma. 2010. Integrating a large domain ontology of species into WordNet. In *Proceedings of LREC 2010*, La Valleta, Malta.

Cuadros, Montse, Lluís Padró, and German Rigau. 2005. Comparing Methods for Automatic Acquisition of Topic Signatures. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'05)*, September.

Cuadros, Montse, Lluís Padró, and German Rigau. 2006. An Empirical Study for Automatic Acquisition of Topic Signatures. In *Proceedings of Third International WordNet Conference (GWC 06)*, pages 51–59, Jeju Island (Korea), January. ISBN 80-210-3915-9.

Cuadros, Montse and German Rigau. 2008a. *Bases de Conocimiento Multilíng ües para el Procesamiento Semántico a Gran Escala.* Procesamiento del Lenguaje Natural, Vol. 40, 35–42.

Cuadros, Montse and German Rigau. 2008b. KnowNet: A proposal for building highly connected and dense knowledge bases from the web. In *First Symposium on Semantics in Systems for Text Processing, STEP'08.*, Venice, Italy, September.

Cuadros, Montse and German Rigau. 2008c. KnowNet: using Topic Signatures acquired from the web for building automatically highly dense knowledge bases. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, August.

Cuadros, Montse and German Rigau. 2008d. Multilingual Evaluation of KnowNet. *Procesamiento del Lenguaje Natural*, 41.

Toral, Antonio, Monica Monachini, Claudia Soria, Montse Cuadros, German Rigau, Wauter Bosma, and Piek Vossen. 2010. Linking a domain thesaurus to WordNet and conversion to WordNet-LMF. In *Proceedings of ICGL 2010*, Hong Kong.

Información General

SEPLN 2013

XXIX CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Universidad Complutense de Madrid- Facultad de Informática – Madrid (España)
18-20 de septiembre 2013
http://www.sepln.org/ y http://nil.fdi.ucm.es/sepln2013/

1 Presentación

La XXIX edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 18, 19 y 20 de septiembre de 2013 en la Facultad de Informática de la Universidad Complutense de Madrid.

La ingente cantidad de información disponible en formato digital y en las distintas lenguas que hablamos hace imprescindible disponer de sistemas que permitan acceder a esa enorme biblioteca que es Internet de manera cada vez más estructurada.

En este mismo escenario, hay un interés renovado por la solución de los problemas de accesibilidad a la información y de mejora de explotación de la misma en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: Extracción y recuperación de información, Sistemas de búsqueda de respuestas, Traducción automática, Análisis automático del contenido textual, Resumen automático, Generación textual y Reconocimiento y síntesis de voz.

2 Objetivos

El objetivo principal del congreso es ofrecer un foro para presentar las últimas investigaciones y desarrollos en el ámbito de trabajo del Procesamiento del Lenguaje Natural (PLN) tanto a la comunidad científica como a las empresas del sector. También se pretende mostrar las posibilidades reales de aplicación y conocer nuevos proyectos I+D en este campo.

Además, como en anteriores ediciones, se desea identificar las futuras directrices de la investigación básica y de las aplicaciones previstas por los profesionales, con el fin de contrastarlas con las necesidades reales del mercado. Finalmente, el congreso pretende ser un marco propicio para introducir a otras personas interesadas en esta área de conocimiento

3 Áreas Temáticas

Se anima a grupos e investigadores a enviar comunicaciones, resúmenes de proyectos o demostraciones en alguna de las áreas temáticas siguientes, entre otras:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Desarrollo de recursos y herramientas lingüísticas.
- Gramáticas y formalismos para el análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Resolución de la ambigüedad léxica.
- Generación textual monolingüe y multilingüe
- Traducción automática
- Síntesis del habla
- Sistemas de diálogo
- Indexado de audio
- Identificación idioma
- Extracción y recuperación de información monolingüe y multilingüe
- Sistemas de búsqueda de respuestas.
- Evaluación de sistemas de PLN.

ISSN 1135-5948

- Análisis automático del contenido textual.
- Análisis de sentimientos y opiniones.
- Análisis de plagio.
- Minería de texto en blogosfera y redes sociales.
- Generación de Resúmenes.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.

4 Formato del Congreso

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, pósters, proyectos de investigación en marcha y demostraciones de aplicaciones. Además prevemos la organización de talleres-workshops satélites para el día 17 de septiembre.

5 Comité ejecutivo SEPLN 2013

Presidente del Comité Organizador

- Luis Hernández (Universidad Complutense de Madrid)

Miembros

- Miguel Ballesteros (Universidad Pompeu Fabra)
- Susana Bautista (Universidad Complutense de Madrid)
- Alberto Díaz (Universidad Complutense de Madrid)
- Pablo Gervás (Universidad Complutense de Madrid)
- Raquel Hervás (Universidad Complutense de Madrid)
- Carlos León (UTAD - Centro Universitario de Tecnología y Arte Digital)
- Henry Anaya-Sánchez (Universitat Jaume I de Castellón)
- Gonzalo Méndez (Universidad Complutense de Madrid)

6 Consejo Asesor

Miembros:

- Prof. José Gabriel Amores Carredano (Universidad de Sevilla)
- Prof. Toni Badia i Cardús (Universitat Pompeu Fabra)
- Prof. Manuel de Buenaga Rodríguez (Universidad Europea de Madrid)
- Prof. Fco. Javier Calle Gómez (Universidad Carlos III de Madrid)
- Prof.ª Irene Castellón Masalles (Universitat de Barcelona)
- Prof.ª Arantza Díaz de Ilarraza (Euskal Herriko Unibertsitatea)
- Prof. Antonio Ferrández Rodríguez (Universitat d'Alacant)
- Prof. Mikel Forcada Zubizarreta (Universitat d'Alacant)
- Prof.ª Ana María García Serrano (UNED)
- Prof. Koldo Gojenola Galletebeitia (Euskal Herriko Unibertsitatea)
- Prof. Xavier Gómez Guinovart (Universidade de Vigo)
- Prof. Julio Gonzalo Arroyo (Universidad Nacional de Educación a Distancia)
- Prof. José Miguel Goñi Menoyo (Universidad Politécnica de Madrid)
- José B. Mariño Acebal(Universitat Politécnica de Catalunya)
- Prof.ª M. Antonia Martí Antonín (Universitat de Barcelona)
- Prof.ª Mª Teresa Martín Valdivia (Universidad de Jaén)
- Prof. Patricio Martínez Barco (Universitat d'Alacant)
- Prof. Paloma Martínez Fernández (Universidad Carlos III de Madrid)
- Profª. Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia)
- Prof.ª Lidia Ana Moreno Boronat (Universitat Politécnica de Valencia)
- Prof. Lluis Padró (Universitat Politécnica de Catalunya)
- Prof. Manuel Palomar Sanz (Universitat d'Alacant)
- Prof. Ferrán Pla (Universitat Politécnica de Valencia)
- Prof. Germán Rigau (Euskal Herriko Unibertsitatea)
- Prof. Horacio Rodríguez Hontoria (Universitat Politécnica de Catalunya)
- Prof. Kepa Sarasola Gabiola (Euskal Herriko Unibertsitatea)

- Prof. Emilio Sanchís (Universitat Politécnica de Valencia)
- Prof. L. Alfonso Ureña López (Universidad de Jaén)
- Prof.ª Mª Felisa Verdejo Maillo (Universidad Nacional de Educación a Distancia)
- Prof. Manuel Vilares Ferro (Universidade de Vigo)
- Prof. Ruslan Mitkov (Universidad de Wolverhampton)
- Prof.ª Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, France)
- Prof. Leonel Ruiz Miyares (Centro de Linguistica Aplicada de Santiago de Cuba)
- Investigador Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica. México)
- Investigador Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica. México)
- Prof. Alexander Gelbukh (Instituto Politécnico Nacional. México)
- Prof. Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa. Portugal)
- Prof. Bernardo Magnini (Fondazione Bruno Kessler. Italia)

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 8 de abril de 2013.
- Notificación de aceptación: 30 de mayo de 2013.
- Fecha límite para entrega de la versión definitiva: 24 de junio de 2013.
- Fecha límite para propuesta de talleres y tutoriales: 15 de marzo de 2013.

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/?page_id=358

Hoja de Inscripción para Instituciones

Datos Entidad/Empresa

Nombre : .....
NIF : ..... Teléfono : .....
E-mail : ..... Fax : .....
Domicilio : .....
Municipio : ..... Código Postal : ..... Provincia : .....
Áreas de investigación o interés: .....
.....

Datos de envío

Dirección : ..... Código Postal : .....
Municipio : ..... Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

Sr. Director de:

Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....
Núm Cuenta : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

.....

Cuotas de los socios institucionales*: 300 €.*

Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Hoja de Inscripción para Socios

Datos Personales

Apellidos : .....
Nombre : .....
DNI : ..... Fecha de Nacimiento : .....
Teléfono : ..... E-mail : .....
Domicilio : .....
Municipio : ..... Código Postal : .....
Provincia : .....

Datos Profesionales

Centro de trabajo : .....
Domicilio : .....
Código Postal : ..... Municipio : .....
Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....
Áreas de investigación o interés: .....
.....

Preferencia para envío de correo:

[ ] Dirección personal [ ] Dirección Profesional

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

En.....a.....de.....de.....
(firma)

Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN

Sr. Director de:
Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

Cuotas de los socios*: 18 €* (residentes en España) o *24 €* (socios residentes en el extranjero).
Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

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

José Gabriel Amores	Universidad de Sevilla
Toni Badía	Universidad Pompeu Fabra
Manuel de Buenaga	Universidad Europea de Madrid
Irene Castellón	Universidad de Barcelona
Arantza Díaz de Ilarraza	Universidad del País Vasco
Antonio Ferrández	Universidad de Alicante
Mikel Forcada	Universidad de Alicante
Ana García-Serrano	UNED
Koldo Gojenola	Universidad del País Vasco
Xavier Gómez Guinovart	Universidad de Vigo
Julio Gonzalo	UNED
José Miguel Goñi	Universidad Politécnica de Madrid

Cartas al director

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

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://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/?cat=21
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/?page_id=1061