

ML-SENTICON: Un lexicón multilingüe de polaridades semánticas a nivel de lemas

ML-SENTICON: *a multilingual, lemma-level sentiment lexicon*

Fermín L. Cruz, José A. Troyano, Beatriz Pontes, F. Javier Ortega

Universidad de Sevilla

Escuela Técnica Superior de Ingeniería Informática, Av. Reina Mercedes s/n
{fcruz,troyano,bepontes,javierortega}@us.es

Resumen: En este trabajo presentamos un conjunto de lexicones de polaridades semánticas a nivel de lemas para inglés, español, catalán, gallego y euskera. Estos lexicones están estructurados en capas, lo que permite seleccionar distintos compromisos entre la cantidad de estimaciones de positividad y negatividad y la precisión de dichas estimaciones. Los lexicones se han generado automáticamente a partir de una mejora del método utilizado para generar SentiWordNet, un recurso ampliamente utilizado que recoge estimaciones de positividad y negatividad a nivel de synsets. Nuestras evaluaciones sobre los lexicones para inglés y español muestran altos niveles de precisión en todas las capas. El recurso que contiene todos los lexicones obtenidos, llamado ML-SENTICON, está disponible de forma pública para su uso.

Palabras clave: Análisis del sentimiento, recursos léxicos, recursos multilingüe

Abstract: In this work, we present a set of lemma-level sentiment lexicons for English, Spanish, Catalan, Basque and Galician. These lexicons are layered, allowing to trade off between the amount of available words and the accuracy of the estimations. The lexicons have been automatically generated from an improved version of SentiWordNet, a very popular resource which contains estimations of the positivity and negativity of synsets. Our evaluations on the English and Spanish lexicons show high accuracies in all cases. The resource containing all the lexicons, ML-SENTICON, is publicly available.

Keywords: Sentiment analysis, lexical resources, multilingual resources

1 Introducción

El Análisis del Sentimiento es una disciplina enmarcada en el Procesamiento del Lenguaje Natural que se ocupa del tratamiento computacional de fenómenos como la subjetividad, las emociones y las opiniones en textos; una buena introducción al estado de la cuestión se encuentra en (Liu y Zhang, 2012). Se trata de un área de investigación muy activa en los últimos años, debido en gran parte a que las opiniones expresadas en Internet por los usuarios constituyen una fuente de información de incalculable utilidad para las Administraciones Públicas, las grandes y medianas compañías y los consumidores. Los contenidos de esta naturaleza generados por usuarios son de un volumen y velocidad de aparición tales que imposibilitan su análisis exclusivamente manual y requieren de la aplicación de métodos automáticos de apoyo. Algunas de las tareas definidas en el área son la detección

de textos subjetivos, la clasificación de textos basada en la polaridad de las opiniones expresadas (positiva/negativa), o la extracción de opiniones individuales y sus participantes. Muchas de las soluciones propuestas a estas tareas se apoyan en lexicones de opiniones o de polaridad (*opinion lexicons* o *sentiment lexicons* en inglés), recursos léxicos que contienen información sobre las implicaciones emocionales de las palabras. Generalmente, esta información consiste en la polaridad a priori de las palabras, es decir, las connotaciones positivas o negativas de dichas palabras en ausencia de contexto.

En este trabajo presentamos nuevos lexicones para inglés, español, catalán, gallego y euskera. Los lexicones están organizados en varias capas, lo que permite a las aplicaciones que los utilicen seleccionar distintos compromisos entre la cantidad de palabras disponibles y la precisión de las estimaciones de

sus polaridades a priori. Para generar estos lexicones, como paso previo, hemos reproducido el método utilizado en (Baccianella, Esuli, y Sebastiani, 2010) para construir SENTIWORDNET 3.0, un recurso léxico construido sobre WordNet y ampliamente utilizado en el área del Análisis del Sentimiento. Hemos incorporado diversas mejoras al método original, que repercutieron positivamente en la calidad del recurso obtenido, según nuestras evaluaciones.

El resto del artículo se estructura como sigue. En la sección 2 repasamos algunos ejemplos representativos de trabajos sobre lexicones de opinión, tanto en inglés como en otros idiomas, centrándonos especialmente en el español. En las secciones 3 y 4 describimos el proceso de construcción de nuestro recurso; dicho proceso está dividido en dos partes fundamentales, descritas y evaluadas cada una por separado. Por último, en la sección 5 resumimos el contenido anterior y señalamos algunas conclusiones y líneas de trabajo futuro.

2 Trabajos relacionados

Existen muchos trabajos que abordan la creación de lexicones de opinión, con enfoques muy distintos. *General Inquirer* (Stone, Dunphy, y Smith, 1966) puede ser considerado, entre otras cosas, el primer lexicón que contiene información afectiva. Se trata de un lexicón construido manualmente, formado por lemas, es decir, unidades semánticas que pueden aparecer en múltiples formas lexicalizadas (por ejemplo, el verbo *approve* es un lema que puede aparecer en los textos con diferentes flexiones, como *approved* o *approving*). *General Inquirer* contiene una gran cantidad de información asociada a cada lema, de tipo sintáctico, semántico y pragmático. Entre toda esta información, existen un total de 4206 lemas que están etiquetados como positivos o negativos. A pesar de su antigüedad, *General Inquirer* es aún usado en muchos trabajos de análisis del sentimiento.

El *MPQA Subjectivity Lexicon* (Wilson, Wiebe, y Hoffmann, 2005) es un ejemplo de lo anterior. Se trata de un lexicón que aglutina palabras subjetivas obtenidas mediante un método automático (Riloff y Wiebe, 2003), términos obtenidos a partir de un diccionario y un *thesaurus*, y las listas de palabras positivas y negativas de *General Inquirer*. En total contiene 8221 palabras, cuyas polaridades

(positiva, negativa o neutra) fueron anotadas manualmente. La lista final contiene 7631 entradas positivas y negativas, y es muy heterogénea, conteniendo palabras flexionadas y otras lematizadas. Al igual que *General Inquirer*, la lista no contiene términos formados por más de una palabra.

Los dos lexicones de opinión más utilizados en la actualidad, atendiendo al número de citas, son *Bing Liu's Opinion Lexicon* (Hu y Liu, 2004; Liu, Hu, y Cheng, 2005) y SENTIWORDNET (Esuli y Sebastiani, 2006; Baccianella, Esuli, y Sebastiani, 2010). Ambos presentan enfoques muy distintos y en cierto sentido opuestos. Mientras que el *Bing Liu's Opinion Lexicon* está formado por unas 6800 palabras flexionadas, incluyendo incluso faltas de ortografía y *slangs* (expresiones informales utilizadas frecuentemente en Internet), SENTIWORDNET está construido sobre WORDNET (Fellbaum, 1998), un recurso léxico en el que las unidades básicas, llamadas *synsets*, aglutinan a distintas palabras que comparten un mismo significado. El *Bing Liu's Opinion Lexicon* está construido a partir de un método automático, pero las listas de palabras positivas y negativas han sido actualizadas manualmente de forma periódica, hasta la versión actualmente disponible en la web, que data de 2011. Por su parte, SENTIWORDNET asigna valores reales de positividad y negatividad, entre 0 y 1, a cada uno de los más de 100 mil *synsets* presentes en WORDNET. Los valores han sido calculados mediante un método automático, partiendo de unos conjuntos semilla de *synsets* positivos y negativos.

Es importante señalar la diferencia entre los lexicones a nivel de palabras o lemas, como *General Inquirer*, el *MPQA Subjectivity Lexicon* o el *Bing Liu's Opinion Lexicon*, y los lexicones a nivel de *synsets*, como SENTIWORDNET. Los primeros están formados por términos con ambigüedad semántica, debido a la polisemia de muchas de las palabras. Sin embargo, los lexicones basados en *synsets* no presentan este problema, puesto que las unidades que los forman representan unívocamente un sólo significado. Por contra, para poder emplear un lexicón basado en *synsets* es necesario aplicar a los textos que se van a analizar una herramienta de desambiguación de significados, las cuales tienen una precisión aún relativamente baja hoy por hoy. La mayoría de los trabajos que utilizan SENTI-

WORDNET optan por calcular valores agregados de polaridad a nivel de palabras o lemas, a partir de los valores de todos los posibles *synsets* correspondientes (Taboada et al., 2011; Denecke, 2008; Martín-Valdivia et al., 2012; Agrawal y others, 2009; Saggion y Funk, 2010; Kang, Yoo, y Han, 2012; Desmet y Hoste, 2013). En este trabajo hemos construido lexicones de ambos tipos: a nivel de *synsets*, introduciendo algunas mejoras al método empleado para la construcción de SENTIWORDNET 3.0, y a nivel de lemas, a partir de los valores obtenidos en el lexicón anterior.

2.1 Lexicones de opiniones en otros idiomas

Aunque aún son pocos los lexicones de opinión disponibles para idiomas distintos al inglés, poco a poco van creciendo en número. Existen trabajos que se centran en la creación de lexicones para idiomas tan diversos como el hindú y el francés (Rao y Ravichandran, 2009), el árabe (Abdul-Mageed, Diab, y Koraem, 2011), el alemán (Clematide y Klenner, 2010), el japonés (Kaji y Kitsuregawa, 2007), el chino (Lu et al., 2010; Xu, Meng, y Wang, 2010), el rumano (Banea, Mihalcea, y Wiebe, 2008) y el español. Creemos que en el caso del español son pocos aún los lexicones disponibles actualmente, teniendo en cuenta que el español es el segundo idioma más hablado en el mundo y el tercero más utilizado en Internet.

En (Brooke, Tofiloski, y Taboada, 2009) se utilizan dos recursos (un diccionario bilingüe¹ y Google Translator²) para obtener, de manera automática y a partir de un lexicón en inglés, dos lexicones en español. El trabajo no muestra resultados de evaluación para los lexicones obtenidos, sino que presenta los resultados de una herramienta de clasificación basada en la polaridad que se apoya en los lexicones. Una técnica similar es empleada en (Molina-González et al., 2013), donde se aplica traducción automática de inglés a español al Bing Liu's Opinion Lexicon. Algunas de las ambigüedades inherentes a la traducción fueron resueltas manualmente. También se añadieron manualmente traducciones de algunas palabras informales y *slangs* contenidos en el lexicón de partida. Tampoco en este trabajo se reportan resul-

tados de evaluación directa del lexicón obtenido, sino que se presentan resultados de evaluación extrínseca basados en clasificación binaria del sentimiento. Pérez-Rosas, Banea, y Mihalcea (2012) parten de dos lexicones en inglés: el MPQA Subjectivity Lexicon (el cual es mapeado sobre WordNet mediante un método automático) y SENTIWORDNET. Después utilizan una versión en español de WORDNET que forma parte del proyecto EuroWordNet (Vossen, 1998) para obtener listas de términos positivos y negativos en español. El trabajo plantea dos variaciones del método, a partir de las que se obtienen dos lexicones en español, uno formado por 1347 términos y otro por 2496. Las evaluaciones realizadas mediante la anotación manual de una muestra de 100 términos de cada uno de los lexicones arrojan un 90% y un 74% de *accuracy* respectivamente.

3 Estimación de la polaridad a nivel de *synsets*

Nuestro objetivo es la obtención de lexicones a nivel de lemas en varios idiomas. Como paso previo a la obtención de dichos lexicones, hemos construido un lexicón en inglés a nivel de *synsets*. Para la creación de este lexicón, partimos del método utilizado para la construcción de SENTIWORDNET 3.0, introduciendo mejoras en cada una de las etapas del método. Una vez obtenido este recurso, generamos lexicones a nivel de lemas para inglés, español, catalán, gallego y euskera. Las evaluaciones llevadas a cabo muestran mejoras significativas tanto en el lexicón a nivel de *synsets* (comparándonos con SENTIWORDNET 3.0), como en el lexicón a nivel de lemas, con valores de precisión y volumen para el lexicón en español superiores a los de (Pérez-Rosas, Banea, y Mihalcea, 2012).

Basándonos en el método empleado por (Baccianella, Esuli, y Sebastiani, 2010) para la generación de SENTIWORDNET 3.0 e incorporando diversas modificaciones, hemos calculado valores reales entre 0 y 1 de positividad, negatividad y objetividad para cada uno de los *synsets* de WORDNET 3.0. Al igual que el método en el que nos basamos, nuestro método se divide en dos partes claramente diferenciadas: un primer cálculo individual de la polaridad, y un segundo cálculo global de la polaridad a partir de los valores obtenidos en la primera etapa.

¹<http://www.spanishdict.com>

²<http://translate.google.com>

3.1 Cálculo individual de polaridad

El cálculo individual de la polaridad se basa en la construcción de clasificadores ternarios, capaces de decidir si un *synset* es positivo, negativo o neutro a partir de los textos de sus glosas (las glosas son definiciones contenidas en WORDNET para cada uno de los *synsets*). Para entrenar estos clasificadores, se parte de distintos conjuntos de *synsets* considerados a priori positivos, negativos o neutros. En (Baccianella, Esuli, y Sebastiani, 2010) se utilizaron los *synsets* correspondientes a palabras positivas y negativas usadas por Turney y Littman (2003)³ En nuestro caso, hemos utilizado también WORDNET-AFFECT (Strapparava, Valitutti, y Stock, 2006) como fuente de semillas positivas y negativas. Los clasificadores entrenados a partir de las distintas fuentes de información, y usando dos algoritmos de clasificación distintos (Rocchio y SVM), fueron combinados en una etapa de meta-aprendizaje, obteniéndose finalmente tres clasificadores regresionales capaces de inducir valores de positividad, negatividad y objetividad en el intervalo $[0, 1]$. Los detalles de esta etapa pueden consultarse en (Cruz et al., 2014).

3.2 Cálculo global de polaridad

El cálculo global de la polaridad trata de refinar en su conjunto los valores de positividad y negatividad asignados a cada *synset*, a partir de distintos tipos de relaciones entre ellos. Estas relaciones se modelan mediante un grafo en el que los *synsets* son representados mediante nodos y las aristas dirigidas indican algún tipo de relación entre los valores de positividad y negatividad de dichos *synsets*. Se asigna a cada nodo un valor numérico (por ejemplo, los valores de positividad), y se aplica entonces al grafo un algoritmo de paseo aleatorio. Estos algoritmos son capaces de computar iterativamente las interacciones que se producen entre los valores asignados a los nodos: los valores inicialmente asignados a los nodos “fluyen” a lo largo del grafo a través de las aristas. Una vez ha convergido, el algoritmo obtiene valores finales para los nodos, que dependen tanto de los valores inicialmente asignados a los mismos como de

³Positivas: good, nice, excellent, positive, fortunate, correct, superior. Negativas: bad, nasty, poor, negative, unfortunate, wrong, inferior.

las relaciones existentes entre los nodos a nivel global.

Las diferencias fundamentales de nuestra propuesta con respecto a SENTIWORDNET 3.0 en este paso son dos. En primer lugar, construimos dos tipos de grafos distintos, uno a partir de las glosas y otro a partir de las relaciones semánticas de WORDNET (en SENTIWORDNET se emplea únicamente un grafo basado en las glosas). En ambos casos, los grafos incluyen aristas con peso positivo, que representan una transferencia directa entre los valores de positividad y negatividad de los *synsets* conectados, y aristas con peso negativo, que indican una transferencia cruzada entre ambos tipos de valores (en SENTIWORDNET sólo se contemplan aristas sin pesos). En segundo lugar, aplicamos POLARITYRANK (Cruz et al., 2012), un algoritmo de paseo aleatorio sobre grafos que permite computar los valores finales de positividad y negatividad en una sola ejecución, existiendo además una interdependencia entre los valores finales positivos y negativos (en SENTIWORDNET se llevaban a cabo dos ejecuciones independientes del algoritmo PAGERANK, una para los valores de positividad y otra para los de negatividad). Los detalles de esta etapa también están explicados más ampliamente en (Cruz et al., 2014).

3.3 Evaluación

En la tabla 1 se muestran los valores de la distancia τ_p de Kendall (Fagin et al., 2004) entre un *gold standard* y los valores obtenidos con nuestro método. Esta medida estima la similitud entre un *ranking* modelo o *gold standard* y otro *ranking* candidato. Cuanto más cercano a cero, más parecidos son ambos *rankings*. Hemos usado el mismo *gold standard* usado en (Baccianella, Esuli, y Sebastiani, 2010), por lo que los resultados son comparables con los mostrados en dicho trabajo. Como puede apreciarse, hemos conseguido mejoras significativas en ambas etapas, con estimaciones finales de positividad y negatividad usando nuestro método más precisas que las de SENTIWORDNET 3.0 (se reduce τ_p un 24,2% y un 7,4%, respectivamente).

4 Inducción de ML-SENTICON

Para facilitar el uso del recurso por parte de aquellos investigadores que no deseen utilizar desambiguación de significados, hemos generado un lexicón a nivel de lemas partiendo del

Etapa	Recurso	Positiv.	Negativ.
1	SWN	0,339	0,286
	ML-SC	0,238	0,284
2	SWN	0,281	0,231
	ML-SC	0,213	0,214

Tabla 1: Valores de τ_p de SentiWordNet (SWN) y ML-SentiCon (ML-SC) obtenidos en cada etapa del método de cálculo de valores de positividad y negatividad de *synsets* (1: Cálculo individual; 2: Cálculo global).

lexicón a nivel de *synsets* anterior. Además, usando recursos que nos permiten conectar los *synsets* con lemas en otros idiomas, hemos generado versiones del lexicón en español, catalán, gallego y euskera. El recurso finalmente obtenido, llamado ML-SENTICON, está disponible de forma pública para su uso ⁴.

Cada lexicón a nivel de lemas está formado por ocho capas. Las capas están ordenadas, desde la primera hasta la octava, de manera que las capas posteriores contienen todos los lemas de las anteriores, y añaden algunos nuevos. Los lemas que conforman cada una de las capas son obtenidos rebajando progresivamente una serie de restricciones, de manera que el número de lemas que las cumplen va aumentando capa tras capa, a costa de una bajada paulatina en la fiabilidad de dichos lemas como indicadores de positividad y negatividad.

4.1 Definición de las capas

Cada *synset* s_i en WORDNET tiene asociado un conjunto de lemas $L_i = \{l_1, l_2, \dots, l_n\}$ (también llamados *variants*), todos con la misma categoría morfosintáctica (nombre, adjetivo, verbo o adverbio). Además, cada *synset* s_i tiene un valor de positividad p_i y un valor de negatividad n_i en el recurso obtenido anteriormente. Diremos que la polaridad del *synset* s_i es $pol_i = p_i - n_i$. De cara a la definición del lexicón a nivel de lemas, consideremos que un *synset* s_i es la tupla formada por el conjunto de lemas y la polaridad, es decir, $s_i = (L_i, pol_i)$. Invirtiendo esta asociación, a cada lema l le corresponde un conjunto de *synsets* $S_l = \{s_i : l \in L_i\}$. Denotamos con \overline{pol}_l a la media de las polaridades pol_i de los *synsets* $s_i \in S_l$. Cada una de las ocho capas está formada por un conjunto de lemas positivos l tales que $\overline{pol}_l > 0$, y otro conjunto de lemas negativos l tales que $\overline{pol}_l < 0$.

⁴<http://www.lsi.us.es/~fermin/index.php/Datasets>

Capa	en	sp	ca	eu	gl
1	157	353	512	138	49
2	982	642	530	278	223
3	1600	891	699	329	370
4	2258	1138	860	404	534
5	3595	1779	1247	538	883
6	6177	2849	1878	888	1429
7	13517	6625	4075	2171	2778
8	25690	11918	7377	4323	4930

Tabla 2: Distribución de lemas por capas en los lexicones en inglés (en), español (es), catalán (ca), euskera (eu) y gallego (gl).

Las dos primeras capas están formadas exclusivamente por lemas $l \in L_i$ de *synsets* s_i que pertenecen a alguno de los conjuntos de entrenamiento usados en la etapa de cálculo individual de la polaridad. El resto de capas están formadas por lemas $l \in L_i$ de cualquier *synset* s_i ; en cada capa se exigen valores mínimos diferentes sobre el valor absoluto de \overline{pol}_l de los lemas l que las conforman. Estos valores mínimos se han escogido tratando de obtener una progresión geométrica en el número de lemas que componen cada una de las capas.

4.2 Obtención de lemas en otros idiomas

Para obtener correspondencias entre los *synsets* y lemas en otros idiomas distintos al inglés, hemos utilizado el *Multilingual Central Repository 3.0* (MCR 3.0) (Gonzalez-Agirre, Laparra, y Rigau, 2012). Este recurso se compone de WordNets incompletos para cuatro idiomas: español, catalán, euskera y gallego. Los *synsets* de estos WordNets están conectados con los de WORDNET 3.0, lo que nos permite replicar el mismo procedimiento de construcción de las capas explicado en la sección anterior. Para el caso del español y el catalán, hemos utilizado también la información generada por el proyecto *EuroWordNet* (Vossen, 1998) a fecha de noviembre de 2006, lo que nos permite aumentar el número de lemas para estos idiomas. *EuroWordNet* se basa en WORDNET 1.6, por lo que hemos tenido que realizar un mapeo a WORDNET 3.0 mediante *WN-Map*⁵ (Daudé, Padró, y Rigau, 2003). En la tabla 2 se muestra la distribución de lemas por capas e idiomas de los lexicones obtenidos.

⁵<http://nlp.lsi.upc.edu/tools/download-map.php>

4.3 Evaluación

Para evaluar la calidad de los lexicones a nivel de lemas, hemos revisado manualmente las listas de lemas positivos y negativos de cada una de las capas, etiquetando cada entrada como correcta o incorrecta. Hemos evaluado de esta forma los lexicones en inglés y español. Para los cuatro primeros niveles (niveles 1-4), hemos revisado las listas completas⁶. Para el resto de niveles (niveles 5-8), hemos revisado una muestra aleatoria estadísticamente representativa de cada uno de los niveles. Hemos calculado el tamaño de la muestra que garantiza un error muestral menor a $\pm 5\%$ en la estimación de la proporción de elementos correctos, suponiendo una distribución binomial de la variable aleatoria con $p = q = 0,5$ (el peor caso posible), y con un intervalo de confianza del 95% ($\alpha = 0,05$). Con estos parámetros, hemos obtenido tamaños muestrales de entre 300 y 400 elementos, según el nivel.

En la tabla 3 se muestra el *accuracy* estimado (porcentaje de elementos correctos frente al total) para cada capa de los lexicones en inglés y español, respectivamente. Los resultados confirman una gran fiabilidad de las listas de lemas positivos y negativos generadas, con valores por encima del 90% en las capas 1-6 del lexicon en inglés y las capas 1-5 del lexicon en español.

Como puede observarse, el *accuracy* del lexicon en inglés es mayor al del lexicon en español, lo cual es lógico puesto que el lexicon en español se ha construido a partir de recursos generados mediante métodos semiautomáticos y por tanto no carentes de errores. La diferencia de *accuracy* entre ambos lexicones va en aumento a lo largo de las capas, pasando de 1,63 puntos porcentuales en la primera capa a 12,27 puntos en la última capa. A pesar de esto, creemos que los valores medidos para el lexicon en español son buenos, si comparamos las capas 5 y 6 (91,75% en un lexicon de 1779 y 86,09% en un lexicon de 2849) con los lexicones en español de (Pérez-Rosas, Banea, y Mihalcea, 2012) (90% en un lexicon de 1347 lemas y 74% en uno de 2496 lemas). Más aún, la siguiente capa de nuestro recurso continúa teniendo un mejor nivel de acierto (77,69%) que el mayor de los lexicones de (Pérez-Rosas, Banea, y Mihalcea,

⁶En el recurso hecho público, hemos incluido las versiones libres de lemas erróneos de las cuatro primeras capas.

Capa	Inglés		Español	
	Acc.	Tam.	Acc.	Tam.
1	99,36%	157	97,73%	353
2	98,88%	982	97,20%	642
3	97,75%	1600	94,95%	891
4	96,24%	2258	93,06%	1138
5	93,95%	3595	91,75%	1779
6	91,99%	6177	86,09%	2849
7	85,29%	13517	77,69%	6625
8	74,06%	25690	61,29%	11918

Tabla 3: Estimación muestral del porcentaje de lemas con polaridad correcta (Acc.) y número de lemas total (Tam.) de cada una de las capas de los lexicones en inglés y español.

2012), con un número de lemas muy superior (6625).

5 Conclusiones

En este trabajo presentamos un nuevo recurso, llamado ML-SENTICON, formado por lexicones a nivel de lemas en inglés, español, catalán, gallego y euskera. Los lexicones están formados por capas, lo que permite seleccionar distintos compromisos entre la cantidad de términos disponibles y la precisión de las estimaciones de sus polaridades a priori. Para cada lema, el recurso proporciona un valor real que representa dicha polaridad, entre -1 y 1, y un valor de desviación estándar que refleja la ambigüedad resultante del cómputo de la polaridad a partir de los valores de los distintos significados posibles del lema. El recurso está disponible de forma pública para su uso⁷.

Como paso previo en la obtención de estos lexicones, hemos presentado también una versión mejorada del método usado para construir SENTIWORDNET 3.0 (Baccianella, Esuli, y Sebastiani, 2010). Hemos llevado a cabo evaluaciones de nuestro método similares a las realizadas en el artículo anterior, cuyos resultados reflejan mejoras significativas en las estimaciones de polaridad obtenidas con nuestro método con respecto al método original. Los detalles de cada una de las mejoras implementadas se pueden consultar en (Cruz et al., 2014).

Creemos que el recurso obtenido puede ser útil en multitud de aplicaciones relacionadas con el análisis del sentimiento, tanto para inglés como para español. Aunque es

⁷<http://www.lsi.us.es/~fermin/index.php/Datasets>

presumible que las conclusiones en cuanto al porcentaje de acierto en la polaridad de los lemas del lexicon en español sean extrapolables al resto de idiomas incluidos (puesto que se han utilizado recursos y métodos equivalentes), sería deseable que otros investigadores más familiarizados con estos idiomas que los autores del presente trabajo estimaran la calidad de dichos lexicones. El método aquí propuesto puede ser reproducido para otros idiomas, siempre que existan WordNets disponibles. En este sentido, puede ser útil el recurso *Open Multilingual WordNet* (Bond y Foster, 2013), que reúne WordNets para multitud de idiomas procedentes de distintos proyectos internacionales.

Agradecimientos

Este trabajo ha sido financiado a través de los proyectos ATTOS (TIN2012-38536-C03-02) y DOCUS (TIN2011-14726-E) concedidos por el Ministerio de Ciencia e Innovación del Gobierno de España, y del proyecto AORESCU (P11-TIC-7684 MO) concedido por la Consejería de Innovación, Ciencia y Empresas de la Junta de Andalucía.

Bibliografía

- Abdul-Mageed, Muhammad, Mona T Diab, y Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. En *ACL (Short Papers)*, páginas 587–591.
- Agrawal, Shaishav y others. 2009. Using syntactic and contextual information for sentiment polarity analysis. En *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, páginas 620–623. ACM.
- Baccianella, Stefano, Andrea Esuli, y Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh conference on International Language Resources and Evaluation*. ELRA, may.
- Banea, Carmen, Rada Mihalcea, y Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. En *LREC*.
- Bond, Francis y Ryan Foster. 2013. Linking and extending an open multilingual wordnet. En *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*.
- Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*, páginas 50–54.
- Clematide, Simon y Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for german. En *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, página 7.
- Cruz, Fermín L, José A Troyano, Beatriz Pontes, y F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.
- Cruz, Fermín L., Carlos G. Vallejo, Fernando Enríquez, y José A. Troyano. 2012. Polarityrank: Finding an equilibrium between followers and contraries in a network. *Inf. Process. Manage.*, 48(2):271–282.
- Daudé, Jordi, Llus Padró, y German Rigau. 2003. Making wordnet mapping robust. *Procesamiento del lenguaje natural*, 31.
- Denecke, Kerstin. 2008. Using sentiwordnet for multilingual sentiment analysis. En *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, páginas 507–512. IEEE.
- Desmet, Bart y Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*.
- Esuli, Andrea y Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. En *Proceedings of Language Resources and Evaluation (LREC)*.
- Fagin, Ronald, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, y Erik Vee. 2004. Comparing and aggregating rankings with ties. En *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, páginas 47–58, New York, NY, USA. ACM.

- Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Gonzalez-Agirre, Aitor, Egoitz Laparra, y German Rigau. 2012. Multilingual central repository version 3.0. En *LREC*, páginas 2525–2529.
- Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *KDD '04: Proceedings of the tenth ACM SIGKDD*, páginas 168–177, New York, NY, USA. ACM.
- Kaji, Nobuhiro y Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. En *EMNLP-CoNLL*, páginas 1075–1083.
- Kang, Hanhoon, Seong Joon Yoo, y Dongil Han. 2012. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010.
- Liu, Bing, Minqing Hu, y Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. En *Proceedings of WWW*.
- Liu, Bing y Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. En Charu C. Aggarwal y ChengXiang Zhai, editores, *Mining Text Data*. Springer US, páginas 415–463.
- Lu, Bin, Yan Song, Xing Zhang, y Benjamin K Tsou. 2010. Learning chinese polarity lexicons by integration of graph models and morphological features. En *Information retrieval technology*. Springer, páginas 466–477.
- Martín-Valdivia, María-Teresa, Eugenio Martínez-Cámara, Jose-M Perea-Ortega, y L Alfonso Ureña-López. 2012. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*.
- Molina-González, M Dolores, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, y José M Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.
- Pérez-Rosas, Verónica, Carmen Banea, y Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. En *LREC*, páginas 3077–3081.
- Rao, Delip y Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 675–682. Association for Computational Linguistics.
- Riloff, Ellen y Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. En *Proceedings of EMNLP*.
- Saggion, Horacio y A Funk. 2010. Interpreting sentiwordnet for opinion classification. En *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*.
- Stone, Philip J, Dexter C Dunphy, y Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.
- Strapparava, Carlo, Alessandro Valitutti, y Oliviero Stock. 2006. The affective weight of lexicon. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, páginas 423–426.
- Taboada, Maite, Julian Brooke, Milan Toftiloski, Kimberly Voll, y Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- Turney, Peter D. y Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.
- Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston.
- Wilson, Theresa, Janyce Wiebe, y Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of the HLT/EMNLP*, páginas 347–354.
- Xu, Ge, Xinfan Meng, y Houfeng Wang. 2010. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. En *Proceedings of the 23rd International Conference on Computational Linguistics*, páginas 1209–1217. Association for Computational Linguistics.