ML-SentiCon: Un lexicón multilíngüe de polaridades semánticas a nivel de lemas

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 en el catálogo de recursos de la Red TIMM.

1 Introducción

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 Sebastiani2010] 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.
A continuación resumimos muy brevemente el proceso de inducción del recurso. Pueden consultarse los detalles de cada una de las etapas del metodo empleado en [Cruz et al.2014]

2 Inducción del recurso

Basándonos en el método empleado 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.

2.1 Cálculo individual de polaridad a nivel de synsets

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 SentiWordNet 3.0 se utilizaron los *synsets* correspondientes a palabras positivas y negativas usadas en [Turney y Littman2003]. En nuestro caso, hemos utilizado también WordNet-Affect[Strapparava, Valitutti, y Stock2006] 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 $[0, 1]$.

2.2 Cálculo global de polaridad a nivel de synsets

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*. 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).

2.3 Cálculo de la polaridad a nivel de lemas

Para facilitar el uso del recurso por parte de aquellos investigadores que no deseen utilizar desambigüación de significados, hemos generado un lexicón a nivel de lemas partiendo del lexicón a nivel de *synsets* anterior. 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.
Para obtener correspondencias entre los *synsets* y lemas para español, catalán, euskera y gallego, hemos utilizado el *Multilingual Central Repository 3.0* (MCR 3.0) [Gonzalez-Agirre, Laparra, y Rigau2012] y la información generada por el proyecto *EuroWordNet* [Vossen1998] a fecha de noviembre de 2006. *EuroWordNet* se basa en WordNet 1.6, por lo que hemos tenido que realizar un mapeo a WordNet 3.0 mediante *WN-Map*[1].

3 Evaluación y conclusiones

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 de nuestro lexicón a nivel de synsets. 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 Sebastiani2010], 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).
Para evaluar la calidad de los lexicones a nivel de lemas, hemos revisado manualmente las listas de lemas positivos y negativos de cada una de las capas, etiquetando cada entrada como correcta o incorrecta. Hemos evaluado de esta forma los lexicones en inglés y español. Para los cuatro primeros niveles (niveles 1-4), hemos revisado las listas completas[2]. Para el resto de niveles (niveles 5-8), hemos revisado una muestra aleatoria estadísticamente representativa de cada uno de los niveles (error muestral $< 5\,\%$, con $p = q = 0{,}5$ y $\alpha = 0{,}05$). En la tabla 2 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] http://nlp.lsi.upc.edu/tools/download-map.php
[2] En el recurso hecho público, hemos incluido las versiones libres de lemas erróneos de las cuatro primeras capas.

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).

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

Tabla 2. 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.

1-6 del lexicón en inglés y las capas 1-5 del lexicón en español. Como puede observarse, el accuracy del lexicón en inglés es mayor al del lexicón en español, lo cual es lógico puesto que el lexicón en español se ha construido a partir de recursos generados mediante métodos semiautomáticos y por tanto no carentes de errores. La diferencia de *accuracy* entre ambos lexicones va en aumento a lo largo de las capas, pasando de 1,63 puntos porcentuales en la primera capa a 12,27 puntos en la última capa. A pesar de esto, creemos que los valores medidos para el lexicón en español son buenos, si comparamos las capas 5 y 6 (91,75 % en un lexicón de 1779 y 86,09 % en un lexicón de 2849) con los lexicones en español de [Pérez-Rosas, Banea, y Mihalcea2012] (90 % en un lexicón de 1347 lemas y 74 % en uno de 2496 lemas). Más aún, la siguiente capa de nuestro recurso continúa teniendo un mejor nivel de acierto (77,69 %) que el mayor de los lexicones de [Pérez-Rosas, Banea, y Mihalcea2012], con un número de lemas muy superior (6625).

Referencias

[Baccianella, Esuli, y Sebastiani2010] 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.

[Cruz et al.2014] 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*.

[Cruz et al.2012] 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.

[Fagin et al.2004] 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.

[Gonzalez-Agirre, Laparra, y Rigau2012] Gonzalez-Agirre, Aitor, Egoitz Laparra, y German Rigau. 2012. Multilingual central repository version 3.0. En *LREC*, páginas 2525–2529.

[Pérez-Rosas, Banea, y Mihalcea2012] Pérez-Rosas, Verónica, Carmen Banea, y Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. En *LREC*, páginas 3077–3081.

[Strapparava, Valitutti, y Stock2006] 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.

[Turney y Littman2003] 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.

[Vossen1998] Vossen, Piek. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston.