Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 5-12 recibido 1-05-2009; aceptado 5-06-2009

Inducción de un Lexicón de Opinión Orientado al Dominio*

Inference of a Domain-Oriented Opinion Lexicon

Fermín Cruz, José A. Troyano, Javier Ortega, Carlos G. Vallejo
Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
{fcruz,troyano,javierortega,vallejo}@us.es

Resumen: En el presente trabajo mostramos la metodología utilizada para la construcción de un lexicón en inglés compuesto de adjetivos y sus orientaciones semánticas. En lugar de calcular un único valor por término, inducimos un conjunto de valores que codifican la orientación semántica de un término cuando es utilizado en distintos dominios. La generación automática del recurso se ha basado en la construcción de grafos (uno por dominio) a partir de expresiones conjuntivas entre adjetivos observadas en un corpus de *reviews* de productos. Posteriormente aplicamos a dichos grafos una versión modificada del algoritmo PageRank adaptada a la utilización en grafos con aristas con pesos positivos y negativos. El método ha sido evaluado utilizando el recurso Micro-WNOp, obteniendo resultados similares o incluso superiores a los de otros trabajos recientes.
Palabras clave: Minería de opiniones, análisis de sentimiento, generación de recursos, algoritmos basados en grafos

Abstract: In this paper we explain the building of an English lexicon of adjectives and their semantic orientations. Rather than defining the semantic orientation of a term as a single value, we induce many real values representing the semantic orientation of that term being used in different domains. The automatic building of this resource has been based on the construction of graphs from conjunctive expressions between adjectives observed in a review corpus. A modified PageRank algorithm, adapted to be used with graphs with positive and negative edges, was applied to these graphs to obtain the values for the semantic orientation of terms. The method has been evaluated using Micro-WNOp, getting similar results or even better than those reported by some recent papers.
Keywords: Opiniong mining, sentiment analysis, resources generation, graph-based algorithms

1. Introducción

Dentro de la disciplina del Procesamiento del Lenguaje Natural (PLN), los trabajos relacionados con textos de carácter subjetivo en los que se vuelcan opiniones o afectos, se ven obligados a tratar con algunas dificultades adicionales, además de los problemas comunes al resto de las tareas del PLN. En los últimos años ha crecido exponencialmente el interés de la comunidad en este tipo de tareas, conocidas en su conjunto como *opinion mining* o *sentiment analysis* (Pang y Lee, 2008). Tareas como la clasificación de documentos basada en la opinión, la extracción de opiniones o el resumen de textos de opinión requieren la resolución de algunas subtareas clave. Una de estas subtareas es el cálculo de la orientación semántica de una palabra o conjunto de palabras, entendida como un valor numérico, discreto o continuo, que codifica la polaridad (y en su caso la intensidad) de las implicaciones afectivas positivas o negativas de la expresión considerada (usaremos los términos *positividad* y *negatividad* para referirnos a dichas implicaciones). Por ejemplo, la palabra *bueno* tiene una orientación semántica positiva de gran intensidad. Si optásemos

* Parcialmente financiado por el proyecto CICYT HUM2007-66607-C04-04.

por expresarla como es habitual, mediante un valor real entre -1 y 1, podríamos asignarle un valor cercano o igual a 1. De igual forma, seguramente asignaríamos a la palabra *malo* un valor cercano o igual a -1, mientras que *aceptable* tendría un valor positivo, pero de menor magnitud.

Como primer paso hacia la construcción de un sistema de extracción de opiniones de *reviews* de productos (Ding, Liu, y Yu, 2008; Liu, 2006), nos planteamos el problema del cálculo de las orientaciones semánticas. En la mayoría de los trabajos revisados, se construye un lexicón en el que se recogen palabras de opinión (generalmente adjetivos) y se calculan sus orientaciones semánticas mediante distintas aproximaciones (ver sección 2). En todos estos trabajos se establece un valor único, independiente del contexto, para la orientación semántica. Esto es, se parte de una presunción de independencia entre el contexto de aplicación y las implicaciones positivas o negativas de las palabras de opinión. Esta premisa está claramente invalidada por las observaciones realizadas por los distintos investigadores en el área; parece claro que la orientación semántica de un término queda definida, además de por el propio término, por el dominio de aplicación del mismo y por el contexto local en el que se aplica. Por ejemplo, el adjetivo *impredecible* tiene implicaciones negativas cuando se aplica a la respuesta en la conducción de un coche, e implicaciones positivas cuando se aplica al guión de una película o a la trama de un libro. Otros ejemplos de esta dependencia entre la orientación semántica y el dominio de aplicación serían aquellos adjetivos relacionados con el tamaño (una habitación de hotel es deseable que sea cuanto más *grande* mejor, mientras que un portátil demasiado *grande* no será del gusto de los usuarios) o aquellos relacionados con el sonido emitido por un dispositivo (se espera de un electrodoméstico que sea *silencioso*, pero no de un dispositivo GPS con voz).

En el presente trabajo definimos la orientación semántica de un término como un conjunto de valores reales, entre -1 y 1, cada uno de los cuales informa de las implicaciones afectivas negativas o positivas del término en cuestión cuando es aplicado a un dominio determinado. Ya que pretendemos utilizar el lexicón en la construcción de un sistema de extracción de opiniones de *reviews* de productos, hemos utilizado como dominios los distintos tipos de productos a incluir en nuestro sistema (en principio, utilizaremos las categorías definidas en la web de la que hemos extraído el corpus). Otras dimensiones necesarias para contextualizar completamente la utilización de un término (por ejemplo, la característica concreta de un producto sobre el que se está expresando una opinión) han sido obviadas en este trabajo, y serán abordadas en futuras ampliaciones de las ideas aquí descritas.

2. Trabajos relacionados

Han sido muchos los autores que han afrontado el cálculo de la orientación semántica, ya sea con la intención de construir un diccionario de orientaciones semánticas o como componente de sistemas de clasificación de documentos basada en la opinión o de extracción de opiniones. Exponemos a continuación algunos de los trabajos que nos han servido de inspiración.

En Hatzivassiloglou y McKeown (1997) se utilizan las apariciones de conjunciones entre adjetivos de un corpus para generar un grafo. La aparición de dos adjetivos coordinados mediante la conjunción *and* sugiere que ambos comparten la misma orientación semántica. De forma análoga, la coordinación de adjetivos mediante *but* sugiere orientaciones semánticas contrarias. Aplicando un algoritmo de clustering al grafo consiguen generar listas de adjetivos con orientaciones semánticas positivas y negativas. En Hu y Liu (2004) se propone un sistema de extracción y resumen de opiniones de *reviews* de productos. Para la determinación de la polaridad de la orientación semántica de las palabras de opinión (sólo se tienen en cuenta los adjetivos), se propone un algoritmo basado en las relaciones de sinonimia y antonimia de WordNet (Fellbaum, 1998). Nótese que tanto en este trabajo como en el anterior, las orientaciones semánticas calculadas son binarias, esto es, un término es clasificado como positivo o negativo, sin entrar a considerar la intensidad de la orientación.

En Kamps et al. (2004) se propone la construcción de un lexicón de orientaciones semánticas de adjetivos basado también en WordNet. Las distancias del adjetivo a clasificar a las palabras *good* y *bad* utilizando las relaciones de sinonimia son utilizadas para calcular un valor real entre -1 y 1. En Ding, Liu, y Yu (2008) se describe un sistema de

extracción y resumen de opiniones evolucionado del propuesto en Hu y Liu (2004). Para la determinación de la orientación semántica de las opiniones, se parte del conjunto inicial de valores generados en dicho trabajo. Dichos valores son contextualizados y corregidos sobre la marcha al evaluar la orientación semántica de todas las opiniones aparecidas en un documento. Para ello, se utilizan reglas sintácticas basadas en las conjunciones, usando las mismas suposiciones que Hatzivassiloglou y McKeown (1997). Estas reglas se aplican sobre *reviews* del mismo producto. De esta manera se consigue adaptar las orientaciones semánticas iniciales al dominio en que se está trabajando. Por tanto este trabajo tiene en cuenta el contexto a la hora de evaluar la orientación semántica de las palabras de opinión, pero lo hace *on-the-fly*, de manera inseparable al sistema de extracción de opiniones.

En Esuli y Sebastiani (2007), se aplica una versión modificada del algoritmo PageRank (Page et al., 1998) (ver sección 3.3) a una red de adjetivos construida a partir de las relaciones de hiponimia de WordNet. En la versión propuesta del algoritmo, se otorga cierta importancia a priori a algunos nodos de la red, de manera que dichos nodos tenderán a alcanzar una puntuación mayor que los demás tras la ejecución del algoritmo PageRank, y propagarán parte de esa puntuación a sus vecinos. En una primera pasada, los nodos correspondientes a las semillas negativas son resaltados, obteniéndose con ello una puntuación para el algoritmo PageRank modificado que indica la negatividad de los adjetivos de la red. Posteriormente se realiza una segunda pasada del algoritmo resaltando los nodos correspondientes a las semillas positivas, obteniendo con ello una puntuación de positividad para los adjetivos.

El uso de PageRank como evaluador de la orientación semántica en lugar de medidas de distancia a las semillas como las utilizadas en Kamps et al. (2004), tiene en cuenta en la evaluación de cada nodo tanto las puntuaciones del resto de los nodos de la red como la propia topología de la misma. Es por ello la propuesta que nos parece más interesante para construir los lexicones de opinión de cada dominio. En la sección 3.3 proponemos una nueva modificación del algoritmo PageRank adaptada a nuestro problema.

3. Construcción del lexicón

3.1. Corpus de reviews de productos

Como base para la construcción automática del lexicón, hemos utilizado *reviews* de productos de diversa índole extraídos de la web *Epinions*[1]. Este portal ofrece documentos de evaluación de productos (*reviews*) escritos por usuarios, ordenados por categorías que van desde aparatos electrónicos a hoteles, pasando por libros o películas. Para cada uno de los *reviews*, se dispone de una puntuación de 1 a 5 otorgada al producto por el autor del mismo. Esta puntuación puede ser utilizada para realizar experimentos de clasificación de documentos basada en la opinión, aunque éste no es el objetivo del presente trabajo.

Empleamos la herramienta WebHarvest[2] para programar el robot que se encargó de la construcción del corpus. Las expresiones XPath necesarias para la extracción de los distintos apartados de cada *review* fueron obtenidas mediante un *plug-in* del navegador *Firefox*, de nombre Solvent[3]. El resultado obtenido consiste en un corpus de más de 134 millones de palabras, formado por casi 234.000 *reviews* de 84 dominios distintos. En la tabla 1 se muestran algunos de los dominios del corpus, junto con el número de *reviews* contenidos en el dominio y la distribución de los mismos con respecto a los 5 valores posibles de evaluación asignados por los usuarios. Como puede observarse, algunos dominios presentan un fuerte desequilibrio entre *reviews* negativos y positivos, lo que puede reflejarse en la distribución entre términos negativos y positivos contenidos en el lexicón que generemos.

Los textos contenidos en los reviews fueron procesados a través de un *pipeline* de procesamiento lingüístico implementado en UIMA (Ferrucci y Lally, 2004), que consta de tokenizador, *sentence splitter*, etiquetador morfosintáctico y lematizador basados en FreeLing (Atserias et al., 2006), un analizador sintáctico superficial basado en Yamcha (Kudo y Matsumoto, 2003) y un analizador de dependencias basado en Minipar (Lin, 1998).

[1] www.epinions.com
[2] http://web-harvest.sourceforge.net/
[3] http://simile.mit.edu/wiki/Solvent

Dominio	*reviews*	**1**	**2**	**3**	**4**	**5**
Aerolíneas	7.107	23	12,7	12,6	23,7	28,1
Libros	11.520	7,08	7,88	1,28	8,57	75,2
Videojuegos	13.625	3,95	7,24	11,7	32	45,2
Destinos turísticos	32.879	3,22	4,05	7,89	27,7	57,1
Cámaras digitales	13.602	7,8	4,91	7,94	29,9	49,5
Reproductores DVD	8.027	15,5	7,92	9,16	31,1	36,3
Teléfonos móviles	7.868	9,02	9,29	12	34,3	35,3
Hoteles	6.171	22,5	12,5	5,77	13,2	46,1
Películas	14.217	21,3	8,26	5,34	8,5	56,6
Reproductores MP3	7.582	11,5	6,55	10,2	30,9	41
Música	14.638	4,71	6,63	1,41	10,3	77
Impresoras	3.925	23,5	10	9,38	21,4	35,8
Coches	23.180	3,74	3,91	6,79	26,2	59,3
...	...	...	...	...	...	...

Cuadro 1: Distribución de puntuaciones asignadas a *reviews* por dominios en el corpus

3.2. Construcción del grafo

Para cada uno de los 82 dominios con los que trabajamos, generamos un grafo a partir de las construcciones conjuntivas entre adjetivos observadas en los textos de los *reviews*. Son estos adjetivos los que aparecen recogidos en el lexicón. Las aristas que conectan los nodos indican la aparición de ambos términos en alguna construcción conjuntiva. Los pesos de dichas aristas expresan, dependiendo del signo, cierta probabilidad de que ambos nodos tengan orientaciones semánticas iguales u opuestas.

Utilizamos patrones de extracción simples para detectar las relaciones conjuntivas entre adjetivos que aparecen en los textos. Diremos que la relación semántica es *directa* cuando ambos adjetivos compartan la misma orientación semántica, e *inversa* cuando posean orientaciones semánticas opuestas. En los ejemplos 1 y 2 se muestran apariciones en el corpus de relaciones conjuntivas directas e inversas, respectivamente.

(1) **Relación directa**

The camera has a **bright *and* accurate** len.
It is a **marvellous**, really **entertaining** movie.
... **clear** *and* **easy to use** interface.
... **easy to get information**, **user-friendly** interface.

(2) **Relación inversa**

The camera has a **bright** *but* **inaccurate** len.
It is a **entertaining** *but* **typical** film.
The driving is **soft** *and not* **aggresive**.

Las palabras que aparecen en negrita corresponden a los términos que serán incluidos en el lexicón. Como puede verse, además de adjetivos se permiten construcciones formadas por adjetivos de la clase *easy/difficult* seguidos de un infinitivo y las palabras que completan el significado del verbo (el comportamiento sintáctico de estas construcciones y los adjetivos que forman parte de la clase *easy/difficult* se encuentran estudiados en Nanni (1980)). Se excluyen los adjetivos comparativos y superlativos. La relación será directa o indirecta en función de la conjunción utilizada y de la aparición o no de negaciones delante de los adjetivos. La aparición de adverbios está contemplada en el proceso de extracción (excluyendo de nuevo a los comparativos y superlativos), aunque en este trabajo no utilizamos la información de intensidad que podrían proporcionar.

Para el cálculo del peso de las aristas hemos utilizado dos aproximaciones. La primera de ellas consiste en asignar un valor entero calculado como la diferencia entre las relaciones directas e inversas observadas en el dominio para el que estemos construyendo el grafo. De esta manera, un valor será positivo cuando se hayan observado más relaciones directas que indirectas entre los términos de los nodos en cuestión, y negativo en caso contrario. Un valor mayor indica más apariciones de dicha relación en el texto, lo cual debe influir en el cálculo de la orientación semántica de los nodos participantes (ver sección 3.3). También hemos experimentado con una versión normalizada del cálculo de los pesos, con un rango de valores entre -1 y 1; de esta forma, aquellas relaciones que aparecen pocas veces en el texto tienen el mismo impacto en la red que aquellas que aparecen en muchas ocasiones.

3.2.1. Limpieza del grafo

A lo largo de algunos experimentos preliminares observamos la necesidad de corregir algunos aspectos de la construcción del grafo. En primer lugar, existen construcciones conjuntivas de las que se extraerían conclusiones erróneas. Por ejemplo, la expresión *"the good and the bad"* no indica que los adjetivos utilizados compartan orientación semántica. Para corregir la introducción de relaciones claramente erróneas en el grafo, utilizamos las relaciones de sinonimia y antonimia de WordNet. Aquellas relaciones conjuntivas encontradas por los patrones de extracción utilizados, de signo opuesto a una relación de sinonimia o antonimia en WordNet, son ignoradas.

WordNet es también utilizado para filtrar adjetivos inexistentes (incorrecciones ortográficas, onomatopeyas, ...).

Disponemos además de un parámetro en la configuración de la construcción del grafo que permite descartar aquellos nodos que aparezcan en menos de un número determinado de relaciones. De esta forma, podemos obviar aquellos términos que sean anecdóticos para el dominio en el que nos encontremos, y se consigue además eliminar gran parte del ruido introducido por relaciones ocasionalmente mal reconocidas por el patrón de extracción. Por último, eliminamos los bucles, aquellas aristas cuyo nodo inicial y final son el mismo.

3.3. Cálculo de las orientaciones semánticas

Una vez construido el grafo, debemos evaluar las orientaciones semánticas de los nodos del mismo, basándonos en la topología de la red y, posiblemente, en algunos valores conocidos a priori (esto es, algunas palabras semilla). Tal como comentamos en la sección 2, nos basaremos en el algoritmo propuesto en Esuli y Sebastiani (2007). Sin embargo dicho algoritmo no tiene en cuenta la existencia de aristas de pesos negativos en el grafo, por lo que proponemos una versión modificada del mismo.

El algoritmo PageRank original se enuncia de la siguiente manera. Dado un grafo $G = (N, E)$ donde N es un conjunto de nodos y E un conjunto de arcos dirigidos entre dos nodos, se definen en primer lugar dos operaciones $E(n_i)$ y $S(n_i)$ que devuelven, respectivamente, conjuntos con los nodos con arcos que entran o salen del nodo n_i. A partir de estas dos operaciones básicas, se define la puntuación (o PageRank) de un determinado nodo mediante la fórmula 1, donde d es un factor de amortiguación que tiene como objetivo incluir en el modelo la probabilidad de que haya un salto aleatorio de un vértice del grafo a cualquier otro. En el contexto de la navegación en Internet, en el que originalmente fue planteado el algoritmo PageRank, dicho factor representa la probabilidad de que un usuario acceda a una página a través de un enlace situado en la página actual, siendo por tanto $(1 - d)$ la probabilidad de que dicho usuario salte a una página aleatoria no enlazada con la página actual. En la definición original de PageRank se recomienda un valor de 0.85 para el factor d.

$$PR(n_i) = (1 - d) + d \sum_{n_j \in E(n_i)} \frac{1}{|S(n_j)|} PR(n_j) \quad (1)$$

La versión modificada del algoritmo propuesta en Esuli y Sebastiani (2007) (nos referiremos a ella como *PageRank ponderado*) añade un vector real e que codifica la importancia a priori de cada nodo (fórmula 2); p_{ji} es el peso de la arista que va de n_j a n_i. En una primera pasada, se asignan valores a e mayores que cero para los nodos correspondientes a las semillas negativas, y se aplica el algoritmo de PageRank ponderado. Se obtiene con ello una puntuación que indica la negatividad de los adjetivos de la red. Posteriormente se realiza una segunda pasada del algoritmo asignando valores de importancia a priori a las semillas positivas, obteniendo con ello una puntuación de positividad para cada nodo.

$$PR(n_i) = (1-d)e_i + d \sum_{n_j \in E(n_i)} \frac{p_{ji}}{\sum_{n_k \in S(n_j)} p_{jk}} PR(n_j) \quad (2)$$

Para poder realizar este cálculo sobre nuestros grafos, debemos tener en cuenta la existencia en los mismos de aristas negativas, que indican una cierta probabilidad de que dos nodos tengan orientaciones semánticas opuestas. De esta forma, un nodo n_1 que tenga una arista negativa que lo conecte con otro n_2, debe verse inclinado a obtener una orientación semántica contraria a dicho nodo, de forma tanto más pronunciada cuanto mayor sea el valor absoluto de la orientación semántica de n_2. Además, debemos tener en cuenta que las aristas que componen el grafo son no dirigidas.

Para cumplir estas exigencias, planteamos la siguiente modificación del algoritmo, a la que nos referiremos a partir de ahora como *PageRank combinado*. Sea un grafo no dirigido $G = (N, E)$ donde N es un conjunto de nodos y E un conjunto de aristas no dirigidas entre dos nodos. Cada arista de E contiene un valor real asociado o peso, distinto de cero, siendo p_{ij} el peso asociado a la arista que conecta los nodos n_i y n_j. Se define la operación $V(n_i)$, que devuelve el conjunto de nodos vecinos de n_i. Se definen las operaciones $V^+(n_i)$ y $V^-(n_i)$, que devuelven los conjuntos de nodos vecinos a n_i cuyas aristas asociadas tengan un peso mayor o menor que cero,

respectivamente. Definimos el PageRank positivo y negativo de un nodo n_i (fórmula 3), donde los valores de e^+ son mayores que cero para los nodos correspondientes a semillas positivas y cero para el resto de nodos, y los valores de e^- son mayores que cero para los nodos correspondientes a semillas negativas y cero para el resto de nodos.

$$\begin{aligned} PR^+(n_i) &= (1-d)e_i^+ + \\ &+ d\Big(\sum_{n_j \in V^+(n_i)} \frac{p_{ij}}{\sum_{n_k \in V(n_j)} |p_{jk}|} PR^+(n_j) + \\ &+ \sum_{n_j \in V^-(n_i)} \frac{-p_{ij}}{\sum_{n_k \in V(n_j)} |p_{jk}|} PR^-(n_j)\Big) \\ PR^-(n_i) &= (1-d)e_i^- + \\ &+ d\Big(\sum_{n_j \in V^+(n_i)} \frac{p_{ij}}{\sum_{n_k \in V(n_j)} |p_{jk}|} PR^-(n_j) + \\ &+ \sum_{n_j \in V^-(n_i)} \frac{-p_{ij}}{\sum_{n_k \in V(n_j)} |p_{jk}|} PR^+(n_j)\Big) \end{aligned} \tag{3}$$

Proponemos que la suma de los valores de e^+ por un lado y de e^- por otro sea igual al número de nodos del grafo; de esta forma obtenemos valores de PageRank con magnitudes similares a las del algoritmo original de PageRank.

Las semillas positivas y negativas que hemos utilizado son las propuestas por Turney y Littman en Turney y Littman (2003)[4], distribuyendo entre ellas los valores de e^+ y e^- de manera uniforme.

Una vez calculados PR^+ y PR^- para cada nodo, obtenemos la orientación semántica asociada a cada término en cada dominio como la diferencia normalizada entre los valores PR^+ y PR^- del nodo que representa al término en el grafo asociado al dominio (fórmula 4). Los valores así obtenidos están acotados entre -1 y 1.

$$SO(n) = \frac{PR^+(n) - PR^-(n)}{PR^+(n) + PR^-(n)} \tag{4}$$

3.4. Evaluación y análisis

La evaluación de los valores obtenidos para las orientaciones semánticas en cada uno de los dominios planteados supone un problema en sí mismo. En otros trabajos, la evaluación se realiza comparando las orientaciones semánticas obtenidas con algunos recursos generados de manera manual o semiautomática; algunos de estos recursos son SentiWordNet (Esuli y Sebastiani, 2006) , General Inquirer (Stone, 1966) o Micro-WNOp (Cerini et al., 2007). Dichos recursos contienen valores para la orientación semántica independientes del dominio: ¿Cómo distinguir entre las divergencias entre los valores de orientación semántica fruto de imperfecciones de nuestro método y aquellas inherentes a peculiaridades del dominio en el que nos encontramos?

Para evitar este problema, construimos un grafo global a partir del corpus completo para generar un lexicón independiente del dominio, comparando los valores obtenidos con los del recurso Micro-WNOp. De esta manera pretendemos medir la bondad del método propuesto, así como la influencia de los parámetros que participan en el mismo. Usamos como *gold standard* el recurso Micro-WNOp (Cerini et al., 2007), que consiste en una muestra de aproximadamente 1100 synsets de WordNet, a cada uno de los cuales les fueron asignados manualmente tres valores reales entre 0 y 1 indicando la positividad, negatividad y neutralidad del mismo. De esos 1100 synsets, nos quedamos con los 284 correspondientes a adjetivos. Dado que en nuestros experimentos no hemos realizado desambiguación de significados, utilizamos el promedio de la orientación semántica de aquellos synsets de adjetivos con un mismo lexema asociado. Con esto obtenemos un *ranking* de 433 adjetivos, que compararemos con el *ranking* obtenido de la aplicación de nuestro método automático de construcción del lexicón al corpus completo. En cada uno de los experimentos, usamos como *gold standard* sólo aquellos adjetivos de Micro-WNOp encontrados en nuestro lexicón.

Para comparar ambos *rankings* hemos utilizado la distancia τ de *Kendall* (τ_p) (Fagin et al., 2004), que mide la similitud entre un *ranking* modelo o *gold standard* y otro *ranking* candidato. La distancia se calcula mediante la siguiente fórmula:

$$\tau_p = \frac{n_d + p * n_u}{Z} \tag{5}$$

siendo n_d el número de pares discordantes (pares ordenados en el *gold standard* y no ordenados en el *ranking* candidato), n_u el

[4]Semillas positivas: good, nice, excellent, positive, fortunate, correct, superior. Semillas negativas: bad, nasty, poor, negative, unfortunate, wrong, inferior.

número de pares ordenados en el *gold standard* e iguales en el otro ranking, y p un factor de penalización asignado a cada uno de estos pares. Z el número total de pares ordenados en el *gold standard*. Usamos un valor de $\frac{1}{2}$ para el factor de penalización. Cuanto más cercano a cero sea el resultado de este cálculo, más parecido es el *ranking* obtenido de nuestro lexicón al *ranking* aportado por Micro-WNOp.

Hemos experimentado con la versión normalizada y sin normalizar del cálculo de los pesos de las aristas del grafo. Para comprobar la mejora obtenida por la aplicación del algoritmo PageRank combinado, hemos reproducido cada uno de los experimentos descartando las aristas negativas (aquellas asociadas a construcciones en las que interviene la conjunción *but*) y utilizando el algoritmo PageRank ponderado. En la tabla 2 se muestran los valores obtenidos en Esuli y Sebastiani (2007) ($Esuli^+$ y $Esuli^-$) junto a los resultados obtenidos en nuestros experimentos (e_1 a e_4). El mejor resultado se obtiene usando el algoritmo PageRank combinado propuesto, junto con la versión normalizada de los pesos. Este resultado supone una mejora del 4,26 % con respecto a la utilización del algoritmo PageRank ponderado. Todos los valores obtenidos mejoran los de Esuli y Sebastiani (2007), trabajo en el que se utilizó un método de evaluación similar al que hemos utilizado, si bien ellos reportaron valores de $\tau_{\frac{1}{2}}$ para sendos *rankings* de positividad y negatividad por separado ($Esuli^+$ y $Esuli^-$ en la tabla de resultados). Téngase en cuenta que en dicha evaluación se utilizó un subconjunto distinto del recurso *Micro-WNOp* al utilizado en nuestros experimentos, por lo que los resultados no son directamente comparables.

	Grafo	**PageRank**	**Pesos**	$\tau_{\frac{1}{2}}$
$Esuli^+$	WordNet	Ponderado	-	0,325
$Esuli^-$	WordNet	Ponderado	-	0,284
e_1	Corpus	Ponderado	Sin norm.	0,239
e_2	Corpus	Ponderado	Norm.	0,235
e_3	Corpus	Combinado	Sin norm.	0,230
e_4	Corpus	Combinado	Norm.	**0,225**

Cuadro 2: Resultados de los experimentos

Para confirmar la hipótesis inicial de este trabajo, en la que consideramos que la orientación semántica de una palabra de opinión es dependiente del dominio en el que se aplica, mostramos algunos ejemplos extraídos del lexicón (tabla 3), que se corresponden con los ejemplos utilizados en la sección 1. Los términos *big*, *small* y *tiny*, tienen connotaciones positivas o negativas dependiendo del dominio de aplicación. Por ejemplo, aplicado a monitores u hoteles, *big* tiene connotaciones positivas (*The room was really big*, *This monitor has a brilliant, big screen.*). Al referirnos a un ordenador portátil, la orientación semántica no está tan clara: es deseable que un portátil sea pequeño en su conjunto (*a cute and tiny laptop*), aunque una pantalla grande, por ejemplo, es una buena característica. Estas diferencias quedan reflejadas en el valor obtenido para la orientación semántica en cada uno de los dominios.

Otros ejemplos de términos cuya orientación semántica depende fuertemente del contexto son *hard to hear* (referido a un lavavajillas: *"It is really pretty quiet, and hard to hear, even when we are standing in kitchen"*; referido a un GPS con voz: *"The speaker of this GPS device is hard to hear."*) y *predictable/unpredictable* (*"The script is unpredictable, funny and shocking all in one!"*; *"unpredictable car behavior"*).

Dominio:	Hoteles	Monitores	Portátiles
big	0,2640	0,1880	-0,5800
small	-0,2400	-0,4570	0,3910
tiny	-0,3120	-	0,7340
Dominio:	Lavavajillas	Móviles	GPS
hard to hear	0,7210	-0,2910	-0,5220
Dominio:	Películas	Coches	Aerolíneas
predictable	-0,262	0,2	0,3260
unpredictable	0,128	-0,559	-0,792

Cuadro 3: Orientación semántica de algunos términos del lexicón

4. Conclusiones

En el presente trabajo hemos descrito la construcción de un lexicón de adjetivos y sus orientaciones semánticas dependientes del dominio. El método utilizado está basado en la construcción de grafos a partir de las relaciones conjuntivas entre adjetivos observadas en un corpus de *reviews* de productos, y la posterior aplicación de un algoritmo de *ranking* de nodos. La evaluación global del método ha proporcionado buenos resultados. Además, hemos mostrado como ejemplo algunos valores de orientación semántica obtenidos para distintos dominios, confirmándose que (1) la orientación semántica de un

término es dependiente del contexto de aplicación del mismo, y (2) el método de inducción de la orientación semántica orientada a dominio funciona correctamente. El algoritmo de *ranking* propuesto mejora significativamente los resultados obtenidos mediante el algoritmo propuesto en Esuli y Sebastiani (2007).

Actualmente, estamos trabajando en ampliar la contextualización de la orientación semántica en el lexicón generado, añadiendo información relativa a la característica concreta del producto sobre la que se está expresando la opinión. De esta manera, la orientación semántica de un término tendrá valores distintos cuando dicho término sea aplicado a distintas características de un mismo producto (por ejemplo, a la pantalla o al teclado de un portátil). Dicha distinción es fundamental para la utilización del lexicón en el sistema de extracción de opiniones.

Bibliografía

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, y Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May. ELRA. http://www.lsi.upc.edu/ñlp/freeling.

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, y G. Gandini, 2007. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics.*, capítulo Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

Ding, Xiaowen, Bing Liu, y Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. En *WSDM '08: Proceedings of the international conference on Web search and web data mining*, páginas 231–240, New York, NY, USA. ACM.

Esuli, Andrea y Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. En *Proceedings of Language Resources and Evaluation (LREC)*.

Esuli, Andrea y Fabrizio Sebastiani. 2007. Pageranking wordnet synsets: An application to opinion mining. En *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, páginas 424–431. Association for Computational Linguistics.

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.

Ferrucci, David y Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

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

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 168–177, New York, NY, USA. ACM.

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

Kudo, Taku y Yuji Matsumoto. 2003. Fast methods for kernel-based text analysis. En *ACL*, páginas 24–31.

Lin, Dekang. 1998. Dependency-based evaluation of minipar. En *Proc. Workshop on the Evaluation of Parsing Systems*, Granada.

Liu, Bing, 2006. *Web data mining; Exploring hyperlinks, contents, and usage data*, capítulo 11: Opinion Mining. Springer.

Nanni, Deborah L. 1980. On the surface syntax of constructions with easy-type adjectives. *Language*, 56(3):568–581.

Page, Lawrence, Sergey Brin, Rajeev Motwani, y Terry Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Informe técnico, Stanford Digital Library Technologies Project.

Pang, Bo y Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

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

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.