Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 95-98 recibido 17-11-2015 revisado 08-01-2016 aceptado 29-01-2016

Generación de recursos para análisis de opiniones en español

Generation of resources for Sentiment Analysis in Spanish

M. Dolores Molina González
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
mdmolina@ujaen.es

Resumen: Tesis doctoral en Informática realizada por Mª Dolores Molina en la Universidad de Jaén (UJA) bajo la dirección de la doctora Mª Teresa Martín Valdivia (UJA). El acto de defensa de la tesis tuvo lugar en Jaén el 28 de noviembre de 2014 ante el tribunal formado por los doctores Luis Alfonso Ureña (UJA), Rafael Muñoz (UA) y Fermín Cruz (U. Sevilla). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: Clasificación de polaridad, corpus de opiniones y lexicón en español.

Abstract: Ph.D. Thesis in Computer Science written by Mª Dolores Molina at the University of Jaén (UJA), under the supervision of Dr. Mª Teresa Martín (UJA). The author was examined on 28th of November 2014 at the University of Jaen by a commission composed by the doctors Luis Alfonso Ureña (UJA), Rafael Muñoz (UA) and Fermín Cruz (U. Sevilla). The unanimously awarded grade was Excellent Cum Laude.
Keywords: polarity classification, Spanish reviews corpus and lexicon.

1 *Introducción*

Esta tesis está centrada en el Análisis de Opiniones (AO) en español, debido a su creciente interés en los últimos años provocado por varios factores, siendo uno de ellos el consumo de datos online, hecho casi imprescindible y rutinario para la toma de decisiones a nivel individual o colectivo.

Aunque son muchas las tareas estudiadas en AO, una de las más consolidadas es la clasificación de la polaridad. Para esta tarea es necesario el uso de recursos léxicos normalmente dependientes del idioma para determinar la polaridad de las palabras.

La mayor parte de los trabajos en AO tratan con documentos escritos en inglés a pesar de que cada vez la cantidad de información subjetiva que publican los usuarios de internet en su propio idioma es mayor. Es por esta razón, que la generación y uso de recursos propios en el idioma de los documentos a tratar se esté convirtiendo en un tema crucial para realizar la clasificación de opiniones mediante orientación semántica.

El idioma español, según *Internet Word State Rank*[1], es el tercer idioma más usado por los usuarios de internet después del chino y el inglés (idioma más usado), así pues, está justificada la generación de recursos lingüísticos nuevos en nuestro idioma para seguir progresando en AO.

La principal contribución de esta tesis es la generación de un lexicón de palabras de opinión independiente del dominio, otros lexicones dependientes del dominio y la generación de un corpus nuevo de opiniones en el dominio turístico, además de la realización de experimentos que certifican la validez de dichos recursos en la clasificación de polaridad de documentos escritos en español.

2 *Organización de la tesis*

La tesis se organiza estructuralmente en cinco capítulos que describen, respectivamente, la justificación y objetivos pretendidos con la ejecución de este trabajo de investigación, los recursos lingüísticos para AO más usados en la clasificación de la polaridad y algunos métodos

[1] http://www.internetworldstats.com/stats7.htm

ISSN 1135-5948

para la generación de los mismos, la información resumida de los resultados obtenidos más interesantes recogidos en las distintas publicaciones, la discusión general de todos los datos en su conjunto y, finalmente, los comentarios sobre futuros trabajos que quedan abiertos en la presente tesis.

El capítulo 1 introduce el interés por el Análisis de Opiniones centrándose en dos técnicas de clasificación de polaridad, como son la aproximación basada en aprendizaje automático o supervisado y la basada en orientación semántica o no supervisada. Tras analizar las ventajas y los inconvenientes de ambas técnicas se explica la decisión tomada para enfocar nuestro interés en la orientación semántica y se expone los objetivos pretendidos con la ejecución de la tesis.

El capítulo 2 ofrece una breve panorámica de recursos lingüísticos existentes, siendo el uso de dichos recursos en el Procesamiento de Lenguaje Natural (PLN) requisito indispensable para la construcción de los clasificadores de polaridad de opiniones. Así se puede ver en este capítulo corpora, cuya definición podría ser la recopilación de textos representativos de una lengua disponible en formato electrónico y lexicones que pueden ser tan sencillos como los consistentes en listas de palabras separadas según su polaridad o tan complejos como las más extensa colección de palabras o n-gramas que llevan asociadas una serie de características que facilitará el conocimiento gramatical y sentimental de dichas palabras o n-gramas. Entre los recursos que se describen se encuentran ejemplos de corpora escritos en inglés, corpora escritos en idiomas distintos del inglés, corpora escritos en el idioma destino de nuestra investigación, los lexicones más usados en la bibliografía, siendo SentiWordNet base de muchos de ellos y lexicones para AO en español. Además en este capítulo se presentan algunos métodos encontrados en el estado del arte para la generación de recursos léxicos adaptados a un dominio.

El capítulo 3 muestra un resumen de las distintas propuestas que se recogen en la memoria de la tesis, que fueron origen de publicaciones y presenta una breve discusión sobre los resultados obtenidos para cada una de ellas. La primera propuesta fue la comparación de la clasificación de polaridad, según el enfoque supervisado y no supervisado para un corpus comparable en inglés y español. Ante las conclusiones obtenidas nuestra segunda propuesta fue la generación de lexicones para realizar la clasificación de polaridad basada en orientación semántica de documentos escritos en español. Se generan dos tipos de lexicones, uno de propósito general y otros adaptados a dominios específicos. La tercera propuesta fue la generación de corpus escritos en español en un dominio distinto de los que ya existían en ese momento para darle más cobertura y experimentación a los lexicones generados, y por último, en la cuarta propuesta se ha querido avanzar en el campo del bilingüismo, usando recursos en inglés para mejorar la clasificación de polaridad para un corpus en español.

El capítulo 4 resume la línea de trabajo totalmente encadenada que comienza con una visión general de la clasificación de polaridad supervisada y no supervisada para corpora comparables en dos idiomas, el inglés y el español. Seguidamente, se centra en la clasificación basada en la aproximación no supervisada sobre corpus en español usando dos métodos distintos, comprobándose que los resultados usando el método basado en lexicón son equiparables a los obtenidos con el basado en grafos, método más complicado y tedioso de implementar. Este hecho es el punto de partida para la creación de recursos lingüísticos en español para la clasificación de la polaridad en nuestro idioma destino, siendo en este capítulo donde se muestran los distintos recursos lingüísticos generados que son el principal aporte que ha suscitado la realización de esta tesis.

Finalmente, el capítulo 5 plantea futuros trabajos ante la necesidad de acotar distancias entre la clasificación de polaridad basada en la aproximación supervisada y no supervisada.

3 Contribuciones

En esta sección se describe brevemente los recursos lingüísticos generados para clasificación de polaridad de opiniones junto con algunas experimentaciones realizadas.

Los distintos tipos de recursos necesarios son los lexicones y los corpora. Así pues, en el primer experimento se generó un lexicón llamado SOL (Spanish Opinion Lexicon) traducido automáticamente del lexicón en inglés de Bing Liu[2]. Con este lexicón se comprobó que los resultados en la clasificación de polaridad eran comparables a los

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

conseguidos con otros métodos más tediosos y complicados (Martínez-Cámara et al., 2013). Ello fue lo que motivó la mejora manualmente de este primer recurso y se creó iSOL (improved SOL) (Molina-González et al., 2013). Dicha mejora fue fruto de un trabajo duro y arduo, con una revisión exhaustiva para adaptar por ejemplo los adjetivos ingleses a las posibles 4 formas españolas según su número y género. Además, se incluyeron también palabras que aunque no están reconocidas en la Real Academia Española son usadas con frecuencia en un entorno de comunicación social. Finalmente, iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas, teniendo por lo tanto 8.135 palabras de opinión.

Los lexicones anteriormente comentados son de propósito general, sin embargo, el AO es una tarea con un cierto grado de interrelación con el dominio tratado. Por consecuencia, surge la idea de generar lexicones adaptados a diferentes dominios. Para la generación de nuevas listas de palabras de opinión se siguió el enfoque basado en corpus. El elemento clave del enfoque basado en corpus es el uso de una colección de documentos etiquetados según su polaridad de donde extraer información.

El primer lexicón adaptado al dominio fue eSOL (enriched SOL) generado a partir del corpus MuchoCine[3] y siguiendo el mismo el supuesto de que una palabra debe ser positiva (o negativa) si aparece en muchos documentos positivos (o negativos), se calculó la frecuencia de las palabras en cada clase de documentos (positivos y negativos). La elección del grupo de palabras para añadir a cada una de las listas (positiva y negativa) fue manual y subjetiva. Dichas palabras fueron añadidas al lexicón de propósito general iSOL. Los resultados obtenidos con estos primeros recursos sobre el corpus MuchoCine comparados con el existente SEL[4] se muestran en la tabla 1.

Lexicón	Macro-Precisión	Macro-F1	Exactitud
SOL	56,15%	56,07%	56,23%
iSOL	62,22%	61,84%	61,83%
eSOL	63,93%	63,33%	63,16%
SEL	52,56%	52,18%	52,64%

Tabla 1. Resultados en la clasificación binaria de corpus MC usando SOL, iSOL, eSOL y SEL

Siguiendo con la generación de lexicones se quiso ampliar a más dominios, por lo que se recurrió al corpus español SFU[5] escrito en español compuesto de 50 opiniones para 8 dominios. Parte del corpus fue usado para la generación de lexicones y el resto para realizar la clasificación de polaridad. Siguiendo el mismo supuesto anteriormente comentado, se calculó la frecuencia de las palabras en cada clase de documentos (positivos y negativos). Para esta generación de nuevos lexicones, la frecuencia hallada siguió dos métodos, a los que se llamaron 'local' y 'global'. El método 'local' contaría la frecuencia absoluta de las palabras por clase (opiniones positivas y negativas) y el método 'global' contaría la aparición de las palabras en cada opinión y en caso de aparecer, independientemente del número de veces que ello ocurra, solo se cuenta como 1. Indistintamente de la metodología empleada, las palabras a ser añadidas al lexicón de propósito general iSOL, debían cumplir el siguiente algoritmo:

Siendo f^+ frecuencia *palabra* (clase positiva)
Siendo f^- frecuencia *palabra* (clase negativa)

Si (f^-=0 AND $f^+\geq 3$) OR ($f^+/f^-\geq 3$)
entonces lista(positiva) ←*palabra*
Si (f^+=0 AND $f^-\geq 3$) OR ($f^-/f^+\geq 3$)
entonces lista(negativa) ←*palabra*

Los nuevos lexicones fueron llamados eSOL*domain*Global y eSOL*domain*Local, siendo *domain* cada uno de los 8 dominios existentes en el corpus SFU. Los resultados obtenidos en la clasificación de polaridad con los lexicones adaptados al dominio generalmente superan los obtenidos con el lexicón de propósito general y pueden ser vistos en Molina-González et al. (2014b).

Una vez generados diversos tipos de lexicones y debido a la dificultad de encontrar corpora distintos a los usados con los que seguir trabajando, se propuso avanzar con la generación de nuevos corpora para el español. Intentando ampliar el número de dominios existentes en la bibliografía, se generó un corpus de opiniones sobre hoteles. Después de estudiar varios portales web, la elección final para extraer las opiniones fue de TripAdvisor[6].

[3] http://www.lsi.us.es/~fermin/index.php/Datasets

[4] http://www.cic.ipn.mx/~sidorov/#SEL

[5] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html

[6] http://www.tripadvisor.es

Se seleccionaron solo hoteles andaluces. Por cada provincia se eligieron 10 hoteles, siendo 5 de ellos de valoración muy alta y los otros 5 con las peores valoraciones, para obtener las mínimas opiniones neutras en el corpus. Todos los hoteles seleccionados debían tener al menos 20 opiniones escritas en español en los últimos años. Finalmente, se obtuvieron 1.816 opiniones. Este corpus se llamó COAH (Corpus of Opinions about Andalusian Hotels) y está disponible libremente[7].

El nuevo corpus dio opción a la generación de otro lexicón adaptado al dominio 'hoteles', llegando con los experimentos realizados a la conclusión de que la inclusión de palabras al lexicón de propósito general iSOL hace mejorar la clasificación de polaridad, como puede verse en Molina-González et al. (2014a).

4 Conclusiones y futuros trabajos

En esta tesis se revela la importancia de disponer de recursos lingüísticos para la clasificación de polaridad en documentos escritos en español. Para tener oportunidad de seguir avanzando en el análisis de opiniones, en esta tesis, se desarrollan diversos recursos siguiendo varias metodologías, algunas ya implementadas para el idioma inglés. Dichas metodologías han permitido acortar distancias entre la clasificación de polaridad en español usando aproximación supervisada y la no supervisada.

Como fututo trabajo se pretende mejorar más el sistema de clasificación usando el lexicón iSOL. Debido a que las palabras no tienen la misma carga de subjetividad positiva y negativa, apoyándose en algún recurso ya existente, en inglés o español, se dará conocimiento a las palabras de opinión contenidas en iSOL.

Para concluir cabe decir que los métodos implementados en esta tesis para el español podrían ser extensibles a otros idiomas con características gramaticales similares y así aumentar los recursos lingüísticos tan necesarios en todos los idiomas para poder realizar el Análisis de Opiniones.

Bibliografía

Mártinez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina González y L. A. Ureña-López. 2013. Bilingual experiments on an opinión comparable corpus. En *Proceeding of 4th Workshop on Computational Approaches to Subjectivity, Sentiment and social Media Analysis,* páginas 87-93, Atlanta, Georgia, USA.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González y J. M. Perea-Ortega. 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, páginas 538-554.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, páginas 7250-7257.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2014a. Cross-Domain semantic analisys using Spanish opinionated words. En *Proceedings of the 19th International Conference on Natural Language Processing and Information Systems,* páginas 214–219.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2014b. A Spanish semantic orientation approach to domain Adaptation for polarity classification. *Information Processing and Management*, páginas 520-531.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2015. eSOLHotel: Generación de un lexicon de opinion en español adapatado al dominio turístico. *Procesamiento del Lenguaje Natural,* 54:21-28.

[7] http://sinai.ujaen.es/coah