Métodos semiautomáticos de generación de recursos de Opinion Mining para el gallego a partir del portugués y el español

Paulo Malvar Fernández[1] y José Ramom Pichel Campos[2]

[1,2] Departamento de Ingeniería Lingüística, imaxin|**software**
Salgueiriños de Abaixo Nº11 L6
Santiago de Compostela

[1]paulomalvar@imaxin.com
Tlfn: 617753454
Fax: 981554988

[2]jramompichel@imaxin.com
Tlfn: 981554068
Fax: 981554988

Abstract

Despite the growth experienced in recent years in the field of Natural Language Processing (NLP), researchers and developers who intend to carry out developments in languages other than English still face the problem that needed resources and applications are limited or even nonexistent. In this paper we propose a semiautomatic method to generate resources to build an Opinion Mining application for Galician by reusing resources originally compiled for Spanish and by taking advantage of Portuguese as a bridge language that, due to its close-relation with Galician, ensures a high rate of lexical transfer.

Keywords: Opinion Mining, Semiautomatic Generation, Resources, Spanish, Galician, Portuguese

Resumen

A pesar del crecimiento experimentado en los últimos años en el ámbito del Procesamiento del Lenguaje Natural (PLN), investigadores y desarrolladores que pretenden llevar a cabo desarrollos para lenguas diferentes del inglés aún se encuentran con el problema de que los recursos y aplicaciones necesarios son escasos, cuando no inexistentes. En este trabajo proponemos una metodología semiautomática para generar recursos para una aplicación de Opinion Mining para el gallego aprovechando recursos del español y utilizando el portugués como lengua-puente que, por su proximidad, asegura para una alta tasa de transferencia léxica con relación al gallego.

Palabras Clave: Opinion Mining, Generación Semiautomática, Recursos, Español, Gallego, Portugués.

1 Introducción

En este apartado introductorio definiremos, en primer lugar, la rama del Procesamiento del Lenguaje Natural (PLN) dentro de la cual se encuadra este artículo. Discutiremos, por otro lado, aspectos relacionados con la carencia de recursos para el desarrollo de aplicaciones de PLN para lenguas diferentes del inlgés, y en especial para lenguas minorizadas, como es el caso del gallego.

1.1 Opinion Mining

Opinion Mining, también conocida como Sentiment Analysis, es una rama del PLN de muy reciente aparición. Así, los primeros estudios en Opinion Mining se remontan a finales de los años 90 y comienzos de la primera década del siglo XIX, [9], [15] y [18].

Las investigaciones llevadas a cabo dentro del ámbito del Opinion Mining se dedican a la extracción automatizada de información subjetiva (esto es, opiniones, sentimientos, juicios de valor, etc) contenida dentro de documentos no estructurados en formato digital: reseñas, artículos de opinión, críticas de hoteles, restaurantes, etc.

A pesar de ser una rama del PLN aún en consolidación, Opinion Mining ha atraído la atención de muchos investigadores en los últimos años gracias al enorme potencial de negocio que se ha observado en sus aplicaciones: vigilancia y protección de marcas, vigilancia del grado de aceptación y satisfacción de productos y/o servicios, gestión de clientes, etc.

Todas estas aplicaciones cobran especial relevancia si tenemos en cuenta el crecimiento exponencial de la información subjetiva experimentado dentro de lo que hoy conocemos como Web 2.0. Foros, blogs, portales de venta de productos y servicios y redes sociales, han cambiado la manera en que se crea y difunde información en Internet.

Empresas, entes públicos, personajes de relevancia dentro de la vida pública, etc necesitan conocer los puntos fuertes y débiles de sus productos y servicios (así como los de la competencia), monitorizar su reputación, o simplemente obtener una valiosa retroalimentación por parte de sus clientes, seguidores y también detractores.

Desde el ámbito del Opinion Mining se pretende, por lo tanto, crear modelos capaces de extraer y clasificar automáticamente toda esa información subjetiva para, de esta manera, ofrecer soluciones que cubran dichas necesidades.

1.2 Falta de recursos

El crecimiento experimentado en los últimos años en el ámbito del Procesamiento del Lenguaje Natural (PLN), no sólo desde el punto de vista de investigación académica, sino también desde el punto de vista de desarrollo de aplicaciones y soluciones comerciales, se apoya en el trabajo realizado durante los últimos 60 años, desde que se comenzaron a desarrollar los primeros traductores automáticos en el contexto de la Guerra Fría.

Es precisamente por esta investigación y desarrollo previo ya realizado durante décadas que hoy en día es posible contar con numerosos y diversos corpora, así como innumerables herramientas, cuya inmensa mayoría se encuentran disponibles sólo para el idioma inglés.

Sin embargo, el problema con se encuentran investigadores y desarrolladores que pretenden llevar a cabo desarrollos de PLN para lenguas diferentes del inglés es que este tipo de recursos y aplicaciones son escasos, cuando no inexistentes. Así por ejemplo, si dentro del ámbito del Opinion Mining en inglés es posible contar con diversos corpora; en español este tipo de recursos es escaso. De hecho después de una intensa búsqueda[1] de recursos realizada previamente al inicio de nuestro proyecto, sólo nos fue posible encontrar un corpus de críticas de cine llamado "Spanish Movie Reviews", que Fermín L. Cruz Mata ha compilado a partir de la página web muchocine.net[2].

Si en español fue posible encontrar un único corpus que se ajustaba a nuestras necesidades de desarrollo, cuando iniciamos el desarrollo de una solución similar para

[1] Durante el proceso de revisión de este artículo se nos informó de la existencia de otro corpus en español especialmente diseñado para el desarrollo de aplicaciones de Opinion Mining. El corpus en cuestión SFU Spanish Review Corpus, desarrollado por Maite Taboada (entre otros) dentro de un proyecto de investigación de Simon Fraiser University. El corpus en cuestión está disponible desde http://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html. Durante el desarrollo de nuestro proyecto de investigación y desarrollo no teníamos constancia de la existencia de este corpus y, por lo tanto, no lo hemos utilizado.

[2] Este corpus puede ser descargado desde http://www.lsi.us.es/~fermin/corpusCine.zip

gallego nos encontramos con que hoy en día este tipo de recurso para gallego es inexistente.

Esta total ausencia de recursos para gallego, fue la razón por la cual nos vimos obligados a encontrar una solución que nos permitiese la generación rápida de este tipo de recursos para gallego.

La decisión más pragmáticamente viable que encontramos fue el aprovechamiento de soluciones y técnicas desarrolladas para otros ámbitos del PLN para generar semiautomáticamente este tipo de recursos para gallego. De esta manera, ideamos una solución para aprovechar la especialmente estrecha relación entre gallego y portugués así como la relación proximidad entre gallego y español.

2 Investigaciones Previas

Una demostración empírica de la abismal distancia que existe en términos de investigación y desarrollo de soluciones de Opinion Mining para inglés y otras lenguas como el español y el portugués, es la amplia diferencia en número existente dentro de la literatura especializada.

Así en el caso del inglés, ya en 1997 [9] desarrollaron una investigación para predecir la orientación semántica de textos teniendo exclusivamente en cuenta los adjetivos presentes en dichos textos. Este tipo de investigaciones se han hecho cada vez más prolíficas para esta lengua como demuestran los números trabajos publicados en años más recientes, por ejemplo, [12], [7], [15], [20], [2] e [3].

En contrapartida, para lenguas como español y portugués los trabajos de investigación dentro del ámbito del Opinion Mining de los cuales tenemos referencia son escasos: [10] y [14], para el español, y [17], [4] y [16], para el portugués.

2.1 Recursos Disponibles

Al igual que ocurre con los trabajos de investigación realizados para el inglés, existen actualmente numerosos vocabularios y corpora disponibles para descarga desarrollados para esta lengua[3].

Además de estos recursos ya precompilados para inglés resulta fácil encontrar en Internet multitud de páginas que por su naturaleza son explotables desde el punto de vista del Opinion Mining. En esta categoría se encuentra epinions.com o amazon.com que almacenan las opiniones de millones de usuarios acerca de muy diversos productos (electrónica, automovilismo, literatura, cine, etc).

Para el español, como ya hemos indicado, sólo nos fue posible encontrar el corpus "Spanish Movie Reviews".

Para portugués, sin embargo, nos resultó imposible localizar ningún corpus precompilado. Téngase en cuenta que los trabajos más prometedores en lengua portuguesa, [17], [4] y [16], se basan en un análisis sintáctico y de detección automática

[3] Dentro del ámbito de un proyecto llamado "Web Mining, Text Mining, and Sentiment Analysis", Bing Liu ha puesto a disposición de la comunidad un corpus de críticas de usuarios de tiendas on-line que puede ser descargado desde http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. John Blitzer también ha puesto a disposición de la comunidad un corpus llamado "Multi-Domain Sentiment Dataset" que puede ser descargado desde http://www.cs.jhu.edu/~mdredze/datasets/sentiment/. Finalmente, mencionar también la contribución de Bo Pang y Lillian Lee, que han puesto a disposición de la comunidad un corpus de críticas de cine, que puede ser descargado desde http://www.cs.cornell.edu/people/pabo/movie-review-data/. Dentro del ámbito de los vocabularios o lexicones etiquetados con información sobre la orientación sentimental, uno de los más famosos y utilizados es General Inquirer, que puede ser descargado desde http://www.wjh.harvard.edu/~inquirer/. En los últimos años, y dada la gran popularidad que están adquiriendo los estudios dentro del ámbito del Opinion Mining, se han creado otros valiosos recursos léxicos, como por ejemplo, SentiWordNet, que aporta información acerca de la positividad, negatividad o neutralidad que vehiculan las unidades léxicas.

de patrones. Para el desarrollo de este tipo de técnicas de Pattern Matching, no es necesario contar con amplios corpora anotados con información semántico-sentimental. Conjeturamos, por lo tanto, que ésta es la razón a la cual se debe la ausencia de recursos precompilados para portugués.

En cualquier caso, al igual que en inglés, para español y portugués es posible encontrar páginas de internet de donde extraer opiniones de manera más o menos automatizada. En los últimos años se han ido acumulando grandes cantidades de información relevante para el ámbito del Opinion Mining. Así, existen páginas, como booking.com o Google Places, por sólo citar dos ejemplos, donde clientes español y portugués hablantes han ido dejando su opinión acerca de productos o servicios que podrían ser explotadas para el desarrollo de recursos y aplicaciones de Opinion Mining.

Sin embargo, el caso del gallego es muy diferente. No siendo este el espacio adecuado para discutir la problemática sociolingüística del gallego, nos limitaremos indicar que, sea por la situación de minorización que el gallego padece en comparación con el español, el inglés y el portugués; sea por que el bajo número de usuarios activos en internet que desarrollan su actividad on-line en esta lengua; la realidad es que el tipo de recursos disponibles para desarrollar una aplicación de Opinion Mining para el gallego que se base en técnicas de Machine Learning es insignificante e, incluso nos atreveríamos a afirmar, inexistente.

Siendo este el caso para el gallego, sería justo preguntar: "¿Por qué optar por una solución basada en Machine Learning para la cual no existen recursos? ¿Por qué no optar en su defecto por una basada en Pattern Matching, que resultaría factible de desarrollar

siendo simplemente necesario contar con lingüistas y/o filólogos que realizasen un trabajo de desarrollo de patrones relevantes para la detección de opiniones?"

Las razón es muy simple. Tras las pruebas iniciales que llevamos a cabo para el desarrollo de una solución basada en Pattern Matching, llegamos a la conclusión de que, en el ámbito de nuestro proyecto, el desarrollo de una solución comercial basada en Pattern Matching, requeriría unos tiempos de desarrollo y una inversión de recursos económicos que nuestra empresa no se podía permitir. Como PYME que somos, en imaxin|**software** nos enfrentamos con fuertes restricciones de tiempo y dinero que poder invertir en desarrollos de I+D. Al no reportar a corto plazo ni los resultados ni los beneficios económicos deseados, un desvío en los tiempos de desarrollo de una solución de este tipo implicaría la cancelación del proyecto.

Por el contrario, si bien es cierto que las soluciones basadas en Machine Learning precisan de recursos enriquecidos específicamente con información pertinente para alimentar los algoritmos de aprendizaje automatizado, es bien sabido que las soluciones basadas en Machine Learning ofrecen resultados aceptables en un muy corto espacio de tiempo.

3 Metodología propuesta

Para el desarrollo de una aplicación de Opinion Mining basada en Machine Learning para el gallego determinamos que acometeríamos el desarrollo de este proyecto utilizando un corpus etiquetado con información de semántico-sentimental y un vocabulario controlado en el cual también se incluyese información de ese mismo tipo acerca de los adjetivos, sustantivos, verbos y adverbios en él contenidos.

Para el desarrollo de la aplicación análoga para español contamos con:

a) Un corpus, "Spanish Movie Reviews", compuesto de un total de 3875 críticas de cine anotadas con la puntuación con la cual sus autores puntuaron la película acerca de la cual versa cada crítica. Del total de 3875 documentos, 351 tienen asociada una estrella de puntuación, 923 dos estrellas, 1253 tres estrellas, 887 cuatro estrellas y 461 cinco estrellas.

b) Un vocabulario controlado, derivado mediante la aplicación del algoritmo explicado en [18] y completado por traducción automática de las formas contenidas en el General Inquirer.

Para la generación de recursos análogos para el gallego se aplicó el siguiente flujo trabajo:

1- Traducción de español a gallego del corpus "Spanish Movie Reviews". Para este paso se optó por el sistema de traducción Opentrad, [11], en cuyo desarrollo imaxin|**software** colaboró.

2- Traducción de español a portugués de las palabras desconocidas por el par es-gl de Opentrad. Para esta tarea se optó por Google Translate[4], que por nuestra experiencia, para este par de lenguas (es-pt), tiene mayor cobertura léxica que Opentrad aunque a costa de una mucho menor corrección gramatical.

3- La lista de palabras traducida a portugués obtenida en el paso anterior fue, en un tercer paso, traducida a gallego utilizando Opentrad pt-gl.

4- En un cuarto paso se detectaron las palabras desconocidas para el par Opentrad pt-gl, las cuales se transliteraron de portugués a gallego utilizando un script de transliteración

[4] http://translate.google.com/#es|pt|

llamado port2gal[5]. El hecho de que portugués y gallego pertenezcan, tal y como afirman [1], [5] y [6], a un mismo conjunto dialectal gallego-portugués, asegura una alta tasa transferencia léxica entre ambas variantes apenas modificando su forma superficial, esto es su ortografía, tal y como demuestra [13].

5- Para la depuración de errores contenidos en la lista final de palabras obtenidos tras los sucesivos pasos explicados, se procedió a una corrección manual de dicha lista que finalmente se utilizó para corregir el corpus generado en el primer paso.

Para generar un vocabulario controlado como el utilizado en para el desarrollo de la aplicación en español se optó por un flujo de trabajo muy similar al utilizado para la generación del corpus:

1- Traducción de español a portugués del vocabulario, utilizando Google Translate.

2- Expansión del vocabulario utilizando la versión portuguesa de OpenThesaurus, conocido como Caixa Mágica[6].

3- Traducción de portugués a gallego del vocabulario, utilizando Opentrad pt-gl.

4- Transliteración de portugués a gallego de las palabras desconocidas, utilizando port2gal.

5- Depuración manual de los resultados derivados de los procesos de traducción automática, transliteración y expansión del vocabulario.

Mediante estos dos flujos de trabajo finalmente se obtuvo, en primer lugar, un corpus de críticas de cine en gallego compuesto, al igual que en el caso del español de 3875 documentos clasificados según el ranking de estrellas asociadas por los usuarios

[5] port2gal, que es un simple script de Perl, que fue inicialmente desarrollado por Alberto García (de la empresa Igalia) e que posteriormente fue mejorado Pablo Gamallo (Departamento de Lengua Española de la Universidad de Santiago de Compostela). Este script simplemente convierte la ortografía del portugués europeo a la ortografía actual del gallego. port2gal está disponible bajo GPL en http://gramatica.usc.es/~gamallo/port2gal.htm.

[6] http://openthesauruspt.caixamagica.pt/

responsables de dichas críticas. En segundo lugar se obtuvo un vocabulario controlado compuesto de un total de 5448 palabras, de las cuales 2293 fueron clasificadas como positivas y 3155 palabras clasificadas como negativas.

4 Configuración del Algoritmo

Tal y como ya ha sido indicado anteriormente, el tipo de estrategia que se adoptó para el desarrollo de este proyecto estuvo condicionada por fuertes restricciones en relación a los recursos que imaxin|**software**, como PYME, podía invertir. Por lo tanto, se optó por una estrategia basada en Machine Learning. Inspirados en los resultados obtenidos en [15], se escogió Support Vector Machines (SVM) como algoritmo a utilizar para el entrenamiento de un módulo de Opinion Mining.

Para la implementación del módulo de SVM se utilizó la versión 2.90 de libSVM [8], en cuya configuración estándar sólo se modificó el tipo de kernel, pasando del estándar RBF kernel a un POLYNOMIAL kernel.

Para la conversión de los textos en vectores de clasificación se utilizaron las siguientes *features*:

1- La presencia de palabras en los textos de entrenamiento que estuviesen contenidas en nuestro vocabulario controlado de términos positivos y negativos. De esta manera, todos aquellos términos contenidos en las lista de términos positivos fueron codificados en los vectores de clasificación con valor 1; y los negativos con valor -1.

2- En cuanto al resto de palabras no contenidas en las listas de términos positivos o negativos, se optó por la codificación con valor 2 para aquellas palabras del conjunto del corpus presentes también en un determinado texto; y la codificación con valor 3 para

aquellas palabras del conjunto del corpus no presentes en un determinado texto. De esta manera el clasificador tendrá en cuenta tanto aquellas palabras presentes como ausentes para determinar la orientación negativa o positiva de los textos de entrada.

3- Por último, en los vectores de clasificación se incluyeron dos coordenadas adicionales: el total de palabras positivas y el total de palabras negativas detectadas.

5 Resultados

Dado el muy reciente auge del Opinion Mining, no existe ni para español ni para gallego ningún *gold standard* con el cual comparar nuestro sistema de clasificación de sentimientos para determinar su rendimiento. Por esta razón, optamos por crear un pequeño corpus de pruebas que construimos extrayendo al azar textos clasificados como críticas positivas o negativas por los usuarios de diversos sitios web. Los sitios web de los cuales se extrajeron los textos fueron: Google Maps[7], booking.com y la tienda de aplicaciones App Store[8] de Apple. Los dominios a los que pertenecen los textos extraídos son los siguiente: 10 textos (5 positivos y 5 negativos) de críticas de hoteles de Santiago de Compostela y Madrid, 10 textos (5 positivos y 5 negativos) de críticas de restaurantes de Santiago de Compostela; y 10 textos (5 positivos y 5 negativos) de críticas de aplicaciones disponibles en la App Store de Apple.

Los textos escogidos estaban escritos en español y fueron traducidos a gallego manualmente. De esta manera, nos es posible realizar una comparativa directa entre los resultados en español y gallego, pues se trata de los mismos textos simplemente escritos en una u otra lengua.

[7] http://maps.google.com/
[8] http://itunes.apple.com/es/

Las medidas que hemos adoptado para realizar la evaluación de ambos motores de clasificación hemos utilizado las medidas estándar de clasificación: precisión[9] (en inglés precisión), cobertura[10] (en inglés recall) y F-measure[11].

En la tabla 1 se presentan los resultados obtenidos por el motor de clasificación para español. Y en la tabla 2 se presentan los resultados obtenidos por el motor de clasificación para gallego.

Tabla 1. Resultados del clasificador SVM para español.

	Precisión	Cobertura	F-Measure
Positivos	0.79	0.73	0.76
Negativos	0.75	0.80	0.77
Global	0.77	0.77	0.77

Tabla 2. Resultados del clasificador SVM para gallego.

	Precisión	Cobertura	F-Measure
Positivos	0.72	0.87	0.78
Negativos	0.83	0.67	0.74
Global	0.78	0.77	0.77

5.1 Discusión de los Resultados

Antes de realizar una discusión de los resultados obtenidos es importante aclarar que, aunque el número de textos de evaluación es reducido y, por lo tanto, no es posible

[9] La precisión mide el porcentaje de clasificaciones correctas realizadas.
[10] La cobertura mide el porcentaje de clasificaciones relevantes realizadas.
[11] El concepto de F-measure que se utiliza en este estudio es el definido por [19: 174], que combina la precisión y la cobertura bajo una única medida de desempeño global.

realizar un generalización de los resultados, creemos que éstos son una indicación del nivel de desempeño de los motores de clasificación que hemos creado.

Como se puede apreciar en las tablas 1 y 2 los resultados son muy similares para gallego y español. La diferencia más significativa entre ambos es la mayor tendencia que tiene el motor de gallego para clasificar como positivos los textos, como sugiere su cobertura del 87% y su precisión del 72%); y la mayor tendencia del motor de español para clasificar los textos como negativos, como se aprecia por su cobertura del 80% y su precisión del 75%.

En cualquier caso, la clasificación de textos positivos y negativos no baja de una precisión del 70% y la cobertura, sólo en el caso de los textos negativos para gallego, se encuentra ligeramente por debajo del 70%.

Sin embargo, es necesario tener en cuenta que los textos que han servido para el entrenamiento de los clasificadores tanto para gallego como para español pertenecen al dominio de la crítica cinematográfica informal, el cual es muy diferente de los dominios representados en los textos de testeo (que recordemos pertenecen al dominio hotelero, hostelero y tecnológico). Este es un factor que, a buen seguro, juega en contra de la precisión de ambos clasificadores. Aún así, como demuestran los resultados globales, que se encuentran tanto para la precisión como para la cobertura ligeramente por debajo del 80%, el desempeño global de ambos motores de clasificación es, en nuestra opinión, muy satisfactorio.

Por otro lado, y en concreto para el clasificador de gallego, existe otro factor que, en nuestra opinión, es responsable de cierta degradación de los resultados. Este factor es la naturaleza del gallego contenido en los textos que han servido como corpus de

entrenamiento. Así, si bien para el español los textos fueron originalmente escritos en esta lengua, en el caso del gallego los textos han sido obtenidos de manera artificial, esto es, mediante un proceso semiautomático de traducción y transliteración. Por lo tanto, podríamos afirmar que mientras para el español contamos con textos naturales, para el gallego contamos con textos escritos en "pseudo-lengua". De cualquier manera, y a la luz de los resultados obtenidos, el clasificador de gallego tiene un desempeño comparable al clasificador de español.

Por lo tanto, consideramos un éxito la experiencia y creemos que queda demostrada la utilidad de la metodología de propuesta de conversión de español a gallego del corpus y demás recursos de entrenamiento.

6 Trabajo Futuro

El trabajo futuro que hemos planificando llevar a cabo con estos dos prototipos tendrá varias líneas de acción:

1- Ampliación del corpus de entrenamiento, tanto desde el punto de vista de tamaño como de cobertura de dominios en él representados.

2- Refinamiento del algoritmo de clasificación. Se experimentará con nuevos conjuntos de *features*, y se aumentará de la complejidad del sistema de clasificación combinando según un modelo de *voting machines* varios algoritmos de clasificación (por ejemplo, Bayes Naïve y Maximum Entropy).

En imaxin|**software** pensamos que la clasificación automática de opiniones será una de las tónicas dominantes en la Internet de la primera mitad de siglo XXI, y por lo tanto, el desarrollo y evolución de esta herramienta es una de nuestras prioridades.

7 Conclusiones

En este artículo hemos mostrado una metodología de conversión a gallego de fuentes de recursos disponibles en español y portugués necesarios para el entrenamiento de un motor SVM de clasificación de opiniones.

La metodología propuesta combina la traducción automática de español a gallego y de portugués a gallego, la expansión de vocabularios mediante tesauros y la transliteración de palabras de portugués a gallego.

Los resultados obtenidos, que rondan el 80% de cobertura y precisión, son comparables a los de herramientas similares disponibles en otras lenguas.

Sin duda, queda demostrado que la metodología propuesta para la obtención de recursos para gallego ha sido un éxito. En nuestra opinión, esta metodología es perfectamente extrapolable a otras lenguas que guardan lazos especialmente estrechos con variedades lingüísticas desarrolladas en términos de recursos de Procesamiento del Lenguaje Natural.

Referencias

1. Aracil, Ll. et al.: Lingüística e sócio-lingüística galaico-portuguesa: reintegracionismo e conflito lingüístico na Galiza. Asociaçom Socio-Pedagógica Galaico-Portuguesa, Ourense (1985)
2. Blitzer, J., Drezde, M., Pereira, F.. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Annual Meeting-Association For Computational Linguistics, vol. 45 (1), pp. 440--448 (2007)

3. Breck, E., Choi, Y., Cardie, C.: Identifying expressions of opinion in context. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 2683--2688 (2007)
4. Carvalho, P., Sarmento, L., Silva, M.J., de Oliveira, E.: Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 53--56 (2009)
5. Coseriu, E.: El gallego en la historia y en la actualidad. In Actas do II Congresso Internacional da Língua Galego-Portuguesa, pp. 793-800 (1987)
6. Cunha, C., Cintra, L.: Nova Gramática do Português Comtemporâneo. Edições João Sá da Costa, Lisboa (2002)
7. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In Proceedings of the international conference on Web search and web data mining, pp. 231--240 (2008)
8. Fan, R.E., Chen, P.H., Lin, C.J.: Working set selection using second order information for training SVM. Journal of Machine Learning Research, vol 6, pp. 1889--1918 (2005)
9. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 174-181 (1997)
10. L. Cruz, F., Troyano, J.A., Enríquez, F., Ortega, J.: Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. In Procesamiento del Lenguaje Natural, vol. 41, pp. 73--80 (2008)

11. Loinaz, I., Aranrtzabal, I., Forcada, M.L., Gómez Guinovart, X., Padró, Ll., Pichel Campos, J.R., Waliño, J.: OpenTrad: Traducción automática de código abierto para las lenguas del Estado español. Procesamiento del Lenguaje Natural, vol. 27, pp 357--360 (2006)

12. Malouf, R., Mullen, T.: Taking sides: User classification for informal online political discourse. In Internet Research, vol. 18 (2) pp. 177--190 (2008)

13. Malvar, P., Pichel Campos, J.R., Senra, Ó., Gamallo, P., García, A.: Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. In Linguamática, vol 2, n. 2, pp. 31--38 (2010)

14. Moreno Ortiz, A., Pineda Castillo, F., Hidalgo García, R.: Análisis de Valoraciones de Usuario de Hoteles con Sentitext*: un sistema de análisis de sentimiento independiente del dominio. Procesamiento del Lenguaje Natural, vol. 45, pp. 31--39 (2010)

15. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79--86 (2002)

16. Sarmento, L., Carvalho, P., Silva, M.J., de Oliveira, E.: Automatic creation of a reference corpus for political opinion mining in user-generated content. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 29--36 (2009)

17. Silva, M.J., Carvalho, P., Sarmento, L., de Oliveira, E., Magalhães, P.: The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. In Proceedings EPIA'09, 14th Portuguese Conference on Artificial Intelligence (2009)

18. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417--424 (2002)

19. van Rijsbergen, C.J.: Information Retrieval. Butterworths: London (1979)

20. Zhuang, L., Jing, L., Zhu, X.Y.: Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43--50 (2006)