# Generación semiautomática de recursos de Opinion Mining para el gallego a partir del portugués y el español

## *Semiautomatic generation of Opinion Mining resources for Galician from Portuguese and Spanish resources*

**Paulo Malvar Fernández**
Departamento de Ingeniería Lingüística,
imaxin|**software**
Salgueiriños de abaixo nº11 L6, 15891,
Santiago de Compostela
A Coruña
paulomalvar@imaxin.com

**José Ramom Pichel Campos**
Departamento de Ingeniería Lingüística,
imaxin|**software**
Salgueiriños de abaixo nº11 L6, 15891,
Santiago de Compostela
A Coruña
jramompichel@imaxin.com

**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 una alta tasa de transferencia léxica con relación al gallego.
**Palabras clave:** Opinion Mining, Generación Semiautomática, Recursos, Español, Gallego, Portugués

**Abstract:** In spite of the growth experienced in recent years in the field of Natural Language Processing (NLP), researchers and developers who intend to carry out developments for languages other than English still have to face the old problem that needed resources and applications are scarce, if not nonexistent. In this paper we propose a semiautomatic method to generate resources for an Opinion Mining application for Galician. For this we drew from Spanish resources and used Portuguese as a bridge-language that, due to its linguistic proximity, ensures a high lexical transfer rate with Galician.
**Keywords:** Opinion Mining, Semiautomatic Generation, Resources, Spanish, Galician, Portuguese

## *1 Introducción*

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.

Sin embargo, el problema con el que se encuentran investigadores y desarrolladores que pretenden llevar a cabo aplicaciones 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 corpora anotados con

información acerca de la orientación de las opiniones, en español, portugués o gallego este tipo de recursos es prácticamente inexistente.

Frente a esta escasez de recursos, ideamos una solución para aprovechar la relación de proximidad entre gallego con el español y de la especialmente próxima relación entre gallego y portugués, para semiautomáticamente generar los recursos necesarios para una aplicación de Opinion Mining basada en Machine Learning.

## 2 *Recursos disponibles*

Demostración empírica de la abismal distancia que existe en términos de investigación y desarrollo de soluciones de Opinion Mining entre el inglés y otras lenguas, como el español y el portugués, es la amplia diferencia en número de recursos disponibles. Así, para inglés existen actualmente numerosos vocabularios y corpora disponibles para descarga desarrollados para esta lengua[1], que han sido generados a partir de inúmeras investigaciones como (Blitzer, J. et al, 2007), (Ding, X. et al, 2008) y (Pang, B. et all, 2002).

Por el contrario, para el español sólo tenemos constancia de un único corpus, "Spanish Movie Reviews"[2], generado dentro de la investigación desarrollada en (Cruz, F. et al, 2008), y para gallego y portugués nos resultó imposible localizar ningún corpus y/o vocabulario precompilado.

Para hacer frente a la ausencia total de recursos para el gallego, en imaxin|software ideamos una solución para la generación rápida de recursos aprovechando la especialmente estrecha relación entre el gallego y el portugués así como la relación de proximidad entre gallego y español.

## 3 *Metodología propuesta*

Para el desarrollo de una aplicación de Opinion Mining basada en Machine Learning para el gallego necesitábamos un corpus etiquetado con información acerca de la orientación de las opiniones y un vocabulario controlado en el cual también se incluyese información de este tipo acerca de los adjetivos, sustantivos, verbos y adverbios en él contenidos.

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

a) El corpus "Spanish Movie Reviews", compuesto de un total de 3875 críticas de cine anotadas con la puntuación que sus autores asignaron la película 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 (Turney, P.D., 2002) y completado por traducción automática de las formas contenidas en el General Inquirer[3].

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" y del vocabulario controlado de español. Para este paso se optó por el sistema de traducción Opentrad, en cuyo desarrollo imaxin|**software** ha colaborado, (Loinaz, I. et at, 2006).

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 en nuestra opinión, subjetiva a todos los efectos, tiene, para este par de lenguas (es-pt), 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 al gallego utilizando Opentrad pt-gl.

[1] 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/.

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

[3] Dentro del ámbito de los vocabularios o lexicones etiquetados con información sobre la orientación sentimental, el más famoso y utilizado es General Inquirer, que puede ser descargado desde http://www.wjh.harvard.edu/~inquirer/.

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

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 llamado port2gal[5]. El hecho de que portugués y gallego pertenezcan, tal y como afirman (Coseriu, E., 1987) y (Cunha, C. y Cintra, L., 2002), 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 (Malvar, P. Et al, 2010).

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.

Mediante este flujo 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 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*

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.

Además, como es bien sabido, las soluciones basadas en *Machine Learning* ofrecen resultados aceptables en un muy corto espacio de tiempo. Por lo tanto, se optó por una estrategia basada en *Machine Learning*, e, inspirados en los resultados obtenidos en (Pang, B. et al, 2002), 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, (Fan, R.E. et al, 2005), 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, codificados con valor 1, y negativos, codificados 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.

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* como rama de investigación dentro del PLN, hoy en día aún 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 nosotros mismos 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[6], booking.com y la tienda de aplicaciones App Store[7] 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

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

[6] http://maps.google.com/

[7] http://itunes.apple.com/es/

aplicaciones disponibles en la App Store de Apple.

Resulta evidente la disparidad entre estos dominios y el dominio de la crítica cinematográfica, al que pertenecen los textos de entrenamiento del clasificador. Sin embargo, en **imaxin**|software queremos aplicar estos modelos de clasificación de opiniones a ámbitos que no se encuentran dentro del dominio de la crítica cinematográfica. Por lo tanto, pensamos que los resultados obtenidos para los textos de evaluación, sin ser en modo alguno concluyentes, podrían ser un indicador de la aplicabilidad a diversos dominios de los modelos de clasificación aprendidos.

Los textos escogidos estaban escritos en español y fueron traducidos a gallego manualmente por los autores de este trabajo. 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.

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.

| | Precisión | Cobertura |
|---|---|---|
| Positivos | 0.79 | 0.73 |
| Negativos | 0.75 | 0.80 |
| Global | 0.77 | 0.77 |

Tabla 1: Resultados del clasificador SVM para español

| | Precisión | Cobertura |
|---|---|---|
| Positivos | 0.72 | 0.87 |
| Negativos | 0.83 | 0.67 |
| Global | 0.78 | 0.77 |

Tabla 2: Resultados del clasificador SVM para gallego

## 5.1 Discusión de los resultados

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 67%.

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 evaluación (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.

## *6 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.

## *Bibliografía*

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)

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)

Cunha, C., Cintra, L.: Nova Gramática do Português Comtemporâneo. Edições João Sá da Costa, Lisboa (2002)

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)

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)

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)

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)

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)

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)

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)