Actas V Jornadas TIMM, pp 19-21

Análisis de sentimientos multilingüe en la Web 2.0

Javi Fernández, José M. Gómez, Patricio Martínez-Barco
Departamento de Lenguajes y Sistemas Informáticos
{javifm,jmgomez,patricio}@dlsi.ua.es

Universidad of Alicante

Resumen

Nuestra propuesta consiste en un sistema de *análisis de sentimientos* híbrido, que consiste una aproximación híbrida, ya que utiliza un léxico de palabras etiquetadas según su polaridad, además de aprendizaje automático. El léxico se genera de manera automática a partir de un corpus etiquetado, y se asigna a cada término del texto una puntuación para cada polaridad. El aprendizaje automático se encarga de combinar las puntuaciones de cada término del texto para decidir la polaridad de ese texto. En nuestro trabajo nos centraremos en la elección de los términos, en la forma de puntuarlos, y en la forma de combinarlos para determinar la polaridad de un texto.

1. Introducción

La creación de la Web 2.0 ha permitido que los usuarios tengan una participación mucho más activa en Internet, creando no sólo nuevos contenidos, sino también a través de sus comentarios y opiniones. Es por eso por lo que podemos encontrar una gran cantidad de información subjetiva sobre un extenso rango de temas. Esta información puede ser muy valiosa tanto para personas como para empresas y organizaciones públicas. Ya que esta información es textual, es muy complicado extraerla y explotarla de la manera adecuada, por lo que se hace necesaria la utilización de técnicas de *procesamiento del lenguaje natural* (PLN). En el caso de la información subjetiva, la rama de *análisis de sentimientos* (AS) se encarga de detectar cuando un texto es positivo o negativo, basándose únicamente en las palabras de ese texto. La tarea del AS se complica cuando se aplica a la Web 2.0, ya que nos encontramos con nuevos problemas como la informalidad o la existencia de nuevos géneros textuales (como los blogs, los foros, los microblogs y las redes sociales), lo que hace necesario actualizar las técnicas de PLN existentes.

El objetivo de esta tesis es el de diseñar nuevas técnicas de AS para mejorar los sistemas actuales. Entre las novedades de esta propuesta cabe destacar el tratamiento de la flexibilidad y secuencialidad del lenguaje humano y la adaptación a diferentes idiomas.

2. Trabajo relacionado

El objetivo del AS es el de identificar y clasificar las opiniones expresadas en un texto [DHJ11]. Existen dos grupos principales de aproximaciones que se pueden seguir [AK08, Liu10, TBT+11]: aproximaciones basadas en *léxicos* (AS no supervisado) y las aproximaciones basadas en *aprendizaje automático* (AS supervisado). Las aproximaciones basadas en léxicos se centran en construir diccionarios de palabras etiquetadas. Este etiquetado

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

asigna una puntuación para cada palabra y para cada polaridad, indicando cómo de estrecha es la relación entre esa palabra y esa polaridad. La manera más común de clasificar un texto utilizando estas puntuaciones es acumular los pesos de cada palabra, sumando los valores positivos y restando los negativos. Si la puntuación final es positiva, el texto se clasifica como positivo, y si es negativa, se clasifica como negativo. Los diccionarios pueden crearse manualmente [SDS66] o automáticamente [Tur02]. Algunos ejemplos de léxicos son *WordNet Affect* [SV04], *SentiWordNet* [ES06] o *JRC Tonality* [BSG$^+$09]. Sin embargo, es difícil recopilar y mantener un léxico universal, ya que una misma palabra en diferentes dominios puede expresar diferentes opiniones. [Tur02, QLBC09].

La segunda aproximación utiliza técnicas de aprendizaje automático. Estas técnicas requieren la utilización de un corpus que contenga textos clasificados, para crear un clasificador capaz de clasificar nuevos textos. La mayoría de trabajos emplean *Máquinas de soporte vectorial* [MC04, PTS09, WHS$^+$05] o *Naïve Bayes* [PL04, WWC05, TCWX09] porque suelen obtener los mejores resultados. En esta aproximación, los textos se representan como vectores de características y, dependiendo de las características utilizadas, los sistemas pueden obtener mejores resultados (lo más común es utilizar bolsa de palabras o características basadas lexemas [PL08]). Estos clasificadores funcionan muy bien en el dominio en el que han sido entrenados pero empeoran cuando se utilizan en un dominio diferente [PL08, TCWX09].

3. Propuesta

Nuestra propuesta consiste en una aproximación híbrida, ya que utiliza un léxico y aprendizaje automático. El léxico se genera de manera automática a partir de un corpus etiquetado, y se asigna a cada término del texto una puntuación para cada polaridad. El aprendizaje automático se encarga de combinar las puntuaciones de cada término del texto para decidir la polaridad de ese texto. En nuestro trabajo nos centraremos en la elección de los términos, en la forma de puntuarlos, y en la forma de combinarlos para diferenciar la polaridad del texto.

Los términos utilizados son palabras, n-gramas y *skipgrams*. La utilización de skipgrams es muy común en el campo del procesamiento del habla. Esta técnica consiste en obtener n-gramas a partir de las palabras del texto, pero permitir que algunos términos puedan ser saltados. Más especificamente, en un *k-skip-n-gram*, n determina el número de términos, y k el máximo número de términos que se pueden saltar. De esta forma los skipgrams son nuevos términos que conservan parte de la secuencialidad de los términos originales, pero de una forma más flexible que los n-gramas. Cabe destacar que un n-grama se puede definir como un skipgram donde $k = 0$.

Las puntuaciones de los términos (palabras, n-gramas y skipgrams) para cada polaridad se obtienen teniendo en cuenta diferentes factores: *(i)* el número de veces que aparece el término en todos los textos del corpus; *(ii)* el número de veces que aparece el término en los textos del corpus de cada polaridad; *(iii)* en el caso de los n-gramas y skipgrams, el número de palabras que contiene el término; y *(iv)* en el caso de los skipgrams, el número de saltos que se ha realizado.

La combinación de los pesos de los términos para obtener la polaridad del texto se realiza mediante aprendizaje automático, donde cada polaridad se considera como una categoría y cada texto del corpus como un ejemplo de aprendizaje. Hemos seguido dos estrategias diferentes para elegir las características del algoritmo de aprendizaje automático. La primera está basada en clasificación de textos, ya que son los propios términos los que se utilizan como características del modelo de aprendizaje, cuyo peso se corresponde con la puntuación del término al que representan. La segunda estrategia realiza la suma de los pesos de los términos para cada polaridad, y cada una de esas sumas serán las características del modelo de aprendizaje automático.

3.0.1. Agradecimientos

Este trabajo de investigación ha sido parcialmente financiado por la Universidad de Alicante, la Generalitat Valenciana, el Gobierno Español y la Comisión Europea a través de los proyectos «Tratamiento inteligente de la información para la ayuda a la toma de decisiones» (GRE12-44), ATTOS (TIN2012- 38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7-611312), FIRST (FP7-287607) y ACOMP/2013/067

Referencias

[AK08] Michelle Annett and Grzegorz Kondrak. A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. In *Proceedings of the 21st Canadian Conference on Artificial Intelligence (CCAI 2008)*, pages 25–35, 2008.

[BSG+09] Alexandra Balahur, Ralf Steinberger, Erik Van Der Goot, Bruno Pouliquen, and Mijail Kabadjov. Opinion Mining on Newspaper Quotations. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526, 2009.

[DHJ11] Maral Dadvar, Claudia Hauff, and FMG De Jong. Scope of negation detection in sentiment analysis. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011)*, pages 16–20, 2011.

[ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.

[Liu10] Bing Liu. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, pages 1–38. 2010.

[MC04] Tony Mullen and Nigel Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 412–418, 2004.

[PL04] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics (ACL 2004)*, page 271, 2004.

[PL08] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[PTS09] Rudy Prabowo, Mike Thelwall, and Wulfruna Street. Sentiment Analysis: A Combined Approach. *Journal of Informetrics*, 3:143–157, 2009.

[QLBC09] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding Domain Sentiment Lexicon through Double Propagation. In *Proceedings of the 21st international Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1199–1204, 2009.

[SDS66] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. The General Inquirer: A Computer Approach to Content Analysis. 1966.

[SV04] Carlo Strapparava and Alessandro Valitutti. WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086, 2004.

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[TCWX09] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. *Advances in Information Retrieval*, pages 337–349, 2009.

[Tur02] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424, 2002.

[WHS+05] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, 2005.

[WWC05] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.