Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 125-128 recibido 24-03-2017 revisado 12-05-2017 aceptado 25-05-2017

Tecnologías de la lengua para análisis de opiniones en redes sociales

Language technologies for opinion analysis in social networks

Manuel Vilares
Elena Sánchez Trigo
Universidade de Vigo
E.S. de Enxeñaría Informática (Ourense) y
Facultade de Filoloxía e Tradución (Vigo)
{vilares, etrigo}@uvigo.es

Carlos Gómez-Rodríguez
Miguel A. Alonso
Universidade da Coruña
Facultade de Informática
Campus de Elviña, A Coruña
{cgomezr, alonso}@udc.es

Resumen: La reciente popularización de los medios web de comunicación social basados en microtextos, entre los que destaca Twitter, ha permitido globalizar la expresión de opiniones. Aunque los microtextos presentan características léxicas y sintácticas propias respecto al lenguaje estándar, ciertos aspectos básicos del lenguaje han de ser respetados para resultar legibles. En este proyecto proponemos explotar este hecho para obtener una mejora del soporte lingüístico integrado en el tratamiento de microtextos en nuestro ámbito de interés natural, el español y el gallego. Para ello será preciso mejorar el rendimiento de las técnicas actuales de análisis sobre texto estándar, diseñar mecanismos de adaptación a microtextos de aquellos modelos y métodos de análisis que son más efectivos en lenguaje estándar; y realizar una proyección de modelos, métodos y recursos efectivos en otras lenguas.
Palabras clave: Análisis del sentimiento, minería de opiniones, análisis sintáctico, dependencias universales

Abstract: The recent popularization of social media based on microtexts, among which Twitter stands out, has enabled a globalization of the expression of opinions. Although microtexts present some specific lexical and syntactic properties that differ from those of standard text, certain basic aspects of language must be respected so that they are intelligible. In this project, we propose to exploit this fact in order to improve the linguistic support for processing microtexts in our natural sphere of interest: the Spanish and Galician languages. To do so, it will be necessary to improve the performance of current parsing and analysis techniques on standard text, to design mechanisms so that models and methods effective for analyzing standard language can be adapted to microtexts, and to project effective models, methods and resources across languages.
Keywords: Sentiment analysis, opinion mining, parsing, universal dependencies

1 Introducción

Cada vez es mayor el número de usuarios que emplean los medios web de comunicación social basados en microtextos para compartir sus opiniones y experiencias acerca de productos, servicios o personas. La popularización de estos medios, entre los que destaca Twitter, ha permitido globalizar la expresión de opiniones inspirándose en la naturaleza de las interacciones humanas, favoreciendo la generación de comunidades virtuales que posibilitan la colaboración remota y dando lugar a una amplia colección de recursos que permite dotarnos de una visión sobre prácticamente cualquier tema. Por ende, la explotación de estos recursos resulta especialmente útil en los ámbitos comercial y administrativo, donde constituyen una fuente de información fiable en la estimación de cómo los artículos o servicios son percibidos por el usuario. Por extensión, proporciona un punto de partida razonable para detectar qué aspectos poseen una buena acogida en un producto o servicio, y cuáles no. Además, dado que es común que los usuarios establezcan comparaciones con otras empresas o administraciones, ello permitirá a estas conocer los puntos en los que necesitan mejorar y en qué sentido.

Esta situación ha despertado un gran interés por el desarrollo de soluciones que po-

ISSN 1135-5948

sibiliten analizar y monitorizar este flujo ingente de datos, algo que pasa por automatizar este proceso, incorporando métodos inteligentes de acceso a la información. Las dificultades añadidas que representan tanto la efímera vida útil de esta información, como la utilización de lenguaje no estándar y en diferentes idiomas, hacen de esta un área emergente de investigación que requiere la conjunción de capacidades en campos como la lingüística computacional, el aprendizaje automático y la inteligencia artificial.

A este respecto, el análisis de sentimiento o minería de opiniones (MO) es un área de investigación centrada en determinar automáticamente si en un texto se opina o no, si la polaridad o sentimiento que se expresa en él es positiva, negativa o mixta; y en extraer automáticamente la percepción de un autor sobre aspectos concretos de un tema. Las soluciones actuales de MO están muy limitadas por su escaso recurso a las tecnologías de la lengua, al basarse en un procesado superficial que no tiene en cuenta las relaciones sintácticas entre palabras ni sus roles semánticos en las oraciones, lo cual resta capacidad de comprensión en unos textos ya de por sí exiguos. Además, la mayoría de estas soluciones adoptan al inglés como lengua base, con la consiguiente ventaja para usuarios, organizaciones y empresas de países angloparlantes.

En este contexto se desarrolla TELEPARES (Tecnologías de la lengua para análisis de opiniones en redes sociales), un proyecto de investigación coordinado entre investigadores del Grupo COLE (www.grupocole.org) de la Universidade de Vigo (UVigo), del Grupo LYS (www.grupolys.org) de la Universidade da Coruña (UDC) y del CITIUS (citius.usc.es) de la Universidade de Santiago de Compostela (USC). Ha obtenido financiación del Ministerio de Economía y Competitividad dentro del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad (FFI2014-51978-C2-1-R y FFI2014-51978-C2-2-R). Manuel Vilares coordina el proyecto y lidera junto con Elena Sánchez el subproyecto en UVigo (en el que también se integran los investigadores de la USC), mientras que Carlos Gómez-Rodríguez y Miguel A. Alonso lideran el subproyecto en la UDC.

2 Desafíos

Describimos brevemente los principales desafíos a los que hemos de enfrentarnos:

1. La utilización masiva de microtextos, a menudo carentes de contexto lingüístico y que necesitan, para su análisis, de un refinamiento y actualización de las técnicas de linguística computacional.

2. El ruido en los textos, manifestado a nivel léxico en forma de escritura no convencional, utilización irregular de mayúsculas y minúsculas; y abreviaciones idiosincrásicas. A nivel sintáctico, en el uso también irregular de signos de puntuación, y en la eliminación de determinantes y otras partículas cuando su inclusión provocaría la superación del tamaño máximo permitido en un tuit (microtexto de Twitter). A nivel semántico, en el uso de emoticonos que ayudan a proporcionar el contexto de textos extremadamente cortos (alegría, tristeza, enfado, etc.) lo que distorsiona el tratamiento. Lo mismo ocurre a nivel pragmático, donde aquellos permiten distinguir expresiones literales de otras que no lo son (ironía, broma, etc.) y ayudan a trasladar al texto aspectos multimodales del lenguaje como las expresiones faciales de cansancio, aburrimiento o interés.

3. El multilingüismo, ya que menos del 50 % de los tuits están escritos en inglés, con una presencia relevante y creciente del español, portugués y japonés (Carter, Weerkamp, y Tsagkias, 2013). Este hecho hace patente la necesidad de desarrollar aplicaciones multilingües en el ámbito de la minería de textos, confrontando la dificultad derivada de que el español sea una lengua con un soporte moderado de las tecnologías del lenguaje, mientras que las restantes lenguas ibéricas varíen entre un soporte fragmentario y uno débil.

3 Objetivos

Mediante el desarrollo de este proyecto tratamos de afrontar los desafíos indicados anteriormente con el fin de desarrollar un sistema efectivo de MO sobre microtextos escritos en español y gallego, para lo cual será preciso:

- Mejorar el rendimiento de los algoritmos de análisis sintáctico sobre texto estándar, ya que de la calidad del análisis realizado depende en gran medida la

aplicabilidad de los resultados a entornos prácticos, como la MO.

- Mejorar el rendimiento de los sistemas de MO mediante la utilización de la estructura sintáctica para extraer la opinión vertida en un texto, con especial atención al tratamiento de las variadas formas de negación, las frases adversativas y la diferenciación entre texto en modo realis (que se refiere eventos o acciones reales) e irrealis (que expresa deseo, potencialidad o condicionalidad).
- Definir modelos de aprendizaje que faciliten la elección de los mejores analizadores, minimizando el coste del proceso de entrenamiento sin perjuicio de la calidad.
- Definir técnicas efectivas que permitan proyectar las herramientas y recursos desarrollados para una lengua, a otra distinta. Ello permitirá, por ejemplo, obtener un analizador sintáctico para un idioma en el que no está disponible un corpus de textos anotados sintácticamente (como es el caso del gallego), a partir de los analizadores obtenidos para otros (como puede ser el español) que sí disponen de tales corpus.
- Definir técnicas efectivas de adaptación de los analizadores a un dominio distinto de aquel para el que fueron concebidos inicialmente, lo que permitirá obtener herramientas para textos no convencionales, como es el caso de los microtextos presentes en los medios web de comunicación social. Ello conlleva también mejorar el rendimiento de los algoritmos de análisis léxico en este contexto, con especial atención al tratamiento de sus peculiaridades léxicas: errores ortográficos, abreviaturas, emoticonos y almohadillas. Todo ello permitirá extraer unidades lingüísticas coherentes que contengan las expresiones de opinión presentes en un enunciado, así como su orientación semántica o polaridad.

4 Resultados alcanzados

Análisis sintáctico: se han realizado desarrollos relevantes en analizadores de dependencias basados en grafos (Gómez Rodríguez, 2016b) y transiciones (Gómez Rodríguez y Fernández-González, 2016). Se ha descrito la relación entre la manera en que funcionan los analizadores basados en transiciones y la forma en que los humanos procesamos el lenguaje (Gómez Rodríguez, 2016a). Se han analizado las dependencias no proyectivas (Ferrer-i-Cancho y Gómez-Rodríguez, 2016a) y se han estudiado las propiedades y distribución estadística de las longitudes de las dependencias (Ferrer-i-Cancho y Gómez-Rodríguez, 2016b; Esteban, Ferrer-i-Cancho, y Gómez-Rodríguez, 2016). Se ha comparado la eficacia de analizadores sintácticos, modelos vectoriales y redes neuronales en tareas de similaridad léxica y analogía (Gamallo, 2017).

Sistemas de MO: se han diseñado e implementado sistemas de minería de opiniones multilingües no supervisados (Vilares, Gómez-Rodríguez, y Alonso, 2017) y supervisados (Vilares, Alonso, y Gómez-Rodríguez, 2017) capaces de proporcionar un análisis de la polaridad de una oración teniendo en cuenta los fenómenos sintácticos que la condicionan (negación, oraciones adversativas, intensificación e irrealis), obteniendo resultados más precisos que los sistemas que se quedan en un nivel léxico. Mediante la aplicación de técnicas de *deep learning* se obtuvo el segundo puesto en las subtareas B y D en la campaña de evaluación SemEval 2016 task 4 (Vilares et al., 2016).

Modelos de aprendizaje: se han diseñado e implementado sendos algoritmos para la predicción del rendimiento en procesos de aprendizaje automático y localización de las instancias para el muestreo (Vilares, Darriba, y Ribadas, 2017).

Recursos lingüísticos: se ha comprobado empíricamente la efectividad de las Universal Dependencies en el procesamiento multilingüe (Vilares, Alonso, y Gómez-Rodríguez, 2016). Se ha creado Galician-TreeGal, un treebank de dependencias universales manualmente revisado para gallego (Garcia, Gómez-Rodríguez, y Alonso, 2016). Se ha creado el corpus EN-ES-CS con tuits en los que se utiliza más de un idioma (Vilares, Alonso, y Gómez-Rodríguez, 2017). Se ha creado el recurso Spanish SentiStrength, cuya eficiencia y utilidad práctica ha sido analizada sobre un conjunto de mensajes de naturaleza política (Vilares, Thelwall, y Alonso, 2015; Vilares y Alonso, 2016).

Normalización de textos: se ha estudiado la robustez de las técnicas basadas en

n-gramas de caracteres para la corrección de palabras en un entorno multilingüe (Vilares et al., 2016a; Vilares et al., 2016b) y se ha experimentado con técnicas de deep learning para la segmentación de palabras (Doval, Gómez-Rodríguez, y Vilares, 2016).

Bibliografía

Carter, S., W. Weerkamp, y M. Tsagkias. 2013. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Doval, Y., C. Gómez-Rodríguez, y J. Vilares. 2016. Segmentación de palabras en español mediante modelos del lenguaje basados en redes neuronales. *Procesamiento del Lenguaje Natural*, 57:75–82.

Esteban, J. L., R. Ferrer-i-Cancho, y C. Gómez-Rodríguez. 2016. The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics: Theory and Experiment*, (2016):063401.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016a. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016b. Liberating language research from dogmas of the 20th century. *Glottometrics*, 33:33–34.

Gamallo, P. Pendiente de publicación. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*.

Garcia, M., C. Gómez-Rodríguez, y M. A. Alonso. 2016. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40.

Gómez Rodríguez, C. 2016a. Natural language processing and the now-or-never bottleneck. *Behavioral and Brain Sciences*, 39:e74.

Gómez Rodríguez, C. 2016b. Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, 42(4):809–817.

Gómez Rodríguez, C. y D. Fernández-González. 2015. An efficient dynamic oracle for unrestricted non-projective parsing. En *Proceedings of ACL-IJCNLP 2015*, páginas 256–261, Beijing, China.

Vilares, D. y M. A. Alonso. 2016. A review on political analysis and social media. *Procesamiento del Lenguaje Natural*, 56:13–23.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. En *Proceedings of ACL 2016*, páginas 425–431, Berlin, Germany.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2017. Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595–607.

Vilares, D., Y. Doval, M. A. Alonso, y C. Gómez-Rodríguez. 2016. Exploiting neural activation values for Twitter sentiment classification and quantification. En *Proceedings of SemEval-2016*, páginas 79–84, San Diego, California.

Vilares, D., C. Gómez-Rodríguez, y M. A. Alonso. 2017. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Vilares, D., M. Thelwall, y M. A. Alonso. 2015. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813.

Vilares, J., M. A. Alonso, Y. Doval, y M. Vilares. 2016a. Studying the effect and treatment of misspelled queries in cross-language information retrieval. *Information Processing & Management*, 52(4):646–657.

Vilares, J., M. Vilares, M. A. Alonso, y M. P. Oakes. 2016b. On the feasibility of character n-grams pseudo-translation for cross-language information retrieval tasks. *Computer Speech and Language*, 36(36):136–164.

Vilares, M., V. M. Darriba, y F. J. Ribadas. 2017. Modeling of learning curves with applications to POS tagging. *Computer Speech and Language*, 41:1–28.