Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 143-146 recibido 15-3-2016 revisado 13-04-2016 aceptado 29-04-2016

SomEMBED: Comprensión del lenguaje en los medios de comunicación social-Representando contextos de forma continua

SomEMBED: Social Media language understanding- EMBEDing contexts

Paolo Rosso, Roberto Paredes
PRHLT, Universitat Politècnica de València
Camino de Vera s/n. 46022
Valencia, España
{prosso,rparedes}@prhlt.upv.es

Mariona Taulé, M. Antònia Martí
CLiC, Universitat de Barcelona
Gran Via 585, 08007
Barcelona, España
{mtaule,amarti}@ub.edu

Resumen: SomEMBED es un proyecto coordinado en el que participan el centro de investigación *Pattern Recognition and Human Language Technology* (PRHLT) de la Universitat Politècnica de València (UPV) y el grupo de investigación *Centre de Llenguatge i Computació* (CLiC) de la Universitat de Barcelona. Se trata de un proyecto del programa de I+D (TIN2015-71147) financiado por el Ministerio de Economía y Competitividad. Paolo Rosso coordina el proyecto SomEMBED y lidera el subproyecto SomEMBED-APP y Mariona Taulé lidera el subproyecto SomEMBED-SLang.
Palabras clave: *Embeddings*, representaciones distribuidas, medios de comunicación social, lenguaje no estándar, aplicaciones

Abstract: SomEMBED is a coordinated project involving the research center of *Pattern Recognition and Human Language Technology* (PRHLT) of the Universitat Politècnica de València and the research group of *Centre de Llenguatge i Computació* (CliC) of the Universitat de Barcelona. This is an R&D project (TIN2015-71147) funded by the Spanish Ministry of Economy and Competitiveness. Paolo Rosso coordinates the SomEMBED project and leads the subproject SomEMBED-APP and Mariona Taulé leads the SomEMBED-SLang subproject.
Keywords: Embeddings, distributed representations, social media, non-standard language, applications

1 Introducción

El proyecto SomEMBED tiene como objetivo general avanzar en el área de la Lingüística Computacional (LC) y el Procesamiento del Lenguaje Natural (PLN) con el fin de afrontar y dar solución a los retos que plantea el uso de la lengua en los medios de comunicación social en la web.

Desde la LC nuestro objetivo es el desarrollo de técnicas y métodos para la modelización de la lengua no estándar a partir de corpus representativos de los medios de comunicación social. Desde el PLN, nuestro objetivo es el desarrollo de nuevas técnicas y métodos a partir del estado actual de los conocimientos científico-técnicos para la resolución de tareas específicas en el marco de aplicaciones concretas.

En este proyecto convergen tres líneas de investigación estrechamente relacionadas: 1) la exploración y producción de diferentes metodologías para la extracción automática de patrones sintáctico-semánticos con el fin de representar semánticamente el contenido de los documentos, teniendo como eje central los métodos basados en representaciones continuas de texto (*embeddings*) que permiten modelar el contexto de un modo eficaz y eficiente; 2) el desarrollo de aplicaciones para la resolución de tareas concretas de PLN que permitan mejorar la comprensión automática del texto (p.e.: la detección del lenguaje figurado), e identificar aspectos claves del perfil de autores - edad, sexo, personalidad, lengua nativa, variedad lingüística (Franco-Salvador et al., 2015)- con especial interés en distinguir a los usuarios de los países de lengua hispana (España, México, Perú, etc.), aspectos que además permiten

ISSN 1135-5948

utilizar su información en tareas como la minería de productos y servicios, en especial para la detección de opiniones falsas, y 3) la creación de recursos lingüísticos, especialmente corpus anotados, orientados al análisis de la lengua no estándar que servirán de base para la metodología de extracción de patrones y para las aplicaciones mencionadas que se van a desarrollar. Estas tres líneas de investigación se concretan en los objetivos que se detallan en el siguiente apartado.

2 Objetivos

1) La exploración y producción de nuevos métodos para la detección de patrones sintáctico-semánticos:

a. Experimentación con técnicas de generación de representaciones continuas de palabras para la obtención de *clusters* de palabras relacionadas[1].
b. Aplicación de la metodología DISCver (desarrollada en el proyecto anterior[2]) a una lengua distinta del español. En concreto, proponemos aplicarla al corpus del inglés *ukWaC* (Baroni et al., 2009) con el objetivo de ratificar la transportabilidad del método.
c. Experimentar con métodos alternativos para la obtención de patrones sintáctico-semánticos, basados en representaciones continuas de conjuntos de palabras (Le y Mikolov 2014).
d. Búsqueda de patrones que cumplan determinadas restricciones. En concreto, detección de patrones relacionados con los operadores de incertidumbre (*backward entailment operators*), fundamentales para la correcta interpretación de la polaridad (Danescu et al., 2009).

2. Creación de la infraestructura básica de recursos lingüísticos orientados al análisis de la lengua no estándar.[3]

a. Creación y anotación lingüística de corpus de lengua no estándar para diferentes variantes del español: *HispaSocialMedia* y *HispaLearners*. Por razones expositivas, denominaremos *HispaSocialMedia* al conjunto de corpus extraídos de distintos medios de comunicación social (microblogs, blogs, *reviews*, fórums) e *HispaLearners* para referirnos al conjunto de corpus formados por producciones orales (transcritas) o escritas producidas por aprendices de español como lengua extranjera.
e. Desarrollo de una base de conocimiento de patrones sintáctico-semánticos organizados de forma jerárquica según diferentes niveles de abstracción lingüística. Los patrones serán el resultado de la aplicación de los diferentes métodos descritos en el primer objetivo a los corpus *HispaSocialMedia* e *HispaLearners*.

3. Desarrollo de aplicaciones para los medios de comunicación social que cubran las tareas siguientes:

a. Identificación de lenguaje figurado (metáfora, analogía, humor, ironía y sarcasmo) frente al lenguaje literal (se utilizará el corpus *HispaSocialMedia*).
b. Detección de comunidades de usuarios en los medios de comunicación social con el objetivo de analizar las similitudes a nivel de patrón sintáctico-semántico entre los grupos de usuarios que utilizan y comparten diferentes tipos de lenguaje, por ejemplo el figurado.
c. Identificación de la variedad lingüística. En concreto, la detección del español peninsular frente al español de Latinoamérica: México, Argentina, etc. (se usará el corpus *HispaSocialMedia*).
d. Identificación de la lengua nativa de usuarios que escriben en español (se usará el corpus *HispaLearners*).
e. Identificación de los rasgos de los autores de textos en medios de comunicación social: sexo, edad, personalidad, tendencia política, entre otras (se usará el corpus corpus *HispaSocialMedia*).

3 Hipótesis

Existe la hipótesis de partida comúnmente aceptada de que la variante de lengua informal usada en los medios de comunicación social difiere sensiblemente de la variante estándar. La mayoría de herramientas de PLN disponibles actualmente están pensadas y adaptadas para la

[1] Véanse los *clusters* resultantes en: http://clic.ub.edu/corpus/es/Diana-Araknion-KB

[2] DIANA-Construcciones (TIN2012-38603).

[3] Se entiende por lengua no estándar todas aquellas variantes lingüísticas que se desvían de la forma oral o escrita estándar y que suelen transgredir la norma gramatical.

lengua estándar. Existen, a nuestro entender, dos maneras posibles de abordar el problema del tratamiento de la lengua no estándar: la primera consiste en la extensión o adaptación de las herramientas existentes, mientras que la segunda aborda el problema desde otra metodología radicalmente distinta. Esta metodología consiste en la extracción de patrones sintáctico-semánticos que modelizan el contenido de los corpus. La primera opción presenta un problema grave ya que la variación en los corpus de lengua informal es impredecible, de manera que hace inviable la adaptación de las herramientas existentes.

Nuestra hipótesis de trabajo es que: a) la modelización de los corpus en base a patrones sintáctico-semánticos permite una adecuada representación de su contenido; b) esta modelización abre la puerta al desarrollo de aplicaciones de PLN más eficientes en medios de comunicación social, y c) se obtiene una información lingüística relevante que sirve de base a estudios teóricos sobre el lenguaje y apunta a un modelo de gramática de la actuación, es decir, una gramática de la lengua en uso. La ventaja es que la mayoría de métodos que se desarrollan en el marco de esta aproximación son independientes de la lengua y de la variante en que se utiliza. Esto es debido a que para cada colección de corpus, nuestra aproximación obtiene una representación en términos de patrones y son estos patrones los que proporcionan las características lingüísticas de la misma. Esta representación constituye el input para el desarrollo de aplicaciones y tareas de PLN.

4 Metodología

La extracción automática de los patrones se realizará a partir del tratamiento masivo de corpus, aplicando principalmente técnicas de *deep learning*. Siguiendo la línea de investigación abierta por Mikolov et al., (2013b), se procederá a generar nuevos patrones, utilizando los conjuntos de herramientas de word2vec[4] y Glove[5]. Se trata de un método basado en las representaciones continuas de palabras (*embeddings*). Estas representaciones consisten en vectores n-dimensionales que han sido generados mediante algoritmos log-lineales de modo que vectores de palabras similares en contexto guardan una similitud próxima. A los modelos originales de generación de vectores continuos de palabras, compilados en el conjunto de herramientas de word2vec, le han seguido otras alternativas igualmente eficientes (Pennington et al., 2014), además de otros modelos que permiten generar representaciones continuas de conjuntos de palabras para su aplicación en frases y documentos (Le y Mikolov, 2014).

Otra línea de investigación se centrará en la obtención automática de operadores de incertidumbre para el español basándonos en la propuesta de Danescu et al., (2010) que, a partir de operadores comúnmente aceptados (por ejemplo: no, jamás, dudar, imaginar, etc.) aplica algoritmos recursivos para extraer nuevos operadores mediante la identificación de contextos prototípicos que los suelen acompañar.

Los diferentes métodos y técnicas que se lleven a cabo se aplicarán para la resolución de tareas específicas en el marco de aplicaciones concretas (véase apartado 3 de la sección 2).

En cuanto a los corpus, se recopilarán mediante el uso de algoritmos de web *crawling* para la compilación del corpus *HispaSocialMedia* y el diseño de tareas (p.e. redacción de textos) para la creación del corpus *HispaLearners*. Una vez obtenidos los corpus se procederá a su procesamiento morfosintáctico automático. Se anotarán manualmente aspectos específicos (p.e.: anotación de los operadores de incertidumbre, especialmente la negación) en una selección de los corpus compilados.

5 Resultados esperables

Destacamos el desarrollo de diferentes técnicas y métodos susceptibles de ser incorporados en aplicaciones de entornos de medios de comunicación social con diferentes finalidades:

- La identificación de la variedad lingüística y la lengua nativa del autor, que permitirá geolocalizar y discriminar la relevancia de noticias y eventos en medios de comunicación social para su posterior análisis y explotación. Dada la activa participación de la población en las redes sociales y la complejidad de distinguir la procedencia geográfica del autor (España, México, Perú, Chile, Argentina, etc.), dicha aplicación tiene mucha importancia en los medios de comunicación (análisis de opinión, difusión e impacto) y la industria (análisis de

[4] https://code.google.com/p/word2vec/

[5] http://nlp.stanford.edu/projects/glove/

opinión, estudio de mercado, explotación de productos, etc.).

- La identificación de rasgos del autor (sexo, edad, personalidad, tendencia política, entre otros) que proporcionará información detallada sobre el perfil del autor del texto ante el que nos encontramos y potencialmente contribuirá notablemente en la lingüística forense de cara a la identificación y estudio de infractores, delincuentes y criminales. Puede ser también de utilidad para los intereses de los medios de comunicación social y la industria mencionados anteriormente.
- La detección de opiniones fraudulentas y otros casos de engaño creados por parte de personas específicamente contratadas para este fin. Nuestra contribución consistirá en el desarrollo de técnicas para la extracción de las construcciones o patrones lingüísticos más recurrentes en la expresión de opiniones fraudulentas y otros casos de engaño sobre personas, organizaciones, productos y servicios, lo cual tendrá un potencial uso para la industria.
- La detección de mecanismos de lenguaje figurado en textos subjetivos (ironía, metáfora, humor, etc.) de los medios de comunicación social y en comunidades de usuarios. Las herramientas que se desarrollarán permitirán superar la barrera que supone el uso de lenguaje figurado, que altera radicalmente el significado del mensaje, y contribuirá a la mejora de tareas mencionadas como el análisis de opinión, además de proporcionar información útil de cómo la información y el lenguaje se comparte dentro de comunidades concretas de usuarios.
- Los nuevos recursos y herramientas que se prevé desarrollar han de incidir en la mejora de las técnicas y métodos de PLN que se usarán en aplicaciones sobre los medios de comunicación social. En concreto, los patrones sintáctico-semánticos generados mediante representaciones continuas servirán para modelizar el contenido de corpus mejorando los resultados obtenidos hasta el momento.
- Por otro lado, los corpus que se prevé construir -*HispaSocialMedia* y *HispaLearners*- sobre la lengua no estándar para las diferentes tareas sobre las que se trabajará en este proyecto, fomentarán y contribuirán a la investigación y avance por parte de otras entidades.

El uso de la lengua no estándar dificulta el uso de las herramientas tradicionales de análisis de la lengua y requiere del estudio detallado de este tipo de variante lingüística en estos medios, y de la investigación y desarrollo de nuevos algoritmos y metodologías para el procesamiento del lenguaje que permitan poder compilar, comprender y explotar el conocimiento que se encuentra dentro de los medios de comunicación social. Los resultados del proyecto coordinado *SomEMBED* podrán incidir favorablemente en el desarrollo de aplicaciones en dicho ámbito (redes sociales, blogs, foros, canales de opinión) y en nuestro conocimiento sobre la lengua no estándar que allí se utiliza.

Bibliografía

Baroni M., S. Bernardini, A. Ferraresi y E. Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43: 209-226, Spinger.

Danescu-Niculescu-Mizil, C., L. Lee, y R. Ducott. 2009. Without a'doubt'?: unsupervised discovery of downward-entailing operators. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, páginas 137-145. Association for Computational Linguistics.

Franco-Salvador, M., F. Rangel, P. Rosso, M. Taulé y A. Martí. 2015. Language Variety Identification using Distributed Representations of Words and Documents. *Proceedings of the 6th International Conerence. of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction* (CLEF 2015), Springer-Verlag, LNCS (9283): 28-40.

Le, Q. V., y T. Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv*:1405.4053.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Pennington, J., R. Socher y C.D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing* (EMNLP 2014), 12: 1532-1543.