Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 209-212 recibido 28-11-2020 revisado 20-01-2021 aceptado 05-02-2021

Buscando robustez en un mundo multilingüe: de pipelines a embeddings

Seeking robustness in a multilingual world: from pipelines to embeddings

Yerai Doval
Grupo COLE, Escola Superior de Enxeñaría Informática,
Universidade de Vigo, España
yerai.doval@uvigo.es

Resumen: Tesis elaborada por Yerai Doval Mosquera bajo la supervisión de los profesores Jesús Vilares Ferro (Universidade da Coruña), Manuel Vilares Ferro (Universidade de Vigo) y la colaboración de Carlos Gómez Rodríguez (Universidade da Coruña). Su defensa tuvo lugar el día 17 de diciembre de 2019 en la Universidade da Coruña, con un tribunal compuesto por los profesores Miguel Ángel Alonso Pardo (Universidade da Coruña), Pavel Bernard Brazdil (Universidade do Porto) y María Lourdes Araújo Serna (Universidad Nacional de Educación a Distancia). Mereció la calificación de Sobresaliente *cum laude* con Mención Doctor Internacional.
Palabras clave: texto ruidoso, pipelines, word embeddings, multilingüismo, sistemas robustos

Abstract: Thesis prepared by Yerai Doval Mosquera under the supervision of professors Jesús Vilares Ferro (Universidade da Coruña), Manuel Vilares Ferro (Universidade de Vigo) and with the collaboration of Carlos Gómez Rodríguez (Universidade da Coruña). Its defense took place on December 17, 2019 at the University of Coruña, with a panel composed of professors Miguel Ángel Alonso Pardo (Universidade da Coruña), Pavel Bernard Brazdil (Universidade do Porto) and María Lourdes Araújo Serna (Universidad Nacional de Educación a Distancia). It obtained the highest qualification with *cum laude* honours and International Mention.
Keywords: noisy text, pipelines, word embeddings, multilingualism, robust systems

1 Introducción

Los usuarios de Internet producen y comparten todo tipo de contenido escrito en una amplia variedad de servicios y plataformas: páginas Web, correos electrónicos, mensajes de chat, publicaciones en redes sociales, etc. El tipo de textos producido y compartido por estos usuarios tiene dos rasgos específicos que lo diferencian de la mayoría de textos escritos, a la vez que lo acercan al lenguaje hablado: su *espontaneidad* e *informalidad*. Esto da como resultado lo que se denominan *textos ruidosos*, con un estilo de escritura fuertemente influenciado por hábitos del habla.

Asimismo, aún cuando el inglés es el idioma predominante, Internet demuestra un claro y creciente *multilingüismo* al acomodar contenidos en prácticamente cualquier idioma. No solo eso, es también habitual el *code-switching* o combinación de palabras en distintos idiomas en una misma frase u oración.

En este trabajo de tesis (Doval, 2019) estudiamos dos enfoques para abordar los desafíos en el procesamiento de contenidos textuales no estándar y multilingües generados por los usuarios, tal y como se pueden encontrar en la Web a día de hoy. Este tipo de textos son denominados a menudo *textos cortos* o *microtextos*.

En primer lugar, presentamos un enfoque tradicional basado en *pipelines* discretos en el que el texto de entrada es preprocesado para facilitar su tratamiento por otros sistemas. Esto implica abordar el problema del multilingüismo identificando el idioma de la entrada para, seguidamente, tratar los fenómenos de escritura no estándar específicos de dicho idioma presentes en dicha entrada. Para ello se aplicarán técnicas de normalización del texto y (re-)segmentación de palabras.

En segundo lugar, analizamos las limitaciones inherentes a este tipo de modelos dis-

ISSN 1989-7553. DOI 10.26342/2021-66-20

cretos, lo cual nos conduce a un enfoque centrado en el empleo de modelos continuos basados en *word embeddings* (i.e., representaciones vectoriales). En este caso, el preprocesamiento explícito de la entrada es sustituido por la codificación de las características lingüísticas y demás matices propios de los textos no estándar en el propio espacio de *embedding* (un espacio vectorial). Nuestro objetivo es obtener modelos continuos que no sólo superen las limitaciones de los modelos discretos, sino que también se alineen con el actual estado de la cuestión del Procesamiento de Lenguaje Natural (PLN), dominado por sistemas basados en redes neuronales.

2 Estructura de la tesis

La memoria, escrita en inglés, está organizada en cinco partes, más apéndices:

Parte 1

- El **Capítulo 1** describe los fenómenos de *texting* que caracterizan el uso del lenguaje en Internet, los cuales constituyen la motivación de este trabajo, y las tareas de preprocesamiento consideradas, a la vez que introduce los enfoques que se estudian en los siguientes capítulos.
- El **Capítulo 2** recoge la terminología relevante para el dominio de nuestro trabajo y, a continuación, introduce dos importantes recursos ampliamente utilizados no sólo aquí, sino en muchos otros sistemas de PLN: los modelos de lenguaje y las *word embeddings*.

Parte 2

- El **Capítulo 3** presenta la tarea de identificación del idioma en el contexto del taller TweetLID (Zubiaga et al., 2014) de identificación del idioma de tuits en el contexto ibérico, y analiza el rendimiento de las herramientas comunes de identificación del idioma para dicha tarea.
- El **Capítulo 4** propone un enfoque sencillo para la normalización de microtextos de cara a la Tarea 2 del W-NUT 2015 (Baldwin et al., 2015), con una estructura clásica en dos pasos (generación y selección de candidatos de normalización), y centrándose en la modularidad y la adaptabilidad de la aproximación.
- El **Capítulo 5** presenta un enfoque de segmentación de palabras basado en un algoritmo de búsqueda y un modelo de lenguaje, además de estudiar su rendimiento cuando este último componente se implementa como una red neuronal recurrente o bien un modelo de n-gramas.

Parte 3

- El **Capítulo 6** analiza, desde un punto de vista teórico, las limitaciones inherentes a los *pipelines* discretos y otros enfoques similares, y cómo el uso directo de *word embeddings* resuelve o evita los problemas resultantes.

Parte 4

- El **Capítulo 7** describe cómo mejorar la integración de espacios de *word embeddings* multilingües obtenidos mediante la alineación de espacios monolingües.
- El **Capítulo 8** describe una técnica de adaptación que mejora el rendimiento de los modelos de *word embeddings* monolingües existentes en el caso de textos ruidosos. También presenta un breve estudio sobre el efecto de la mala segmentación de palabras en el rendimiento de las *word embeddings*.

Parte 5

- El **Capítulo 9** cierra el trabajo de tesis presentando las conclusiones más relevantes y las futuras líneas de trabajo.

Apéndices La memoria de tesis incluye a mayores una serie de apéndices que, si bien aportan conclusiones significativas, no son necesarias para seguir la línea argumental principal de la disertación.

- El **Apéndice A** analiza el rendimiento de una amplia gama de algoritmos fonéticos en el proceso de generación de candidatos para normalización (Doval, Vilares, y Vilares, 2018), proceso descrito en el Capítulo 4.
- El **Apéndice B** presenta un amplio análisis de los factores que suelen intervenir en la alineación bilingüe de los espacios de *embedding* monolingües descritos en el Capítulo 7.

3 Contribuciones

Resumimos a continuación las contribuciones más relevantes de la tesis, correspondientes a las Partes 2, 3 y 4 de la misma.

3.1 Enfoque discreto: el *pipeline*

Nuestro *pipeline* de preprocesamiento está formado por las siguientes etapas o módulos:

Identificación del idioma Nuestra solución propuesta (Doval, Vilares, y Vilares, 2014) pasa por adaptar y reentrenar las herramientas de identificación automática del idioma existentes utilizando varios corpus de nuestra elección, de modo que todos compartan el mismo punto de partida.

Normalización de microtexto Nuestro enfoque (Doval, Vilares, y Gómez-Rodríguez, 2015) se basa en un proceso en dos fases: (1) generación de normalizaciones candidatas empleando correctores ortográficos y diccionarios; (2) selección de candidatas, implementada a través de un modelo de lenguaje a nivel de palabra y un algoritmo de búsqueda.

Segmentación de palabras Nuestra solución para corregir las segmentaciones incorrectas (Doval y Gómez-Rodríguez, 2019; Doval, Gómez-Rodríguez, y Vilares, 2016) consta de dos componentes: (1) un algoritmo de *beam search*, que genera y elige entre los posibles candidatos de segmentación de forma incremental; y (2) un modelo de lenguaje a nivel de byte o carácter que permite que el algoritmo pueda clasificar los candidatos, e implementado como una red neuronal recurrente o modelo de n-gramas.

3.2 Limitaciones del enfoque discreto y transición a un modelo continuo

Tras un estudio de las limitaciones inherentes al enfoque discreto tradicional, hemos comprobado que podemos resolver los problemas derivados de la propagación de errores y la fragmentación del contexto cambiando a un modelo continuo centrado en *word embeddings*. Estas representaciones vectoriales permiten mejorar la integración de nuestro sistema a la vez que constituyen un *lenguaje intermedio* para soportar entornos multilingües. Asimismo, se pueden utilizar para codificar las particularidades derivadas de los fenómenos propios de los textos generados por los usuarios, haciendo así innecesario su procesamiento explícito.

3.3 *Embeddings* multilingüe

Hemos desarrollado un método de postprocesamiento (Doval et al., 2018; Doval et al., 2019), que hemos denominado Meemi (por "*Meeting in the Middle*"), que mejora la integración de espacios monolingües inicialmente aislados y posteriormente alineados mediante herramientas como VecMap (Artetxe, Labaka, y Agirre, 2018) y MUSE (Conneau et al., 2018). Para mejorar dicha integración, aplicamos sobre estos alineamientos una transformación lineal no restringida que se aprende haciendo corresponder las traducciones de palabras con sus representaciones promedio.

De manera notable, hemos ido más allá de la configuración bilingüe habitual en este tipo de herramientas y hemos mostrado también cómo Meemi puede extenderse, de forma natural, a un número arbitrario de idiomas, los cuales acaban integrados en un único espacio vectorial compartido. En este caso, utilizamos métodos ortogonales en el primer paso de alineación que solo transforman el espacio de *embedding* de uno de los lenguajes (origen) mientras deja intacto el otro espacio (destino), que se convierte en el espacio vectorial multilingüe. Este proceso se repite para los espacios de origen correspondientes a cada idioma restante.

3.4 *Embeddings* robustas para microtextos

Los modelos de *word embeddings* como word2vec (Mikolov et al., 2013) o fastText (Bojanowski et al., 2016), son de por sí capaces de agrupar variantes estándar y no estándar de palabras si se les proporciona un corpus de entrenamiento lo suficientemente grande que incluya tales variantes (Sumbler et al., 2018). Sin embargo, nosotros proponemos ir un paso más allá con una modificación del modelo *skipgram* de fastText que permite no solo mejorar el rendimiento de las *word embeddings* resultantes en textos ruidosos, sino que además permite preservar su rendimiento en los textos estándar (Doval, Vilares, y Gómez-Rodríguez, 2020).

Para ello, introducimos un nuevo conjunto de palabras en el proceso de entrenamiento, que denominamos *bridge-words* (*palabras puente*), y cuyo objetivo es actuar a modo de guía a la hora de enlazar una palabra estándar con sus contrapartidas con ruido.

La robustez de las *embeddings* resultantes frente al ruido presente en el texto, se hace especialmente patente en el caso de modelos entrenados de extremo a extremo. El uso de estas *embeddings* nos evita tener que preprocesar la entrada original para modificarla, co-

mo venía ocurriendo hasta ahora, lo que solía llevar a introducir errores que luego se propagarían a otras partes de nuestros sistemas.

Bibliografía

Artetxe, M., G. Labaka, y E. Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. En *Proc. of the 32th AAAI Conf. on Artificial Intelligence, AAAI 2018*, páginas 5012–5019.

Baldwin, T., M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, y W. Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. En *Proc. of the 1st Workshop on Noisy User-generated Text, W-NUT 2015*, páginas 126–135.

Bojanowski, P., E. Grave, A. Joulin, y T. Mikolov. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, y H. Jégou. 2018. Word translation without parallel data. En *Proc. of the 6th Int. Conf. on Learning Representations, ICLR 2018*.

Doval, Y. 2019. *Seeking robustness in a multilingual world: from pipelines to embeddings*. Ph.D. tesis, Universidade da Coruña, 12.

Doval, Y., J. Camacho-Collados, L. E. Anke, y S. Schockaert. 2019. Meemi: A simple method for post-processing cross-lingual word embeddings. *arXiv preprint arXiv:1910.07221*.

Doval, Y., J. Camacho-Collados, L. Espinosa-Anke, y S. Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. En *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2018*, páginas 294–304.

Doval, Y. y C. Gómez-Rodríguez. 2019. Comparing neural-and n-gram-based language models for word segmentation. *Journal of the Association for Information Science and Technology*, 70(2):187–197.

Doval, Y., C. Gómez-Rodríguez, y J. Vilares. 2016. Spanish word segmentation through neural language models. *Procesamiento del Lenguaje Natural*, 57:75–82.

Doval, Y., D. Vilares, y J. Vilares. 2014. Identificación automática del idioma en Twitter: adaptación de identificadores del estado del arte al contexto ibérico. En *Proc. of the Tweet Language Identification Workshop co-located with the 30th Conf. of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014*, páginas 39–43.

Doval, Y., J. Vilares, y C. Gómez-Rodríguez. 2015. LYSGROUP: Adapting a Spanish microtext normalization system to English. En *Proc. of the 1st Workshop on Noisy User-generated Text, W-NUT 2015*, páginas 99–105, Beijing, China.

Doval, Y., J. Vilares, y C. Gómez-Rodríguez. 2020. Towards robust word embeddings for noisy texts. *Applied Sciences*, 10(19):6893.

Doval, Y., M. Vilares, y J. Vilares. 2018. On the performance of phonetic algorithms in microtext normalization. *Expert Systems with Applications*, 113:213–222.

Mikolov, T., G. Corrado, K. Chen, y J. Dean. 2013. Efficient estimation of word representations in vector space. *Proc. of the International Conference on Learning Representations, ICLR 2013*, páginas 1–12.

Sumbler, P., N. Viereckel, N. Afsarmanesh, y J. Karlgren. 2018. Handling Noise in Distributional Semantic Models for Large Scale Text Analytics and Media Monitoring. *Proc. of the 4th Workshop on Noisy User-generated Text, W-NUT 2018*.

Zubiaga, A., I. S. Vicente, P. Gamallo, J. R. Pichel, I. Alegría, N. Aranberri, A. Ezeiza, y V. Fresno. 2014. Overview of TweetLID: Tweet Language Identification at SEPLN 2014. En *Proc. of the Tweet Language Identification Workshop co-located with the 30th Conf. of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014*, volumen 1228, páginas 1–11. CEUR-WS.org.