proceedings

Extended Abstract

On the Processing and Analysis of Microtexts: From Normalization to Semantics [†]

Yerai Doval [1,*] and David Vilares [2]

[1] Grupo COLE, Departamento de Informática, Escuela Superior de Ingeniería Informática, Universidade de Vigo, Campus As Lagoas, 32004 Ourense, Spain
[2] FASTPARSE Lab, Grupo LyS, Departamento de Computación, Facultade de Informática, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain; david.vilares@udc.es
[*] Correspondence: yerai.doval@uvigo.es; Tel.: +34-988-387-280

[†] Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 18 September 2018

Abstract: User-generated content published on microblogging social platforms constitutes an invaluable source of information for diverse purposes: health surveillance, business intelligence, political analysis, etc. We present an overview of our work on the field of microtext processing covering the entire pipeline: from input preprocessing to high-level text mining applications.

Keywords: microtext normalization; Language Identification; sentiment analysis; text preprocessing; text mining; semantics

1. Introduction

Extracting information from microtexts (e.g., tweets) requires the use of Natural Language Processing (NLP) techniques. Unfortunately, their performance is sensitive to the so-called texting phenomena (shortenings, substitutions, word concatenation, etc.) present in these texts. Thus, we first need to adapt the input to writing standards in a process called microtext normalization.

2. Microtext Normalization

One of the most usual approaches when implementing a microtext normalization system is decomposing it into two steps [1]: normalization candidate generation, where domain dictionaries, phonetic algorithms [2], as well as other spell checking techniques are used to obtain standard words to replace in the input text; and candidate selection, where the most likely normalized sequence according to some language model is constructed.

Notably, this approach works at the word level, as candidates are generated and selected for each word in the input text. However, word boundaries (in this case, blank spaces) are also affected by texting phenomena, hence their positioning cannot be assumed to be correct.

To address this issue we can add, as an early step in the normalization pipeline, a word segmentation subsystem that will try to normalize the positioning of word boundaries. In particular, we have experimented with character-based n-gram language models paired with a beam search algorithm, obtaining state-of-the-art results [3].

On top of this, in order to support multilingual environments such as most microblogging social platforms, it becomes essential to know in advance the language or languages in which the texts we want to normalize are written in, so that we can choose the right modules for the task. Consequently, we have added an automatic language identifier to our normalization pipeline. In this regard, we have tested and adapted well-known tools for the task [4].

The ongoing work is currently focusing on obtaining an accurate candidate selection mechanism, where language models play again a key role.

3. Sentiment Analysis

Normalization systems have many applications in downstream NLP tasks, such as Sentiment Analysis (SA) in Twitter, where the goal is to predict the polarity of a text being positive, negative or neutral. In this context, we have studied symbolic systems that compute the sentiment of sentences by taking into account their syntactic structure. The hypothesis is that syntactic relations between pairs of words are helpful to process linguistic phenomena such as negation, intensification or adversative subordinate clauses, very relevant for the task at hand. Our experiments suggest that our approach better deals with these phenomena than lexical-based systems. We also have developed machine learning models that have been evaluated in international evaluation campaigns [5,6].

These techniques are usually applied to monolingual environments, but their application to multilingual and code-switching texts, where words coming from two or more languages are used indistinctly, is gaining increasing interest [7].

Normalization and sentiment analysis might also be useful in higher level text mining applications. Political analysis, where the main goal is to use social media to estimate the popularity of politicians, is of special interest as it can be used as an alternative to traditional polls [8].

Furthermore, NLP techniques can be used in social analysis to study the cultural differences across different countries. More in particular, in [9] we explore the semantics of part-of-day nouns for different cultures in Twitter, which can be helpful to understand how different societies organize their day schedule.

Author Contributions: Y.D. conceived, designed and performed the normalization experiments; Y.D. analyzed the data from the normalization experiments; D.V. conceived, designed and performed the sentiment analysis experiments; D.V. analyzed the data from the sentiment analysis experiments; Y.D. and D.V. wrote the paper.

Acknowledgments: Research partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through projects FFI2014-51978-C2-2-R, TIN2017–85160–C2–1-R and TIN2017–85160–C2–2-R; the Spanish State Secretariat for Research, Development and Innovation (which belongs to MINECO) and the European Social Fund (ESF) under a FPI fellowship (BES-2015-073768) associated to project FFI2014-51978-C2-1-R; and by the Galician Regional Government under project ED431D 2017/12. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 714150 – FASTPARSE), which covered the costs of open access publishing. We gratefully acknowledge NVIDIA Corporation for the donation of a GTX Titan X GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Doval, Y.; Vilares, J.; Gómez-Rodríguez, C. LYSGROUP: Adapting a Spanish microtext normalization system to English. In Proceedings of the Workshop on Noisy User-generated Text, Beijing, China, 31 July 2015; pp. 99–105.
2. Doval, Y.; Vilares, M.; Vilares, J. On the performance of phonetic algorithms in microtext normalization. *ESWA* **2018**, *113*, 213–222.
3. Doval, Y.; Gómez-Rodríguez, C. Comparing Neural- and N-gram-based Language Models for Word Segmentation. *JASIST* **2018**, accepted.

4. Doval, Y.; Vilares, D.; Vilares, J. Identificación Automática del Idioma en Twitter: Adaptación de Identificadores del Estado del Arte al Contexto Ibérico. In Proceedings of the Tweet Language Identification Workshop co-located with the 30th Conference of the Spanish Society for Natural Language Processing (SEPLN 2014), Girona, Spain, 16 September 2014; Zubiaga, A., Vicente, I.S., Gamallo, P., Pichel, J.R., Alegria, I., Aranberri, N., Ezeiza, A., Fresno, V., Eds.; CEUR Workshop Proceedings, 2014; Volume 1228, pp. 39–43.
5. Vilares, D.; Doval, Y.; Alonso, M.A.; Gómez-Rodríguez, C. LYS at SemEval-2016 Task 4: Exploiting neural activation values for Twitter sentiment classification and quantification. In Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, CA, USA, 16–17 June 2016; pp. 79–84.
6. Vilares, D.; Doval, Y.; Alonso, M.A.; Gómez-Rodríguez, C. LyS at TASS 2015: Deep Learning Experiments for Sentiment Analysis on Spanish Tweets. In Proceedings of the Workshop on Sentiment Analysis at SEPLN, Alicante, Spain, 15 September 2015; Villena-Román, J., García-Morera, J., García-Cumbreras, M.A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A., Eds.; CEUR Workshop Proceedings, 2015; Volume 1397, pp. 47–52.
7. Vilares, D.; Alonso, M.A.; Gómez-Rodríguez, C. Supervised sentiment analysis in multilingual environments. *IPM* **2017**, *53*, 595–607.
8. Vilares, D.; Thelwall, M.; Alonso, M.A. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *JIS* **2015**, *41*, 799–813.
9. Vilares, D.; Gómez-Rodríguez, C. Grounding the Semantics of Part-of-Day Nouns Worldwide using Twitter. In Proceedings of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, Louisiana, 6 June 2018; pp. 123–128.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).