information

Editorial

Editorial for the Special Issue on "Natural Language Processing and Text Mining"

Pablo Gamallo [1,*] and Marcos Garcia [2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15705 Santiago de Compostela, Spain
[2] CITIC, LyS Group, Universidade da Corunha, 15001 Corunha, Spain; marcos.garcia.gonzalez@udc.gal
* Correspondence: pablo.gamallo@usc.es

Received: 5 Septermber 2019; Accepted: 5 Septermber 2019; Published: 6 Septermber 2019

Natural language processing (NLP) and Text Mining (TM) are a set of overlapping strategies working on unstructured text. On the one hand, NLP consists of linguistically motivated strategies focused on building an interpretable representation from free text. The core of NLP is linguistic analysis, including tasks such as lemmatization, PoS tagging, syntactic analysis, anaphora resolution, semantic role labeling, and so on. TM, on the other hand, can be seen as a set of Text2Data techniques for discovering and extracting relevant and salient knowledge from large amounts of unstructured text. Its main objective typically is not to understand all or even a large part of what a given speaker/writer has uttered, but rather to extract items of knowledge or regular patterns across many documents, especially Web content and social media.

Following recent advances in NLP, machine learning, deep learning, and big data, TM is now an even more valuable method for connecting linguistic theories with real-world NLP applications aimed at building organized data from unstructured text. Both hidden and new knowledge can be discovered by making use of NLP techniques and TM methods, by relying on supervised or unsupervised learning strategies within big data environments.

This special issue on "Natural Language Processing and Text Mining" aims at promoting new approaches and techniques for mining text to extract information by making use of core NLP strategies. The 8 contributions accepted for the current special issue address many topics in information extraction based on linguistic strategies: for instance, open information extraction, named entity recognition, or event extraction. The multilingual factor is very important, as the papers accepted cover different languages, such as Chinese, Uyghur, Basque, Portuguese, Indi, or French. The application domains are also varied, including social media, forensic reports, or financial and biomedical documents.

In the paper "Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents" [1], the authors deal with the problem of applying named entity recognition (NER) models on different domains. More precisely, they train NER models with deep learning approaches and these models are applied on two specific domains: biomedical and financial ones. In their experiments, the most successful language model turns out to be BERT (Bidirectional Encoder Representations from Transformers) [2], an attention method that learns contextual relations between words (or sub-words) in a text.

In the article "Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship" [3], the authors analyze over 3000 reports of missing persons in São Paulo city during the Brazilian dictatorship. For this purpose, they make use of unsupervised methods of information extraction in Portuguese that let them identify named entities and terminology. In addition, this analysis allows the authors to observe terminological patterns which are relevant for people identification. They claim that their text mining strategy could assist researchers working in pattern search among autopsy reports.

In the study "Multilingual Open Information Extraction: Challenges and Opportunities" [4], the authors analyze the approaches on multilingual open information extraction by exploring the state-of-the-art systems for English and Portuguese. They explore the benefits of multilingual approaches to allow transfer knowledge between languages. More precisely, they carry out an experiment on parallel corpora and relation extraction systems to improve the effectiveness of open information extraction.

In the article "Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case" [5], the authors study the linguistic and social aspects of the behavior of young and adult people. This study is carried out by of means of the linguistic analysis of the content of their tweets and the social relations that arise from them. More precisely, they gather about 10 million tweets from more than 8000 users and used NLP techniques, including topic modelling, to process the corpus and find the social parameters. They work on a low-resourced language: Basque language.

The paper "Event Extraction and Representation: A Case Study for the Portuguese Language" [6] describes an event extraction system from Portuguese documents which is based on a pipeline of specialized natural language processing tools: namely part-of-speech tagging, dependency parsing, named entity recognition, semantic role labeling, and a knowledge extraction module. The developed system is evaluated with a corpus of Portuguese texts and compared with the existing tool LinguaKit [7].

In the article "Spelling Correction of Non-Word Errors in Uyghur–Chinese Machine Translation" [8], the authors describe three spelling correction methods based on machine translation by considering that most spelling problems are caused by out-of-vocabulary issues. The aim is to improve the quality of Uyghur–Chinese machine translation, by solving the out-of-vocabulary problem caused by Uyghur spelling errors in Uyghur–Chinese machine translation.

The paper "An Improved Word Representation for Deep Learning Based NER in Indian Languages" [9] describes a named entity recognition system based on deep learning approaches for Indian languages. For this purpose, they use a novel combined word representation, including several levels of embeddings, namely character-level, word-level, and affix-level embeddings.

Finally, the article "Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities" [10] describes how to enrich the Istex project by allowing queries on named entities contained in scientific papers. More precisely, this paper describes the implementation of a text mining tool which uses the recognition of named entities for the open access resource Istex, a French database of twenty million scientific papers.

As a whole, the articles published in this special issue are an interesting contribution at the intersection between NLP and TM, giving special relevance to the analysis and application of state-of-the-art technologies in different languages.

References

1. Francis, S.; Landeghem, J.; Moens, M.F. Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents. *Information* **2019**, *10*, 248. [CrossRef]
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
3. Martin-Rodilla, P.; Hattori, M.; Gonzalez-Perez, C. Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship. *Information* **2019**, *10*, 231. [CrossRef]
4. Claro, D.; Souza, M.; Xavier, C.; Oliveira, L. Multilingual Open Information Extraction: Challenges and Opportunities. *Information* **2019**, *10*, 228. [CrossRef]

5. Fernandez de Landa, J.; Agerri, R.; Alegria, I. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information* **2019**, *10*, 212. [CrossRef]
6. Quaresma, P.; Nogueira, V.; Raiyani, K.; Bayot, R. Event Extraction and Representation: A Case Study for the Portuguese Language. *Information* **2019**, *10*, 205. [CrossRef]
7. Gamallo, P.; Garcia, M.; Piñeiro, C.; Martinez-Castaño, R.; Pichel, J.C. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In Proceedings of the 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018; pp. 239–244. [CrossRef]
8. Dong, R.; Yang, Y.; Jiang, T. Spelling Correction of Non-Word Errors in Uyghur–Chinese Machine Translation. *Information* **2019**, *10*, 202. [CrossRef]
9. AP, A.; Mary Idicula, S. An Improved Word Representation for Deep Learning Based NER in Indian Languages. *Information* **2019**, *10*, 186. [CrossRef]
10. Maurel, D.; Morale, E.; Thouvenin, N.; Ringot, P.; Turri, A. Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities. *Information* **2019**, *10*, 178. [CrossRef]