Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 119-122 recibido 18-10-2019 revisado 12-01-2020 aceptado 15-01-2020

Adverse Drug Reaction extraction on Electronic Health Records written in Spanish

Extracción de Reacciones Adversas a Medicamentos en Historias Clínicas Electrónicas escritas en español

Sara Santiso González
IXA group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1, 20018, Donostia
sara.santiso@ehu.eus

Abstract: PhD thesis on Language Analysis and Processing written by Sara Santiso González at the University of the Basque Country (UPV/EHU) under the supervision of Dr. Arantza Casillas (Department of Electricity and Electronic) and Dr. Alicia Pérez (Department of Computer Languages and Systems). The thesis defense was held on June 13, 2019 and the members of the commission were Dr. Raquel Martinez (President, National Distance Education University (UNED)), Dr. Arantza Díaz de Ilarraza (Secretary, University of the Basque Country (UPV/EHU)) and Dr. Lluis Padró (Vocal, Technical University of Catalonia (UPC)). The thesis obtained excellent grade with Cum Laude mention.
Keywords: Adverse Drug Reactions, Electronic Health Records, Text mining, Supervised machine learning

Resumen: Tesis doctoral en Análisis y Procesamiento del Lenguaje defendida por Sara Santiso González en la Universidad del País Vasco (UPV/EHU) y realizada bajo la dirección de las doctoras Arantza Casillas Rubio (Departamento de Electricidad y Electrónica) y Alicia Pérez Ramírez (Departamento de Lenguajes y Sistemas Informáticos). La defensa de la tesis tuvo lugar el 13 de junio de 2019 ante el tribunal formado por los doctores Raquel Martinez (Presidenta, Universidad Nacional de Educación a Distancia (UNED)), Arantza Díaz de Ilarraza (Secretaria, Universidad del País Vasco (UPV/EHU)) y Lluis Padró (Vocal, Universidad Politécnica de Cataluña (UPC)). La tesis obtuvo la calificación de sobresaliente con mención Cum Laude.
Palabras clave: Reacciones Adversas a Medicamentos, Historias Clínicas Electrónicas, Minería de textos, Apredizaje automático supervisado

1 Introduction

This thesis was developed on the IXA group of the University of the Basque Country (UPV/EHU) and is related with the extraction of Adverse Drug Reactions (ADRs).

An ADR is defined by the World Health Organization (WHO) as 'a response to a medicine which is noxious and unintended, and which occurs at doses normally used in man'. The WHO informed about the importance of reporting ADRs to understand and treat the diseases caused by drugs and, as a result, improve the patients care. However, ADRs are still heavily under-reported, which makes their prevention difficult. This was the **motivation** to automatically extract ADRs on Electronic Health Records (EHRs). Given that information stored digitally by the hospitals is growing, Natural Language Processing (NLP) techniques can be used to create a system that helps the doctors to analyze the ADRs of the patients in a given EHR, facilitating the decision making process and alleviating the work-load. As a consequence, the patients' health could improve and the pharmaco-surveillance service would be informed about the detected ADRs. The ADR extraction was defined as a relation extraction task. That is, the aim is to detect ADR relations between the entities (drugs and diseases) recognized in a given text. For the ADR extraction developed in this work, we

ISSN 1135-5948. DOI 10.26342/2020-64-14

distinguished two steps that were developed using a pipeline approach:

1. Medical Entity Recognition (MER) to find "drug" entities and "disease" entities. The "drug" entity encompasses either a brand name, a substance or an active ingredient and the "disease" entity encompasses either a disease, a sign or a symptom.
2. ADR detection to discover the relations between "drug" entities and "disease" entities that correspond to ADRs. The "drug" entity would be the causative agent and the "disease" entity would be the caused adverse reaction.

In the ADR extraction process, we had to overcome some challenges that make this supervised classification task difficult. On the one hand, the ADRs are minority relations because generally the drug and the disease are either unrelated or related as treatment and, thus, the ADRs are rare cases. On the other hand, the EHRs show multiple lexical variations. In addition, our EHRs are written in Spanish whereas the majority of biomedical NLP research has been done in English.

In this way, we tackled some drawbacks present in state-of-the-art works: the class imbalance, the lexical variability and the few resources and tools to apply NLP in the medical domain for Spanish and other languages different to English.

The main **objective** of this work is the creation of a model able to detect automatically ADRs in EHRs written in Spanish. This, in turn, encompasses the sub-objectives stated below:

- Detect ADRs by discovering relations between the causative drug and the caused diseases.

  The aim is to detect drug-disease pairs related as ADRs and not only the disease caused by the drug. Indicating explicitly the entities involved in an ADR can result more useful for their study.

- Discover approaches to overcome the class imbalance.

  Given that ADRs are rare events, it is frequent to find the class imbalance problem. Machine learning algorithms tend to expect balanced class distributions and learning the minority class is difficult for them. For this reason, our intention is to explore different techniques that could help to tackle this issue improving the ADR detection or find approaches that could be robust against imbalanced distributions of the class.

- Discover robust representations to cope with the lexical variability and the data sparsity.

  This is a challenge goal due to two factors. First, the EHRs are written during consultation time and each doctor uses different terms or expressions, producing lexical variations. Second, due to confidentiality issues, there is a lack of available EHRs. Then, our intention is to explore different representations to make the most of the annotated corpus.

2 Thesis Overview

This thesis was organized in eight chapters.

Chapter 1 explains the motivation to develop the ADR extraction and the framework. It also presents the objectives to achieve together with the main research question to address.

Chapter 2 makes a review of the works related with the ADR extraction task. It focuses on the definition of ADR extraction, the techniques and features employed for the ADR classification, the corpora and the evaluation schemes used for ADR extraction.

Chapter 3 presents the corpora employed in this work (IxaMed-GS, IxaMed-CH, IxaMed-E). Furthermore, it describes the schemes and metrics employed for the evaluation of our systems.

Chapter 4 describes the features employed to create the symbolic characterizations of the ADR events, our first approach. It presents the Random Forest classifier used for ADR detection of intra-sentence as well as inter-sentence ADRs. It also explains the approaches explored to tackle the class imbalance (Santiso et al., 2014; Santiso et al., 2016; Casillas et al., 2016b; Casillas et al., 2016a; Santiso, Casillas, and Pérez, 2019). With this approach, the best results are a precision of 34.0, a recall of 59.3 and an f-measure of 43.2.

Chapter 5 explains the dense characterizations created from embeddings that were used together with the Random Forest classifier overcoming the class imbalance, our second approach. Moreover, it proposes dif-

ferent smoothing techniques that were applied to the dense representations in order to improve the proximity between semantically related words (Santiso, Pérez, and Casillas, 2019b). With this approach, the best results are a precision of 47.4, a recall of 66.7 and an f-measure of 55.4.

Chapter 6 explains the neural networks used for ADR detection, Joint AB-LSTM networks, as our third approach. It includes the core-features employed to infer the dense representations. It also presents the techniques explored to overcome the class imbalance suited for neural networks (Santiso, Pérez, and Casillas, 2019a). With this approach, the best results are a precision of 72.4, a recall of 71.4 and an f-measure of 71.9.

Chapter 7 discusses the results obtained with the best performing approach, using slightly different corpora and incorporating the automatic detection of medical entities. Until now, we have just focused on the ADR detection step and we have tried different representations and classifiers. The best results are obtained with the higher corpus, yielding a precision of 74.4, a recall of 76.0 and an f-measure of 75.2.

Chapter 8 gives the final conclusions, which include the response to the research questions and the main contributions. It explains the future lines of work regarding the ADR extraction. It also shows the publications related to this work.

In addition, it includes three appendices.

Appendix A explains the two approaches explored to detect negated entities automatically. These negated entities are used to discard negative ADR candidates (Santiso et al., 2017; Santiso et al., 2019).

Appendix B briefly explains some experiments developed to detect medical entities automatically. These entities are those used to observe the influence of MER step on ADR detection.

Appendix C gives detailed results of the experiments developed in Chapter 5 for ADR detection using dense representations and the Random Forest classifier.

3 Contributions

The main **contribution** of this work is that the ADR extraction was developed using EHRs written in Spanish. To the best of our knowledge, for ADR extraction in texts written in Spanish, we are the first employing EHRs. Other contributions derived from the tasks carried out during this work are:

- Combination of approaches to tackle the high class imbalance.

  We made a step ahead in the development of NLP methods that deal with ADR extraction defined as relation extraction task. As a first approach we tackled both inter- and intra-sentence ADR extraction, even though the mainstream in the related works just focused on intra-sentence relations. In this context, inference algorithms should be suited to cope with the challenge of an extremely high class imbalance. Although the imbalance problem diminishes considerably in intra-sentence scenarios, we explored classical approaches to tackle the class imbalance (sampling, cost-sensitive learning, ensemble learning, one-class classification) in the context of inter- and intra-sentence ADR extraction. We observed that the combination of them, precisely sampling and cost-sensitive learning, was beneficial in our framework.

  Besides, in an attempt to discard non-ADR instances and alleviate the class imbalance, we also tried negation detection. We developed two ways of detecting negated medical entities in EHRs: an adaptation of the NegEx tool and a Conditional Random Fields algorithm using dense characterizations. We corroborated, however, that class imbalance can be tackled in intra-sentence ADR extraction, while there is room for improvement in inter-sentence relation extraction.

- Mechanisms to deal with lexical variability.

  NLP in the medical domain dealing with EHRs has, among others, the challenge of high lexical variability (large specialized vocabularies, non-standard abbreviations, misspellings, etc.) and lack of available corpora. Quantitatively, there is a reflect of the lexical variability in the remarkable ratio of Out-Of-Vocabulary elements. To cope with this issue it results crucial to propose not only competitive inference algorithms but also robust characterizations of the instances.

Throughout this work we analyzed two classification techniques (Random Forest, Joint AB-LSTM) and two representations (symbolic, dense). We experimentally corroborated that context-aware embeddings (dense representations created taking into account the embeddings of the context-words) are useful to preserve the lexical nuances in this domain. In addition, to alleviate the influence that the lack of training samples might have in the quality of the inferred dense representations, we proposed the use of smoothing techniques. Smoothing helps to avoid superficial variations and, hence, makes different (but close) points in the space to be equivalent.

Moreover, we observed that dense spaces of lemmas also helped to tackle the lexical variability. In fact, lemmatization was particularly effective in the neural networks used for ADR extraction.

- Tolerance to external noise.

  We exposed the ADR extraction system to two types of noise. On the one hand, we assessed the impact of corpora from slightly different sources (different hospitals with different services or specializations). On the other hand, we analyzed the influence of miss-recognized medical entities into the ADR detection step leading to a fully automatic ADR extraction system. We corroborated that the Joint AB-LSTM is able to cope with these types of noise although, naturally, there is a small decrease in its performance due to the missed entities involved in the ADR pairs.

Acknowledgements

The author would like to thank the staff of the Pharmacy and Pharmacovigilance services of the Galdakao-Usansolo and Basurto hospitals. This work was partially funded by the Spanish Ministry of Science and Innovation (PROSAMED: TIN2016-77820-C3-1-R) and the Basque Government (BERBAOLA: KK-2017/00043, Predoctoral Grant: PRE 2018 2 0265).

References

Casillas, A., A. Díaz de Ilarraza, K. Fernandez, K. Gojenola, M. Oronoz, A. Pérez, and S. Santiso. 2016a. IXAmed-IE: Online medical entity identification and ADR event extraction in Spanish. In *2016 IEEE International Conference on Bioinformatics and Biomedicine*, pages 846–849.

Casillas, A., A. Pérez, M. Oronoz, K. Gojenola, and S. Santiso. 2016b. Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications*, 61:235–245.

Santiso, S., A. Casillas, and A. Pérez. 2019. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Informatics Journal*, 25(4):1768–1778.

Santiso, S., A. Casillas, A. Pérez, and M. Oronoz. 2017. Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 177–188.

Santiso, S., A. Casillas, A. Pérez, and M. Oronoz. 2019. Word embeddings for negation detection in health records written in Spanish. *Soft Computing*, 23(21):10969–10975.

Santiso, S., A. Casillas, A. Pérez, M. Oronoz, and K. Gojenola. 2014. Adverse drug event prediction combining shallow analysis and machine learning. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 85–89.

Santiso, S., A. Casillas, A. Pérez, M. Oronoz, and K. Gojenola. 2016. Document-level adverse drug reaction event extraction on electronic health records in Spanish. *Procesamiento del Lenguaje Natural*, 56(0):49–56.

Santiso, S., A. Pérez, and A. Casillas. 2019a. Exploring joint ab-lstm with embedded lemmas for adverse drug reaction discovery. *IEEE Journal of Biomedical and Health Informatics*, 23(5):2148–2155.

Santiso, S., A. Pérez, and A. Casillas. 2019b. Smoothing dense spaces for improved relation extraction between drugs and adverse reactions. *International journal of medical informatics*, 128:39–45.