Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 155-158 recibido 23-03-2016 revisado 06-04-2016 aceptado 27-04-2016

Deteami research-transference project: natural language processing technologies to the aid of pharmacy and pharmacosurveillance

Proyecto de transferencia tecnológica Deteami: tecnologías de procesamiento del lenguaje natural para la ayuda en farmacia y en farmacovigilancia

A. Casillas[(1)]∗, **A. Díaz de Ilarraza**[(1)], **K. Gojenola**[(1)],
L. Mendarte[(2)], **M. Oronoz**[(1)], **J. Peral**[(3)], **A. Pérez**[(1)]
[(1)]IXA research group (UPV-EHU);
∗arantza.casillas@ehu.eus
[(2)]Basurto University Hospital;
[(3)]Galdakao-Usansolo Hospital

Abstract: The goal of the Deteami project is to develop tools that make clinicians aware of adverse drug reactions stated in electronic health records of the clinical digital history.The records produced in hospitals are a valuable though nearly unexplored source of information among others due to the fact that are tough to get due to privacy and confidentiality restrictions. To leverage the clinicians work of reading and analyzing the health records looking for information about the health of the patients, in this project we explore the records automatically, identify among others disorder and drug entities, and infer medical information, in this case, adverse drug reactions. In this project a research-framework was settled with the Galdakao-Usansolo and Basurto Hospitals from Osakidetza (the Basque Health System). Osakidetza provided both the texts and the final user feedback, as well as, specialists that annotate the corpora, an in this way, we obtained a gold-standard.
Keywords: Technological transference, clinical text mining, entity recognition

Resumen: El objetivo del proyecto Deteami es el desarrollo de herramientas para ayudar al personal clínico a identificar reacciones adversas a medicamentos en informes médicos electrónicos de la historia clínica digital. Los informes que se generan en los hospitales son una valiosa fuente de información aún no debidamente explotada debido principalmente a restricciones de privacidad y confidencialidad. Con el objetivo de aliviar el trabajo del personal clínico que se dedica a leer y analizar los informes médicos buscando información sobre la salud de los pacientes, en este proyecto analizamos automáticamente los informes, identificamos entre otras entidades que describen enfermedades y medicamentos, y finalmente, inferimos información médica; en este caso, reacciones adversas a medicamentos. En este proyecto hemos establecido un marco de colaboración con los hospitales de Galdakao-Usansolo y Basurto pertenecientes a Osakidetza (Servicio Vasco de Salud). Osakidetza participa mediante la provisión de los textos y retroalimentándo el trabajo técnico con su experiencia, así como expertos que anotan el corpus para la obtención del gold-standard.
Palabras clave: Transferencia tecnológica, minería de textos clínicos, reconocimiento de entidades

1 Introduction

Typically, clinical documentation is produced in natural language on a free text basis, basically without or with little structure. From the research point of view, clinical text pro-

∗ This work was partially supported by the Spanish Ministry of Science and Innovation (EXTRECM: TIN2013-46616-C2-1-R, TADEEP: TIN2015-70214-P) and the Basque Government (DETEAMI: M inistry of Health 2014111003, IXA Research Group of type A (2010-2015), ELKAROLA: KK-2015/00098).

ISSN 1135-5948

cessing dates from the early eighties (Friedman et al., 1983). There are easy understandable examples of the interest of natural language processing in this domain such as in (Taira, Soderland, and Jakobovits, 2001) where they proposed an approach to extract valuable information in the framework of radiology. The information extracted can be used as a decision support system.

The research goes ahead with evidences of the potential use of these techniques, while it is steadily being implemented in the hospitals. The goal of this project is twofold: on the one hand, to develop and transfer NLP technologies to the clinical domain and, on the other hand, to extract automatically adverse drug reactions from electronic health records. According to the World Health Organization, an Adverse Drug Reaction (ADR) "is a response to a drug which is noxious and unintended, and which occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease...".

The detection of ADRs is a key issue for the pharmacy and pharmacovigilance services that attempt to prevent medicine-related adverse effects in humans, in order to promote patient safety, but also the rational use of medicines. (Henriksson et al., 2015) state that ADRs cause the 3-5% of hospital admissions world-wide.

While there is some research in the clinical language processing, mainly in the language used in journals, there are still few applications integrated in the health services and that deal with the language used in patient's records. One of the reasons, is the difficulty to access information about patients due to privacy and confidentiality restrictions. For the Deteami project, although all the records were collected without any private data about the patients (without any name, age, address...) three ethical committees were passed and a confidentiality agreement was signed between the UPV-EHU University and Osakidetza.

From the point of view of computational linguistics, adverse drug reactions can be represented as a pair of entities (drug, disease) in which the drug was the causative agent of the disease. Hence, the task can be formulated as finding cause-effect events from a drug to a disease.

To work on this task, a set of real data was collected and manually annotated by two experts. This corpus serve as i) a gold-standard for the development of FreeLing-Med, a clinical entity recognition tagger and ii) training and evaluation sets for an ADR event retrieval system. The ADR retrieval tool shall serve to the aid of the Pharmacy and Surveillance services in their task of detecting adverse drug reactions. Besides, it will be an attempt to carry out technological transference. To do so, the prototype shall be validated by virtue of experts from the two involved hospitals. The ultimate goal of such a prototype would be to contribute in the early diagnosis of diseases, with its impact on the wellbeing of the society.

An added value to this project is that it shall be developed in Spanish, while there are few tools available for the clinical domain.

2 *Materials and Methods*

Our first purpose has been to acquire a corpus representative of the ADR event detection task and get it annotated by experts (clinicians, pharmacists, etc.) following the process in (Oronoz et al., 2015). In parallel, we are developing FreeLing-Med (Oronoz et al., 2013). Together with this, we mean to use NegEx adapted to this domain in order to detect negation and hence, help discarding negated events early. While we focus on the semantic tags provided by FreeLing-Med that correspond to medical entities, the remaining information would feed the subsequent stages, mainly the event retrieval system. Figure 1 summarizes the processes involved in the Deteami project, each of which is explained in the following sub-sections. As a result, the system extracts structured information from free text in the clinical domain in Spanish.

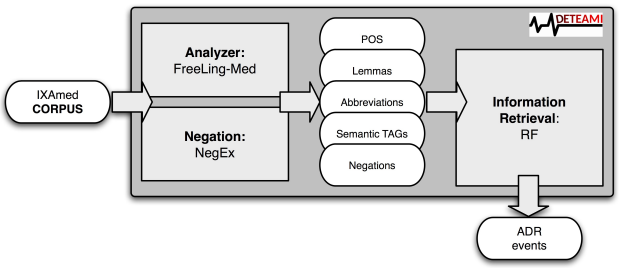

Figure 1: Deteami project overview.

2.1 Corpus acquisition

While there are several efforts in the literature that aim at the clinical domain, due to the lack of data, the majority focus on clinical journals and abstracts from PubMed. There are a few examples in the literature aiming at clinical text notes, the most of them for English, but there is also a remarkable one for Swedish (Dalianis et al., 2012). For the Spanish language this is the first corpus composed of real electronic records, we are aware of. In the Deteami project, the two hospitals involved, the Galdakao-Usansolo and Basurto Hospitals, contribute with medical records. Each of the hospitals make use of different platforms to store the data and most of the records are fully unstructured. At the IXA group, we collect and classify the records, we properly encode them and analyze them with FreeLing-Med. After that, we filter the terms indicating disorders and drugs, and we obtain the information that is shown to the human annotators in the "Brat Rapid Annotation Tool" format. In this way, the expert annotators following the guidelines can create the gold-standard.

2.2 Analyzer

A key issue in the detection of (drug, disease) events is to appropriately recognize clinical entities (particularly: drugs, signs, symptoms, substances, etc.).

There are a few examples of analyzers well adapted for the clinical domain, an example of them is GENIA (Tsuruoka et al., 2005), an analyzer for English. There is a version of MetaMap Transfer (MMTx) for Spanish (Carrero et al., 2008) that translates into English the text in Spanish by means of Google Translate and next applies the English version of MMTx in order to extract the medical entities that appear in the English SNOMED-CT. The arising question is the impact of error propagation.

In this project, instead, we adapted a general purpose linguistic analyzer FreeLing (L. Padró, 2011) to the medical domain. We enhanced the dictionaries with samples from the manually annotated corpus, terminology from standard medical ontologies (SNOMED CT, CIE-9-CM, ATC classification...), abbreviations within the medical domain and also tackle ambiguity. This tool analyzes the records and obtains valuable features such as tokens, lemmas, POS tags and semantic tags (disorder, body part...). As we said before, some of these features are shown to the annotators, but also serve for the event detection classifier.

2.3 Negation and speculation

Negation detection is crucial to discard early several potential ADR pairs. The negation in the medical domain has been tackled with two approaches:

1. Rule-based: NegEx (Chapman et al., 2001) is an outstanding toolkit in this area. It consists of a set of regular expressions built automatically from trigger phrases (pre-, post- and pseudo-negation) that can negate a clinical finding on which the negation is focused. It provides competitive results with 94.5% precision in clinical abstracts. The tool was adapted for Spanish as well (Costumero et al., 2014).

2. Machine Learning: there are alternatives that mean to infer negation patterns, such as (Averbuch et al., 2004). While this technique is language-independent, it was tested for English with an F1-score of 99.7%.

We are studying the approach that fits better the kind of texts we have, and we started using NegEx.

2.4 Information retrieval

The approaches used to extract information from clinical texts can be divided into two main groups: those that make use of rules and those based on machine learning. Rules, in comparison with inferred classifiers, tend to have better precision while lower recall. Our idea is to combine both methods in an hybrid system. First, we aim to use Kybots or abstract schemas defined in the Kyoto project (Vossen et al., 2008), to define adverse drug event patterns. Second, for event retrieval, we are making use of random forests, a supervised ensemble classification strategy.

3 Concluding remarks

In this project two hospitals attached to Osakidetza and a research group with experience in NLP are cooperating in an attempt to develop and transfer technological solutions based on NLP to the clinical domain,

and hence, fill the technological gap in clinical text mining in Spanish. The project is carried out for real clinical texts that lack of structure and so as to support language technologies in Spanish. This is the first year out of three of the Deteami project. Some of the mentioned tools have been already developed but are being improved. That is the case of the analyzer FreeLing-Med, currently in use but under development in order to improve i) the identification of non standard or local medical language and ii) the disambiguation of semantic tags and abbreviations. The detection of negation and speculation is in its early stage. Some experiments in the retrieval of adverse drug events have been carried out, but need to solve the problem of having an unbalanced set of disorder-drug pairs (there are very few pairs indicating ADRs and many indicating prescriptions). All the analyzers will improve with the manual annotation of new medical records. The two expert clinicians that are already annotating the corpus will continue with this task and reviewing the results automatically obtained for one more year. In this way, both the health system and the NLP researching group will benefit for this fruitful collaboration.

References

Averbuch, M., T.H Karson, B. Ben-Ami, O. Maimon, and L. Rokach. 2004. Context-sensitive medical information retrieval. *MedInfo*, page 282.

Carrero, F.J., J. Carlos Cortizo, J.M. Gómez, and M De Buenaga. 2008. In the development of a spanish metamap. In *Proceedings of the 17th CIKM conference*, pages 1465–1466. ACM.

Chapman, W.W., W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Costumero, R., F. Lopez, C. Gonzalo-Martín, M. Millan, and E. Menasalvas. 2014. An Approach to Detect Negation on Medical Documents in Spanish. In *Brain Informatics and Health*. Springer, pages 366–375.

Dalianis, H., M. Hassel, A. Henriksson, and M. Skeppstedt. 2012. Stockholm EPR corpus: A clinical database used to improve health care. In *Swedish Language Technology Conference*, pages 17–18. Citeseer.

Friedman, C., N. Sager, E.C. Chi, E. Marsh, C. Christenson, and M. S. Lyman. 1983. Computer Structuring of Free-Text Patient Data. In *Symposium on Computer Applications in Medical Care*, pages 688–691. American Medical Informatics Association.

Henriksson, A., M. Kvist, H. Dalianis, and M. Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics*, 57:333–349.

L. Padró. 2011. Analizadores Multilingües en freeling. *Linguamatica*, 3(2):13–20, December.

Oronoz, M., A. Casillas, K. Gojenola, and A. Pérez. 2013. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. *Lecture Notes in Computer Science*, 8259:536–547.

Oronoz, M., K. Gojenola, A. Pérez, A. Díaz de Ilarraza, and A. Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56(0):318 – 332.

Taira, R., S. Soderland, and R. Jakobovits. 2001. Automatic structuring of radiology free-text reports 1. *Radiographics*, 21(1):237–245.

Tsuruoka, Y., Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392.

Vossen, P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, and J. VanGent. 2008. KYOTO: a System for Mining, Structuring and Distributing Knowledge across Languages and Cultures. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 23–37, may.