On-Line Analysis of Clinical Texts in Spanish

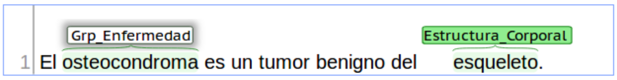
A. Atutxa, A. Casillas, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, A.Pérez, O.Perez-de-Viñaspre

Ixa Taldea, UPV/EHU
ixa.si.ehu.es
maite.oronoz@ehu.eus

1. Introduction and objectives

The scientific and industrial community is showing an increasing interest in the automatic text processing of biomedical documents, electronic health records or drug labels due to the rich information contained in this type of documents. Clinicians can benefit from the mining of the medical information in clinical texts, for instance by using decision support tools.

In this work we present an on-line prototype built with the aim of assisting information extraction from electronic health records in Spanish. The prototype (freely available at http://ixa2.si.ehu.es/freeling2openbrat/) consists of a web-interface that allows the user to enter a piece of clinical text. This piece of texts is next analyzed by a linguistic analyzer called FreeLing-Med so as to extract medical entities (see Figure 1) and finally the text is shown in a friendly interface using the Brat Rapid Annotation Tool [1]. The prototype is in its early stage of testing and validation by experts from the Galdakao-Usansolo and Basurto hospitals from the Basque Sanitary System.



1 El osteocondroma es un tumor benigno del esqueleto.

Figure 1: *Output of freeling2openbrat.*

2. Materials and Methods

2.1. FreeLing-Med

FreeLing-Med [2] is an adaptation of an open analyzer, FreeLing [3], suitable, among other languages, for Spanish. FreeLing-Med incorporates ontologies and dictionaries to identify medical entities, such as the following ones: the distribution in Spanish of the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), the International Statistical Classification of Diseases and Related Health Problems in its 10th version (ICD-10), a database with brand names of drugs, a dictionary of medical abbreviations, non standard medical terms provided by experts in pharmacy, and so on. Being FreeLing-Med a linguistic analyzer, it produces tokenization, syntactic analysis and also semantic annotation. The semantic annotation adds a tag to identify the type of medical entity for each term added to the dictionaries. The semantic tags of SNOMED-CT identify whether an entity describes a disorder, a finding, a substance, a product, a body structure or any class of concept that appears in the hierarchies of SNOMED-CT. Gathering the findings and disorders together (as diseases) and drug brand names and substances (as drugs) FreeLing-Med achieved an F-score of 36.35 and 59.91 on diseases and drugs respectively. As a result, given free text FreeLing-Med enables clinical entity recognition.

2.2. Brat Rapid Annotation Tool

The result of FreeLing-Med is presented in a human-friendly interface called the Brat environment. Medical entities as drug brand names, diseases, body parts or qualifiers are marked in the output of freeling2openbrat.

3. Conclusions

The system is straightforward to use, simple and friendly. New utilities can be easily added as for example a summary of all the entities mentioned in the document together with their frequency. Using statistics extracted by the analysis of a set of texts, very valuable data can be obtained: the most frequent treatment for a specific disorder, the procedure used with higher frequency to treat a disorder etc. The output of the tool has already been used in a non graphical manner to work in the automatic detection of Adverse Drug Reactions with good results [4] (F-score of 27.2 on the positive class and a weighted average F-score of 88.5).

This work makes a step ahead on the development of Natural Language Processing tools adapted for the clinical domain in Spanish.

Acknowledgements

This work has been partially funded by the Spanish ministry (TADEEP: TIN2015-70214-P, PROSA-MED: TIN2016-77820-C3-1-R) and the Basque Government (DETEAMI: 2014111003; ELKAROLA: KK-2015/00098).

References

[1] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou and J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*

[2] M. Oronoz, A. Casillas, K. Gojenola, and A. Pérez, "Automatic annotation of medical records in Spanish with disease, drug and substance Names,". *Lecture Notes in Computer Science*, vol. 8259, pp. 536–547, 2013.

[3] L. Padró, S. Reese, E. Agirre, and A. Soroa, "Semantic Services in Freeling 2.1: WordNet and UKB," in *Global Wordnet Conference*, Mumbai, India, 2010.

[4] A. Casillas., A. Pérez, M. Oronoz, K. Gojenola, S. Santiso (2016) Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications*, Volume 61, 1 November 2016, Pages 235-245