A web prototype for detecting chemical compounds and drugs

Daniel Sánchez-Cisneros[1], Sara Lana-Serrano[2], Isabel Segura-Bedmar[1], Leonardo Campillos[3], Paloma Martínez[3]

[1]ComputerScienceDepartment, Universidad Carlos III de Madrid, Spain
{dscisner, isegura,pmf}@inf.uc3m.es
[2]Universidad Politécnica de Madrid, Spain
slana@diatel.upm.es
[3]Universidad Autónoma de Madrid
leonardo.campillos@uam.es

Abstract. This paper introduces a web prototype for named entity recognition of chemical compounds and drugs. The tool is based on a system developed to participate in the ChemDNER task organized as part of Biocreative 2013 workshop. The system combines the ChemSpot tool as well as a set of semantic-based rules, which were defined according to the guidelines provided to task participants. The prototype is available at http://multimedica.uc3m.es:8080/biocreative2013demo/

Keywords: Drug named entity recognition, information extraction

1 Introduction

Most research on named entity recognition (NER) in the biomedical domain are based on dictionary based methods and Supervised Machine Learning (SML) methods. The main problems with the former approach are their domain dependency and their inability to recognize terms not included in the dictionaries. Machine learning techniques build classification models based on annotated corpus and produce the best results [1], although they require annotated corpora.

Current trends try to develop hybrid systems that combine best of two approaches. In this work we present a prototype that combines existing systems such as ChemSpot [2] and Metamap [3] with gazetteers extracted from biomedical resources such as

MeSH[1], DrugBank[2], Wikipedia[3] and ChEBI [4]. Lastly, based on error analysis of the development set, we defined a set of semantic rules to detect false negatives and discard false positives generated by the previous processes. In this paper, we present a web tool designed on this system. The tool allows user to introduce a text and then detect chemical compounds and drugs occurring in the text.

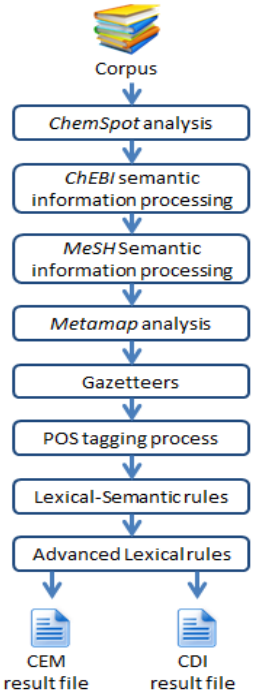

Fig. 1. Pipeline architecture

2 Description of the prototype

Figure 1 shows the pipeline architecture of the prototype. In a first step, texts are processed by the ChemSpot tool. This tool is able to identify mentions of chemicals. The next three processes are responsible for extracting semantic knowledge from the CheBI ontology, the MeSH vocabulary and the UMLS Metathesaurus (using the MetaMap tool). In particular, the semantic features used are: MeSH semantic types, MeSH type, MeSH_TreeNumbers, UMLS semantic types, and ancestors from ChEBI

[1] http://www.nlm.nih.gov/mesh/meshhome.html
[2] http://www.drugbank.ca
[3] http://wikipedia.org

by traversing recursively the relationships: *is_a*, *has_role*, *is_conjugate_acid_of* and *is_conjugate_base_of*. In the next phase, a gazetteer tagger implemented in the GATE[4] environment is used. Based on error analysis of the development set, a set of 27 gazetteers with more than 340,000 entries have been compiled to process texts in order to rule out false positive instances and to annotate false negative instances that were not recognized in the previous steps. The sixth module is the ANNIE PoS tagger included in GATE. Pos tags are used to discard some instances as well as to define the rules used in the last two steps to classify the entities according to PoS tagging, affix processing and multiword processing. More information about the processes and resources used can be found at [5].

Figure 2 shows a screenshot of the web tool. The tool allows users to write a text to be processed by the system. As result of the processing, chemical compounds and drugs appear highlighted in text. Also, the identified chemical compounds are linked to the ChEBI database.

Fig. 2. A web tool for identifying chemical compounds and drugs.

4 http://gate.ac.uk

The system was evaluated on the test dataset provided by the BioCreative IV (CHEMDNER 2013 task[5]). It was able to recognize chemical and drug named entities with an F-measure of 0,594 over Chemical Entity Mentions (CEM) evaluation. As future work, we plan to conduct an evaluation with users to measure the usability of our tool.

Acknowledgments

TThis work was supported by the EU project TrendMiner [FP7-ICT287863], by the project MultiMedica [TIN 2010-20644-C03-01)], and by the research network MA2VICMR [S2009/TIC-1542].

References

1. Rocktschel, T., Huber, T., Weidlich, M., Leser, U.: WBI-NER: The impact ofdomain-specific features on the performance of identifying and classifying mentions of drugs. Proceedings of SemEval 2013, pp. 356-363, (2013).
2. Rocktschel, T., Weidlich, M., Leser, U.: Chemspot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), pp. 1633-1640, (2012).
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium, American Medical Informatics Association, pp. 17-21, (2001)
4. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research, 36, pp. 344-350, (2008).
5. Lana-Serrano, S., Sánchez-Cisneros, D., Campillos, L. and Segura-Bedmar, I. Recognizing chemical compounds and drugs: a rule-based approach using semantic information. Proceedings of the fourth BioCreative challenge evaluation workshop, vol 2, (2013)

[5] http://www.biocreative.org/tasks/biocreative-iv/chemdner/