IICE: web tool for automatic identification of chemical entities and interactions

Andre Lamurias[1,2⋆], Luka A. Clarke[1], and Francisco M. Couto[2]

[1] BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisboa, Portugal
alamurias@lasige.di.fc.ul.pt,laclarke@fc.ul.pt,fcouto@di.fc.ul.pt

Abstract. Automatic methods are being developed and applied to transform textual biomedical information into machine-readable formats. Machine learning techniques have been a prominent approach to this problem. However, there is still a lack of systems that are easily accessible to users. For this reason, we developed a web tool to facilitate the access to our text mining framework, IICE (Identifying Interactions between Chemical Entities). This tool annotates the input text with chemical entities and identifies the interactions described between these entities. Various options are available, which can be manipulated to control the algorithms employed by the framework and to the output formats.

Keywords: Text Mining, Machine Learning, Ontologies, Named Entity Recognition, Relation Extraction

1 Introduction

The amount of information about chemical compounds that is published in the form of scientific literature is growing at an unprecedented rate [1]. To update the chemical interactions described in databases, such as DrugBank [4] and IntAct [3], relies on manual reading and parsing the literature. This means that this update will always lag behind scientific publications, as experts extract the relevant information from the papers. For this reason, there is a growing need for automatic methods that transform biomedical text into machine-readable structured data, such as an interaction between compounds.

Information extraction systems applied to the biomedical domain have been developed and are available to the community [5]. However, their performance depends on the machine used by the user, usually requiring external libraries and specific installation instructions. A more practical solution is releasing the system as a web tool, with a front-end enabling any user to test and experiment with it.

We developed the IICE framework (Identifying Interactions between Chemical Entities), for automatic annotation of biomedical documents. IICE is based

⋆ Corresponding author

on supervised machine learning algorithms and semantic similarity between ontology concepts. We have evaluated the framework with the CHEMDNER [7] dataset, for the recognition of chemical entities, and with the DDIExtraction dataset [8], for extraction of drug-drug interactions. The F-measure obtained for each dataset was of 78.26% and 72.52%, respectively, which can be considered nearly state-of-the-art.

The IICE framework can be accessed by a web tool[3], with several configuration options available to the user. These options enable the user to obtain different results by adjusting the methods and thresholds applied. As such, it is possible to set the options for higher recall or precision, depending on the specific needs of the user. The results may be given in the form of HTML tables, or a XML file.

2 Architecture overview

The IICE framework is based on three components which take as input biomedical text, to accomplish distinct tasks. With this modular approach, it is possible to run the framework only using some of the modules, which may be useful if the text was already partially annotated, or if it is going to serve as input to another framework.

Entity recognition This module recognizes the chemical entities mentioned in each sentence. If the input consists of more than one sentence, we split the text in sentences, and process each one individually. The input text is classified using Conditional Random Fields classifiers [2], trained with data sets from community challenges [7, 8]. We have trained classifiers for specific types of chemical entities, in order to obtain higher recall and also provide a type for each entity recognized.

Validation The chemical entities recognized in the text are normalized to ChEBI ontology [6] identifiers. Using the ChEBI ontology, it is possible to validate the entities recognized in the same sentence. Our assumption is that entities that were correctly recognized in a given sentence should share more similarity than recognition errors. Therefore, we implemented a filter to exclude entities with low semantic similarity to other recognized entities in the same fragment of text. This approach obtained high precision values.

Relation extraction The relation extraction module identifies pairs of entities in the text that are described as interactions. We trained a classifier with the DDIExtraction dataset [8], using kernel-based learning algorithms [9, 10]. This type of algorithm has been successfully applied to other relation extraction tasks. The input text for this module should be already annotated with chemical entities, either by the previous module, with a different framework, or manually. Each interaction is also labeled with one of the types of interactions considered in the DDIExtraction dataset.

[3] `http://www.lasige.di.fc.ul.pt/webtools/iice/`

3 Web tool

The IICE web tool can be used to automatically annotate the abstract of a scientific article with chemical entities and interactions. This can be useful for applications such as developing a network of interactions based on the literature, or finding articles relevant to a particular chemical compound.

Figure 1 shows the options that are available to the user. Using these options, it is possible to recognize only the chemical entities in the text (NER), or only the chemical interactions (RE) if the text is already annotated with chemical entities, or both. The input text can be annotated with the "<entity>" tag. We have trained classifiers for entity recognition with two datasets, annotated with different criteria: the CHEMDNER corpus considers various types of chemical entities, while the DrugNER corpus is focused only on drugs. The user may choose to use only the set of classifiers trained with one of the datasets, if annotations similar to that dataset are preferred. We also provide several options related to the validation module, in order to tune the framework for higher recall or precision. Finally, it is also possible to choose which types of machine learning algorithms to use for Relation Extraction. The user may select the classifier we have trained with the Shallow Language kernel [9] or with the Subset Tree kernel [10]. The ensemble classifier combines the results of the kernel classifiers with other domain-specific features to obtain better results. We have previously described in detail how these algorithms were applied and how they may influence the results [11].

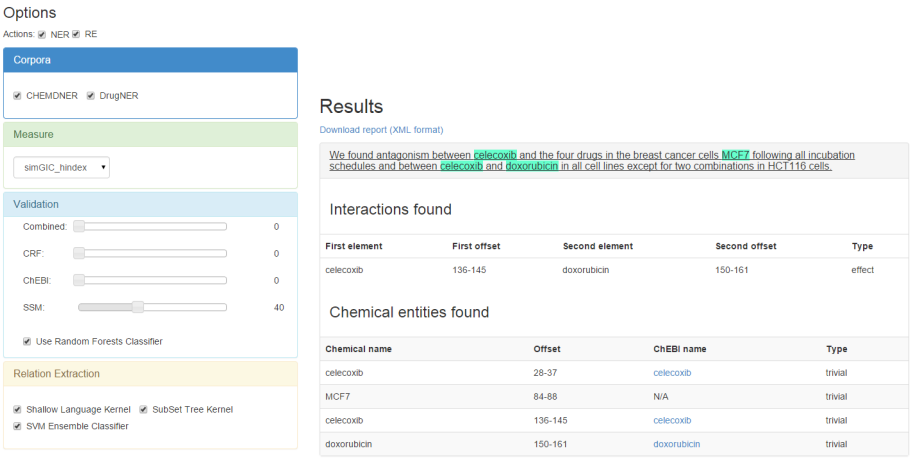

Fig. 1. Screenshot of the options panel and results obtained with the IICE web tool.

The results obtained with the web tool are shown on Figure 1. First we provide a link to the results in the XML format used by the DDI Extraction

dataset. Then, the original text is shown, with the chemical entities highlighted. We organize the interaction and the chemical entities found in two distinct tables. The interactions table provides the two elements of the interactions, and the type of chemical interaction. The entities table provides the name, offset and type of chemical entity, as well as the ChEBI ontology identifier mapped to that entity.

Using only one set of NER classifiers, the system takes about 10 seconds to process one sentence. This value increases as more options are activated, taking as long as 60 seconds if all classifiers are used. We plan on improving the speed performance of the web tool by pre-loading the classifiers and ontologies and deploying the tool on the server as a background service.

Acknowledgments. This work was supported by the Fundação para a Ciência e a Tecnologia (https://www.fct.mctes.pt/) through the PhD grant PD/BD/106083/2015 and LaSIGE Unit Strategic Project, ref. PEst-OE/EEI/UI0408/2014 and by the European Commission (http://ec.europa.eu) through the BiobankCloud project under the Seventh Framework Programme (grant #317871).

References

1. Hunter, L., & Cohen, K. B.: Biomedical language processing: what's beyond PubMed?. Molecular cell, 21(5), 589-594 (2006)
2. McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit (2002)
3. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., ... & Hermjakob, H.: The IntAct molecular interaction database in 2012. Nucleic acids research, gkr1088 (2011)
4. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., ... & Wishart, D. S.: DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research, 42(D1), D1091-D1097 (2014)
5. Leaman, R., & Gonzalez, G.: BANNER: an executable survey of advances in biomedical named entity recognition. In Pacific Symposium on Biocomputing (Vol. 13, pp. 652-663) (2008)
6. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., ... & Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic acids research, 41(D1), D456-D463 (2013)
7. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., ... & Segura-Bedmar, I.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform, 7(Suppl 1), S2 (2015)
8. Herrero-Zazo, M., Segura-Bedmar, I., Martnez, P., & Declerck, T.: The DDI corpus: An annotated corpus with pharmacological substances and drugdrug interactions. Journal of biomedical informatics, 46(5), 914-920 (2013)
9. Giuliano, C., Lavelli, A., & Romano, L.: Exploiting shallow linguistic information for relation extraction from biomedical literature. In EACL (Vol. 18, pp. 401-408) (2006)
10. Moschitti, A.: Making Tree Kernels Practical for Natural Language Learning. In EACL (Vol. 113, No. 120, p. 24) (2006)
11. Lamurias, A., Ferreira, J. D., & Couto, F. M.: Identifying interactions between chemical entities in biomedical text. Journal of integrative bioinformatics, 11(3), 247 (2014)