Identifying Chemical Entities based on ChEBI

Tiago Grego [1]*, Francisco R. Pinto [2] and Francisco M. Couto [1]

[1]Lasige, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal
[2]Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade de Lisboa, Portugal

ABSTRACT

This software demonstration paper presents Identifying Chemical Entities (ICE), a platform composed by algorithms for chemical entity recognition, entity resolution to a reference database, namely ChEBI, and validation using chemical semantic similarity. It aims to provide the users with an improved display of entity recognition results, exposing outliers which are possible recognition errors and displaying evidence that corroborates consistent chemical entities in the entity recognition and resolution process.

1 INTRODUCTION

Chemical information in the scientific literature is increasing at a fast pace, making it difficult for researchers to keep up to date with what is being published. Chemical abstracting services addressed this issue by having chemistry experts manually extract the necessary information from the literature. However, with the exponential growth in publication rate, automatic methods for chemical entity identification are increasingly needed.

The lack of available chemical terminologies has been an important aspect for the slow development of chemical text mining systems, but the recent release of ChEBI (de Matos *et al.*, 2010) allowed their development and some chemical entity recognition tools are already available. ChEBI is however more than a dictionary of molecular entities, it is an ontology that provides a structured classification of the molecular entities. The ontology structure is essentially a directed acyclic graph (DAG) and comprises three separate sub-ontologies (Chemical Entity, Role, and Subatomic Particle). The Chemical Entity sub-ontology provides a structural relationship between terms while the Role ontology provides a functional relationship between them, allowing for a thorough comparison of chemical entities.

ICE (Identifying Chemical Entities) is a software platform that integrates algorithms for chemical entity recognition in biomedical literature, resolution of named entities to the ChEBI database, and validation of annotations using semantic similarity in the ChEBI ontology to gather annotation evidence. Figure 1 shows an outline of the ICE architecture.

2 RECOGNITION

The software platform uses the algorithms for chemical entity recognition presented by Grego *et al.* (2009). This entity recognition system follows a machine learning approach using an implementation of Conditional Random Fields (CRF) to build a classification model based on a manually annotated patent document corpus.

The first step in the entity recognition process is the splitting of the input text into a sequence of tokens, which are then classified according to the previous model. With this method chemical named entities are located in the input text, however there is no mapping to a reference database and thus an entity resolution module is required.

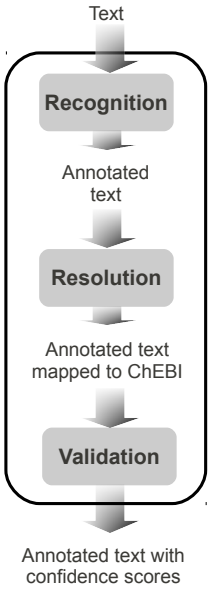

Fig. 1. Architecture of ICE with the three modules for Recognition, Resolution and Validation of chemical entities.

3 RESOLUTION

For entity resolution to the ChEBI database this software platform uses the algorithms presented in Grego *et al.* (2012). This module takes as input the string identified as being a chemical compound name and returns the most relevant ChEBI identifier along with a confidence score.

A lexical similarity method is used to compare the constituent words in the input string with the constituent words of each ChEBI term, to which different weights have been assigned according to its frequency in the database. A final score is provided with the mapping and a minimum score threshold can be used to allow for no mapping to be made in cases where the provided mapping score is too low, which might be an indication that the term is absent from ChEBI.

4 VALIDATION

A novel algorithm was developed for this software platform to perform validation of named entities mapped to ChEBI. The

*To whom correspondence should be addressed: tgrego@fc.ul.pt

underlying assumption is that most often a text fragment such as a paragraph has a limited scope, and therefore normally contains entities that are somehow related to each other. With this in mind, this algorithm takes as input entities mapped to ChEBI within a text fragment and searches for relationships between them. The output is for each input entity the most similar entity within the text fragment, with the corresponding similarity score.

Using the ontology structure of ChEBI we are able to compare chemical entities according to both structural and functional characteristics through several possible semantic similarity measures (Ferreira and Couto, 2010). The BOA framework (Tavares *et al.*, 2011) offers an implementation for chemical semantic similarity calculation.

Based on the maximum similarity score of each entity we can filter outliers and corroborate consistent entities. This is performed using two thresholds. The algorithm has thus four parameters that can be tuned according to the user requirements: the text fragment window, which can be the full document, paragraph or sentence; the semantic similarity measure to be used for the comparison of the entities; and the two thresholds, which can be tuned to allow for more precision or recall.

The final result is advantageous in semi-automated tasks by providing an improved view over the entity recognition results, because the user will have an indication of which entities have increased consistency and are most probably correctly identified, as well as which entities are outliers in the sense that no similar entities could be find, which might be an evidence of a recognition error.

5 EXAMPLE

As an example lets consider the following sample sentence and follow the steps of ICE.

A mixture of ethanol, propanol and acetic acid with a small amount of sodium chloride.

In the entity recognition step four entities can be found are now highlighted.

A mixture of ethanol, propanol and acetic acid with a small amount of sodium chloride.

In entity resolution, ChEBI identifiers are assigned to the entities.

A mixture of ethanol [CHEBI:16236], propanol [CHEBI:28831] and acetic acid [CHEBI:15366] with a small amount of sodium chloride [CHEBI:26710].

Entities mapped to ChEBI are compared to each other, and those with high maximum similarity are considered consistent. That is the case of ethanol and propanol, which have high semantic similarity. Sodium chloride has a low similarity with the other entities, and is thus considered an outlier. Acetic acid in the example has reasonable similarity to both ethanol and propanol, but not high enough to be considered consistent. The final result would highlight differently the new three classes of entities.

A mixture of ethanol [CHEBI:16236], propanol [CHEBI:28831] and acetic acid [CHEBI:15366] with a small amount of sodium chloride [CHEBI:26710].

In this example, the text fragment window is very small and thus there are few entities to be compared, which can provide misleading results. All four entities were in fact correct, but the indication of consistent and outlier entities still provide interesting and meaningful explanation that users can use in a semi-automated fashion.

6 CONCLUSION

This software demonstration paper presents ICE, a framework that performs chemical entity recognition and resolution to the ChEBI database. The entity recognition results are then further processed using the ChEBI ontology to identify outliers and consistent entities, providing the user with annotation evidence.

The goal is to provide the user, potentially a curator performing semi-automatic annotation tasks, different layers of certainty in the recognized entities for better analysis of automatic chemical entity recognition results, as well as providing evidence based on semantic similarity.

The framework is being extended to other ontologies in addition to ChEBI using ontology matching techniques that can align shared concepts between ontologies such as the Gene Ontology and ChEBI (Cruz *et al.*, 2011).

Also, this software platform will be used in the context of the project SPNet. This project aims to uncover network motifs associated with virulence in *Streptococcus pneumoniae*. This requires an extensive analysis of the transcriptional and metabolic networks involved in virulence, and the integration of chemical data with genomic and proteomic data will be required.

ACKNOWLEDGEMENTS

The authors want to thank the Portuguese Fundação para a Ciência e Tecnologia through the financial support of the SPNet project (PTDC/EIA-EIA/119119/2010), the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD grant SFRH/BD/36015/2007 and through the LASIGE multi-annual support.

REFERENCES

Cruz, I., Stroe, C., Pesquita, C., Couto, F., and Cross, V. (2011). Biomedical ontology matching using the AgreementMaker system. In *Software Demonstration at the International Conference on Biomedical Ontologies (ICBO)*.

de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. *Nucleic Acids Research*, **38**, D249–D254.

Ferreira, J. and Couto, F. (2010). Semantic similarity for automatic classification of chemical compounds. *PLoS computational biology*, **6**(9), e1000937.

Grego, T., Pęzik, P., Couto, F. M., and Rebholz-Schuhmann, D. (2009). Identification of chemical entities in patent documents. In *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, volume 5518 of *Lecture Notes in Computer Science*, pages 942–949.

Grego, T., Pesquita, C., Bastos, H., and Couto, F. (2012). Chemical entity recognition and resolution to ChEBI. *ISRN Bioinformatics*, **2012**.

Tavares, B., Bastos, H., Faria, D., Ferreira, J., Grego, T., Pesquita, C., and Couto, F. (2011). The Biomedical Ontology Applications (BOA) framework. In *Software Demonstration at the International Conference on Biomedical Ontologies (ICBO)*.