Recognition and resolution of chemical entities to ChEBI.

Tiago Grego[1], Francisco M. Couto[2]

[1,2] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa

[1] tgrego@fc.ul.pt (corresponding author)

[2] fcouto@di.fc.ul.pt

Abstract

Virtually all biological phenomena have their roots in chemical processes, and it is impossible to understand the complex signaling and metabolic networks of living systems without taking into account the chemical entities that take part in it.

Since the late 1960's, Bioinformatics emerged and was applied in the creation and maintenance of public biomedical databases, but only recently public chemical databases such as ChEBI (Chemical Entities of Biological Interest) were released providing high quality manually curated chemical data organized in an ontology.

However, the largest repository of scientific data is the scientific literature, containing millions of documents such as scientific publications and patents in unstructured natural language text.

Within the literature exists extensive information that is not covered in other knowledge resources, thus database curators manually analyze and annotate the literature to grow and validate the data in the databases, but this is a tedious, time consuming and costly process.

Fortunately, this process have been addressed by automatic text mining systems that already shown to be helpful in speeding up some steps of this process, namely performing named entity recognition and entity resolution.

Two main approaches are used by such systems:

Dictionary based approaches require domain terminologies to find and map matching entities in the text and depend on the availability and completeness of these terminologies.

Whatizit is a popular text processing system that uses a dictionary based approach for identifying a wide variety of biomedical terms by using several pipelines based on specific terminologies, including ChEBI.

Machine learning based approaches require an annotated corpus which is used to learn a model that can be applied in the named entity recognition of new text. An entity resolution module is required to perform the mapping of recognized entities.

Recently, a joint team of curators from ChEBI and the European Patent Office released a gold standard corpus composed by 40 patent documents manually annotated with 18061 chemical entities, from which 9696 could be mapped to ChEBI. With the availability of this this corpus we developed a Conditional Random Fields based machine learning method for chemical entity recognition and a lexical similarity based method for chemical entity resolution to the ChEBI database, and compared our methods with Whatizit [1].

We found that dictionary based method can already provide competitive results in recognizing chemical named entities obtaining F-measures of up to 70% for partial matching and 32% for exact matching assessment in the full gold standard. However the developed machine learning method consistently outperformed it obtaining F-measures of 77% for partial matching and 57% for exact matching.

A known drawback of dictionary based methods is the inability to recognize entities not present in the dictionary used, so an evaluation was performed using only the mapped entities in the gold standard. Still, the machine learning method outperformed the dictionary based method obtaining 66% and 57% F-measure for respectively partial and exact matching against the 54% and 39% obtained using Whatizit.

Taking into consideration not only entity recognition but also resolution, our methods were able to obtain an F-measure of 51% for partial matching and 47% for exact matching, while the dictionary method could only obtain 37% and 32% respectively.

Overall, we demonstrated that a completely dictionary independent machine learning entity recognition method and a lexical similarity resolution method can surpass dictionary based methods in recognizing chemical compounds and mapping them to the ChEBI database.

Tools like the one presented here [2] can be helpful to gather more knowledge about biomedical entities involved in complex biological processes, such as metabolic networks. Thus, we intend to apply our tool to the project SPNet, which aims to uncover network motifs associated with virulence in pneumococcus. After the identification of genes with an impact in virulence, the analysis of the transcriptional and metabolic network for neighbor genes can locate other direct or indirect target genes involved in the virulence.

The metabolic networks comprise the chemical reactions within the organism, thus to understand them a characterization of their chemical entities is required. This will be the aim for the application of our tool to SPNet, in order to identify the chemical entities in literature, which may provide more information for the integration with the genomic and proteomic data that may uncover the network motifs associated with the virulence of the bacteria.

References

1. Rebholz-Schuhmann et al, Text processing through Web services: calling Whatizit. Bioinformatics, 24(2) 296-298, 2008
2. Grego T et al, Identification of chemical entities in patent documents. Lecture Notes in Computer Science, vol.5518, 942–949, 2009