UEM-UC3M: An Ontology-based named entity recognition system for biomedical texts.

Daniel Sanchez-Cisneros
Universidad Carlos III de Madrid
Avda. de la Universidad, 30
28911 Leganés - Madrid - Spain
dscisner@inf.uc3m.es

Fernando Aparicio Gali
Universidad Europea de Madrid
C/ Tajo, s/n. Urb. El Bosque
28670-Villaviciosa de Odón- (Madrid)
fernando.aparicio@uem.es

Abstract

Drug name entity recognition focuses on identifying concepts appearing in the text that correspond to a chemical substance used in pharmacology for treatment, cure, prevention or diagnosis of diseases. This paper describes a system based on ontologies for identifying the chemical substances in biomedical text. The system achieves an F-1 measure of 0.529 in the task.

1 Introduction

Named entity recognition (NER) involves processing text and identifying certain occurrences of words belonging to particular categories of named entities. In recent years, much attention has been paid to the problem of recognizing gene and protein mentions in biomedical abstracts for different purposes such as information extraction, relation extraction or information retrieval. In this case we focus on the pharmacological domain. Furthermore, some initiatives have promoted the evaluation of different systems of named entity recognition and relation extraction in the pharmacological domain. This is the case of *Semeval 2013: Recognition and classification of drug names* task[1] (Segura-Bedmar et al., 2013), where the system presented in this communication has been evaluated.

Following the annotation guidelines of the task, a drug is a substance that is used in the treatment, cure, prevention or diagnosis of disease. Moreover, each drug name entity can be classified in four different types: *drug*, *brand*, *drug_n* and *group*. Our system uses biomedical ontologies and external resources (containing biomedical information) as input to determine whether we are treating a drug name entity or not.

The resource integration seems to represent an improvement since the knowledge available for identifying entities is higher. Some biomedical resources such as *Drugbank*[2], *Kegg*[3], *Pubchem*[4] or *Drugs.com*[5] focus on providing a compound of information collected from different sources.

Section 2 exposes some related work in the field of NER. In section 3 we describe the system used for identifying drug name entities. Section 4 presents the results obtained by the system and a little comparison with other approaches. In section 5 we outline some conclusions obtained and ideas for future work.

2 Related work

The field of NER has been very studied in recent years, and has been faced in many approaches. Since text structures are frequently used to characterize documents in text mining algorithms, there only stand out those based in terms and

[1] http://www.cs.york.ac.uk/semeval-2013/task9/data/uploads/task-9.1-drug-ner.pdf

[2] http://www.drugbank.ca/
[3] http://www.genome.jp/kegg/
[4] http://pubchem.ncbi.nlm.nih.gov/
[5] http://www.drugs.com/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 622–627, Atlanta, Georgia, June 14-15, 2013.

concepts. This is due to that concept-based systems represent the semantic content with a smaller number of characteristics, opposite to the term-based systems based on characters or words. Concept-based and term-based representations mainly differ in the implicit or explicit appearance, respectively, of the words identified in the document. This fact implies that concept-based extraction techniques are more complex, requiring the use of more advanced computational linguistics techniques and a greater dependence on knowledge domain.

One reference system that focuses on concept recognition in the biomedical domain is *MetaMap* (Aronson, 2001). MetaMap is a program developed by the National Library of Medicine (NLM) that uses the UMLS Metathesaurus for annotating the concepts in a given text. The program is designed to obtain the concept that best fits a particular phrase, finding its origin in an attempt to improve the retrieval of biomedical literature indexed in MEDLINE/PubMed. MetaMap is a program with many strengths, such as the power of linguistic analysis, the high performance setting possibilities and the variety of processing algorithms included. On the other hand, MetaMap shows some weaknesses such as the algorithms developing focused on English grammar texts, or high processing time lapse due to the complexity of the algorithms (not suitable for real-time systems). Metamap analysis time periods goes from less than a minute for short simple text to long hours for complex sentences.

Gimli (Campos et al., 2013) is an open source and high-performance solution for biomedical named entity recognition on scientific documents, supporting the automatic recognition of gene, proteins, DNA, RNA, and cell domain names. This tool implements a machine learning approach based on conditional random fields (CRF).

On the other hand, there exists a more recent concept extraction techniques based on ontologies. Ontologies link concept labels to their interpretations, ie specifications of their meanings including concept definitions and relations to other concepts. Apart from relations such as *isa* and *part-of*, generally present in almost any domain, ontologies also model domain-specific relations, eg *clinically-associated-with* and *has-manifestation* are specific associations for the biomedical domain. Therefore, ontologies reflect the structure of the domain and constrain the potential interpretations of terms. Thus, ontologies can provide rich concept knowledge of domain specific name entities. This is the case of *Open Biomedical Annotator* (*OBA*) (Jonquet et al., 2009), an impressive annotation system using ontologies, which provides online access for users and for other systems as a Web service. There are other examples of utilities for extracting concepts using ontologies (e.g. Terminizer (Hancock et al., 2009), Whatizit (Rebholz-Schuhmann et al., 2008) or Reflect (Pafilis et al., 2009)). However, the magnitude of ontologies and resources integrated under the OBA Web service is difficult to reach by other systems (Whetzel et al., 2011): in three years

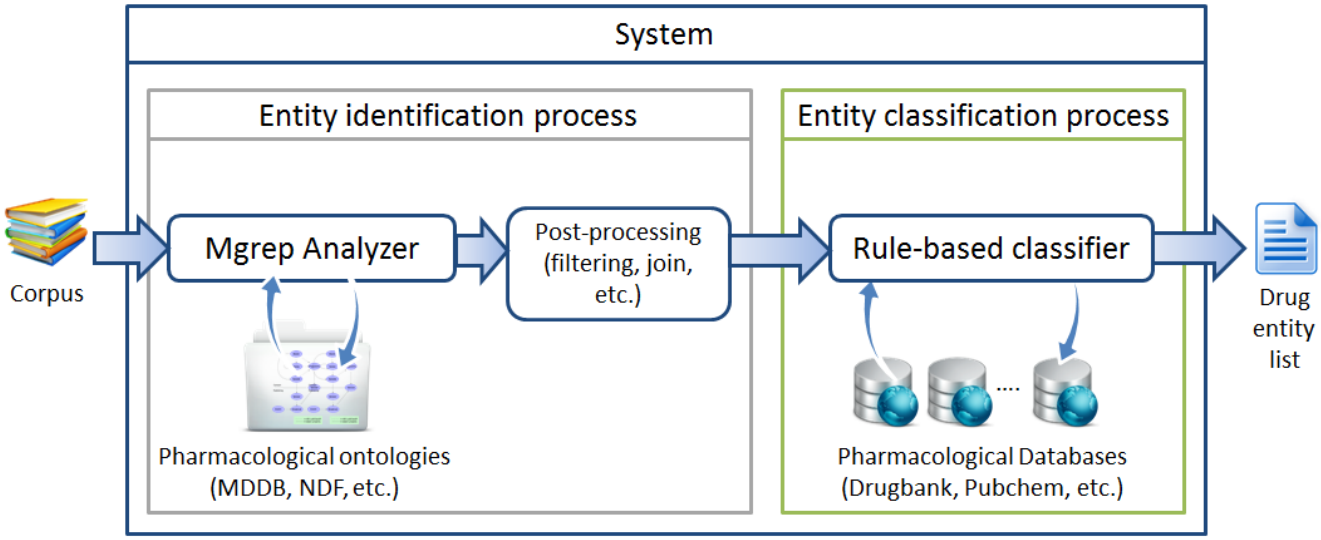

Figure 1: Architecture of the system.

(from 2008 to 2011), they have increased from 72 to 260 biomedical ontologies.

The concept recognition tool used by the OBA system -in order to find ontology concepts matching the terms extracted from texts- is called *Mgrep*. Although Mgrep is not a free tool, some results are presented in (Jonquet et al., 2008). A comparison between Mgrep and MetaMap can also be found in (Shah et al., 2009), where they make an evaluation over a biological and disease terms dictionaries with precision (0.87 to 0.71 respectively) and recall (1548 to 1730 recovered terms respectively) metrics. Thus, we decided to use Mgrep for identifying drug name entities in the system.

3 Description of the system

The system (see figure 1) is divided in two phases: (i) in one hand, the system must scan drug name entities without specifying any further information. This is the so-called entity identification process; (ii) on the other hand, the system classifies by using a rule-based process the type of the entities discovered previously. This is the so-called entity classification process.

The corpus is processed sentence by sentence, using the identification tag provided for each sentence.

3.1 Entity identification process

In this phase we analyze each sentence of the corpus with Mgrep analyzer. This tool allows us to set the ontologies we want to use in the analysis. All additional ontologies used in the analysis increases the computational complexity required.

The ontologies used in this first drug name identification phase belong to UMLS collection, and more specifically to the pharmacological domain:

- Master Drug Data Base[6] (MDDB): National Drug Data File ontology provides a codified drug dictionary, drug vocabulary, and drug pricing for prescription drugs and medication-based over-the-counter products in the United States. It supports the ever-changing world of drug information in healthcare.
- National Drug File[7] (NDF): this ontology contains information about a comprehensive set of drug database elements and clinical information approved by the U.S. Food and Drug Administration (FDA), and dietary supplements information.
- National Drug Data File (NDDF): this is an extension of the NDF ontology that includes chemical ingredients, clinical kinetics, diseases, dose forms, pharmaceutical preparations, physiological effects and

Annotations

TERM filter	ONTOLOGY filter	TYPE filter	CONTEXT
Pharmaceutical Preparations	National Drug File	ancestor	medicine containing kaolin or attapulgite - Ketoconazole - Central
Pharmaceutical Preparations	National Drug File	ancestor	affect the effect of Pirenzepine or whose effects may
Pharmaceutical Preparations	National Drug File	ancestor	Drug Interactions: Pirenzepine may interact with the

Figure 2a: Result of analysis with the Mgrep analyzer.

TERM filter	ONTOLOGY filter	TYPE filter	CONTEXT
Drug Products by Generic Ingredient Combinations	National Drug File	ancestor	Aventyl, Surmontil) - Potassium chloride (e.g., Kay Ciel)
Drug Products by Generic Ingredient Combinations	National Drug File	ancestor	Tofranil, Aventyl, Surmontil) - Potassium chloride (e.g., Kay Ciel)

Figure 2b: Example of multiword drug entity divided.

[6] http://www.medispan.com/medi-span-electronic-drug-file.aspx
[7] http://www.fdbhealth.com/fdb-medknowledge/

therapeutic categories.

- Ontology for Drug Discovery Investigations: this ontology contains information about description of drug discovery investigations from OBO[8] relation ontology.
- MESH Thesaurus[9]: this ontology contains sets of terms naming descriptors in a hierarchical structure. There exist 26,853 descriptors and over 213,000 entry terms in 2013 MeSH.

For each drug name entity identified the Mgrep analyzer provides information about the ontology concept recognized, term information, snippet of original text (see figure 2a). After identifying drug name entities we noticed some errors in the recognized concepts, thus we held a post-processing of the analysis results. Some entities are recognized by several ontologies at the same time, so it is necessary to filter repeated instances.

Biomedical complex name entities are not identified. To solve this, we join compound name entities by following the charoffset of the sentence. The system only links two or more drug entities that were next to each other, without punctuation between them. For example, *potassium chloride* (see figure 2b) is recognized separately in potassium and chloride, so we group it as *potassium chloride* concept.

As a result of this process we obtain a list of clear drug name entities that conforms our run 1 approach in the task. However, we elaborate a second filter based in a gazetteer containing terms with no useful meaning for our drug name entity identification purpose. This gazetteer contains terms such as *agent*, *compound* and *blocker*. The results of this second filter conforms our run 2 approach in the task. As a result of entity identification phase we obtain a list of drug name entities, but they are not identified as any type yet.

3.2 Entity classification process

In this phase we classify the list of pharmaceutical terms obtained from analysis phase. To do so, we elaborate a rule-based system following the annotation methods described in the task guidelines. This annotation method was based in biomedical resources, such as DrugBank, for determining aspects as if the drug entity is approved for human use, or if the drug entity is registered as a brand name. We can organize the general rules of the classification process by resources used:

- DrugBank: These rules search the drug entity in DrugBank resource and obtain several information:
  - Drug information: information about approval state of the drug (a*pproved*, *experimental*, *illicit*). A rule classifies a drug entity as *drug_n* when *experimental* or *illicit* state is found in a drug, otherwise the drug entity is catalogued as *drug* type.
  - Synonym list: list of possible registered names of the entity. A recursive process searches each synonym in DrugBank (obviating the synonym list this time), and applies the rules as if original drug entity were treated. The result of the recursive process affect to the original drug entity.
  - Brand name list: list of registered commercial brand names of the entity. If a drug name entity is found in the brand name list, then it is catalogued as a *brand* type.
  - Categories: information about general category of drug. If the drug is found as a category, then it is classified as *group* type.
- Pubchem: These rules search the drug entity and obtain information of drug identification and compound information and IUPAC name.
- ATC Index[10]: These rules look for the drug entity in ATC Index resource and determine whether the entity is *drug* or *group* depending on the level of ATC code found.
- Kegg: These rules search the drug entity in this resource and obtain information of drug categories. If the drug is found as a category, then it is classified as *group* type.
- MeSH[11]: These rules search information about MeSH tree categories classification of the drug entity. If the drug is found as a category, then it is classified as *group* type. Another rule makes a naïve processing of the MeSH

[8] http://www.obofoundry.org/ro/
[9] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[10] http://www.whocc.no/atc_ddd_index/
[11] http://www.ncbi.nlm.nih.gov/mesh/67055162

description text to evaluate if the drug entity were used in humans. If this information is found in the text, then the drug entity is classified as a *drug* type.

The described rules are representative examples of the complete rule-based system. There were assigned priorities to the rules, since some rules are more certain to describe a drug type than others. Thus, if a drug entity is found to be approved for using in humans after processing the MeSH text, but when looking the DrugBank state is found as illicit state, then the drug is classified as *drug_n* type since DrugBank offers a certain state of the drug, instead of a natural text description that may be classified as a false positive. Depending on the values collected on these biomedical resources the rule-based system determines whether the type of an entity is a *drug*, *group*, *brand* or *drug_n*.

4 Results

The best result in entity identification (exact matching) obtained by the system correspond to run 2, achieving a F1 measure of 0.609. On the other hand, the best results achieved in strict matching (boundary and type evaluation) correspond to run 2 again, with 0.529 F1 score.

	Partial matching			*Exact matching*			*Strict matching*		
Team	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Run 1	0,502	0,7	0,585	0,454	0,633	0,528	0,393	0,548	0,458
Run 2	0,653	0,685	**0,669**	0,594	0,624	**0,609**	0,517	0,542	**0,529**

Table 1: Results obtained by the system.

These results contrast with the result obtained by run 1, achieving a F1 measure of 0.528 and 0.458 in entity identification and strict matching evaluation respectively. Thus we can quantify the advantage of using a filter based on gazetteer in an average increment of 0.079 F1 measure.

We have noticed that the higher results are obtained in partial matching evaluation because of the relaxed conditions of the charoffset. This seems reasonable since complex multiword entity is hard to parse and define an exact charoffset.

On the other hand, we also noticed that evaluating the classification of the type decrement the best results obtained by the system from 0.609 to 0.529 of F1 score. This indicates that there is still a lot of improvement work in the rule-based system for type classification. A little error analysis was done in a set of 10 documents of the training dataset. The results show errors in conflictive entities that show multiples categories in DrugBank resource. Thus, for example *cocaine* drug entity contains tags of *illicit* and a*pproved* in DrugBank database, so the system classify this entity as *drug_n* instead of *drug*.

5 Conclusions and future work

In this paper we present a system for drug name entity recognition based on ontologies as participation for "Semeval 2013: Recognition and classification of drug names" task. The system is based on integration of biomedical resources for identification and classification of pharmacological entities. The best result of the system obtained an F1 measure of 0.529.

The usage of ontologies in named entity recognition task seems to be a good choice since we can select specific ontologies. A possible future work includes an improvement of rule-based system, including a bigger collection of biomedical resources. The entity classification could increase the results by creating an hybrid approach between rule-based methods and machine learning techniques. On the other hand, in the entities identification task, the system could include other biomedical text analyzers and establish a vote system. This would improve whether we consider an entity or not. Finally, in error analysis were noticed problems related to rule-based module. Therefore, an insightful improve could pass through making a context analysis in order to clear the ambiguity surrounding the drug entity.

Acknowledgments

This work has been funded by MA2VICMR project (S2009/TIC-1542) and MULTIMEDICA project[12] (TIN 2010-20644-C03-01).

References

[12] http://labda.inf.uc3m.es/multimedica/

Aronson, A.R. 2001. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*.Proc AMIA Symp. 17–21.4.

Campos, D., Matos, S., Oliveira J.L.. 2013. *Gimli: open source and high-performance biomedical name recognition*. BMC Bioinformatics 14:54.

Hancock, D., Morrison N., Velarde G., Field D. 2009. *Terminizer - Assisting Mark-Up of Text Using Ontological Terms*. Nature Precedings.

Jonquet C., Musen M.A., Shah N. 2008. *A System for Ontology-Based Annotation of Biomedical Data*. Data Integration in the Life Sciences, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 144–152.

Jonquet, C., Shah N.H., Musen M.A. 2009. *The Open Biomedical Annotator*, Summit on Translat Bioinforma. 56–60.

Pafilis E., O'Donoghue S.I., Jensen L.J., Horn H., Kuhn M., Brown N.P., et al. 2009. *Reflect: augmented browsing for the life scientist*. Nature Biotechnology, 27, 508–510.

Rebholz-Schuhmann D., Arregui M., Gaudan S., Kirsch H., Jimeno A. 2008. *Text processing through Web services: calling Whatizit*. Bioinformatics. 24, 296–298.

Segura-Bedmar I., Martínez P., Herrero-Zazo M. 2013. *SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)*. Proceedings of Semeval 2013.

Shah N.H., Bhatia N., Jonquet C., Rubin D., Chiang A.P., Musen M.A. 2009. *Comparison of concept recognizers for building the Open Biomedical Annotator*. BMC Bioinformatics.10, S14.

Whetzel P.L., Noy N.F., Shah N.H., Alexander P.R., Nyulas C., Tudorache T., et al. 2011. *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications*. Nucleic Acids Research. 39, W541–W545.