Extraction of Drug-disease relations from MEDLINE Abstracts

Aida Bchir
High institute of management, Tunisian University
Cité Bouchoucha, Tunis, Tunisia
aidabchir@gmail.com

Wahiba Ben Abdessalem Karaa
High institute of management, Tunisian University
Cité Bouchoucha, Tunis, Tunisia
wahiba.abdessalem@isg.rnu.tn

Abstract—Biological research literature, as in many other domains of human activity, is a rich source of knowledge. MEDLINE is a huge database of biomedical information and life sciences; it provides information in the form of abstracts and documents. However, extracting this information leads to various problems, related to the types of information such as recognition of all terms related to the domain of texts, concepts associated with them, as well as identifying the types of relationships. In this context, we suggest in this paper an approach to extract disease-drug relations: in a first step, we employ Natural Language Processing techniques for the abstracts’ preprocessing. In a second step we extract a set of features from the preprocessed abstracts. Finally we extract a disease-drug relation using machine learning classifier.

Information extraction; SVM; disease-drug relation;

I. INTRODUCTION

Medical computing systems have seen an explosive growth over the past two decades and are used to store information, to access to this information to discover new knowledge or provide decision support to improve the quality of care. Within this general framework, the information used is mainly medical records of patients, such as clinical summaries account and the minutes of consultation, which contain information on many medical patients and also the medical literature. Every day, hundreds of biomedical papers are written. Many of these papers are available online. MEDLINE for example is an online bibliographic database on biomedical domain that contains over 22 million references of papers. It is used in many researches[1]. The large part of needed information is in textual form and we are able to automatically extract it from texts. The need to convert all this information in a structured form is a major challenge and is the starting point for the development of appropriate tools for querying and automatic processing of that information. Extracting this information leads to various problems, related to the types of information such as recognition of all terms related to the domain of texts, concepts associated with them [2] [3] [4], as well as identifying the types of relationships. The extraction of relationships is a process whose goal is to determine the existence of a semantic link between two entities and, where possible, to characterize the nature of this relationship. Many researches are interested by this domain like extraction of drug-drug interaction [5], Drug-protein interactions, Protein-protein.

We are particularly interested in this study to the extraction of relations between drug and disease and building of a knowledge base to help users discovering relation between these two concepts. We use, for this purpose, a classification approach to detect these relations.

II. RELATED WORK

Information extraction in biomedical domain encompasses biomedical entities recognition such as extraction of disease, treatment, protein, and extraction of relations between these entities. There are many approaches used to extract relation between two entities from abstracts. One class of these approaches is based on statistical methods, B.J. Stapley and al. [6] developed a system that retrieve and visualize knowledge from database and literature using gene names. In this approach a co-occurrence method is used.

In general, this approach has low precision and high recall and it is difficult to identify the semantic type of the extracted relation. Another type of approach is a rules based approach [7]. In this study the aim is to identify and extract bimolecular relation using the inhibit relation as for example. They obtain 90% of precision and 57% of recall. They first develop the semantic automata from the UMLs database, they create the syntactic patterns, and then they extract entities that are connected by some form of the patterns. This method is mostly driven with human efforts.

The third approach is based on Machine learning methods [8]. It uses Naive Bayes classifier. Its object is the extraction of the relation subcellular-localization (Protein,Subcellular-Location), this relation represents the protein and the subcellular where it found. Michele and al. [9], Guillaume Tisserant and al. [10] use also this method in their work. Angus and al.[11] apply SVM to extract relation between biomedical entities from oncology narratives. BioPPISVMExtractor system is developed by Zhihao Yang and al., [12] to extract protein-protein interactions from biomedical literature, it is svm-based system. They obtain 41.84% of recall and 55.41% of precision.
Ontology-based approach can be also used to extract biomedical relation. ONBIRES is a system that is created by Minlie Huang and al. [13], it is used to extract biological relation from Medline texts. Gene-disease, gene-gene and protein-protein are the type of relation extracted by this system. It is an ontology-based system. They use four external ontology including MeSH, OMIM, GO and LocusLink.

Several methodologies are available for the extraction of semantic relations, especially for the extraction of relation between drug and disease. Rosario and Hearst [14] were interested in disambiguation of seven types of relationships. They compared five generative models and a neural network model and found that the latter allows for better results. Lee et al. [15] applied patterns manually built on medical abstracts to identify drug-disease relationships. They obtain 48.1% as precision and 84.8 as recall.

III. OUR APPROACH WORKFLOW

The whole process of our approach can be detailed as follows:

a) Pre-processing: In this step, we initially employ several Natural Language Processing techniques in text analysis. These techniques include splitting, tokenization, part-of-speech tagging, parsing and semantic interpretation.

b) Feature extraction: The feature extraction consists on finding a set of features from the preprocessed data set. The goal of this step is to find out relevant indicators that may help us capturing the accurate class of the relation between drug and disease situated in different sentences.

c) Extraction drug-disease relation: This step consists on training and testing classification models on our data set in order to learn all relations between drug and disease.

A. Feature extraction

We use the output of the pre-processing step to construct feature vectors for the machine learning algorithm. In the following we will present all features that can be used to extraction drug-disease relation:

a) Frequency features:
   - Number of named entity in the sentence.
   - Number of drugs in the sentence.
   - Number of diseases in the sentence.
   - Number of verbs between each two named entity in the sentence.
   - The number of the words between each two named entities.
   - Bag-of-word count of the word in sentence.

b) Lexical features:
   - Combination of words of the named entity.
   - Combination of words between each two named entities.
   - Combination of 3 words before each named entity.
   - Combination of 3 words after each named entity.
   - Combination of lemmas of the words between each two named entities.
   - Combination of lemmas of the 3 words before each named entity.
   - Combination of lemmas of the 3 words after each named entity.

c) Syntactic features:
   - The POS of each named entity.
   - Combination of POS of words between each two named entities.
   - Combination of POS of 3 words before each named entity.
   - Combination of POS of 3 words after each named entity.
   - Combination of verbs between each tow named entities.
   - First verb before each named entities.
   - First verb after each named entities.

d) Semantic role:
   - Combination of semantic type of the words in the sentence: The possible values are DIS (disease), TREAT (treatment) and NONE.

B. Training and Testing relation Detection

We use a SVM to extract and classify relation between drug and disease.

The classification process is based on the features extracted with the feature extraction module. These features constitute our feature vector used to train our classification model.

IV. CONCLUSION

An enormous amount of information exists in biomedical literatures but to analyze and process this information automatically, it is a very hard task because it is characterized by a rich and specific vocabulary. Biomedical Information extraction is to automatically extract structured information from texts.

In this paper, we propose a disease-drug relation extraction approach. We initially employ Natural Language Processing techniques in text analysis. Then, we extract feature vector from each sentence. Finally, these vectors are used in the training and testing step in order to extract relations between...
treatment and disease. In this step, we use SVM a machine learning classifier.

In the current work, we have only extracted relation between drug and disease; we plan to extract other type of relation between other concepts.

REFERENCES


