ELSEVIER

Drug name recognition in biomedical texts: a machine-learning-based method

Linna He, Zhihao Yang, Hongfei Lin and Yanpeng Li

College of Computer Science and Technology, Dalian University of Technology, Dalian, 116024 Liaoning, China

Currently, there is an urgent need to develop a technology for extracting drug information automatically from biomedical texts, and drug name recognition is an essential prerequisite for extracting drug information. This article presents a machine-learning-based approach to recognize drug names in biomedical texts. In this approach, a drug name dictionary is first constructed with the external resource of DrugBank and PubMed. Then a semi-supervised learning method, feature coupling generalization, is used to filter this dictionary. Finally, the dictionary look-up and the condition random field method are combined to recognize drug names. Experimental results show that our approach achieves an *F*-score of 92.54% on the test set of DDIExtraction2011.

Introduction

The pharmaceutical industry is increasingly becoming a knowledge-based discipline [1]. Scientists need to access relevant information and knowledge in the process of developing drugs. The deluge of published research has overwhelmed most healthcare professionals because it is not possible to remain up-to-date with everything published about, for instance, drug interactions [2]. Industry estimates suggest that 90% of drug targets are derived from the literature [3]. One such source is PubMed with more than 23 million MEDLINE journal article references and abstracts dating as far back as the mid-1960s. With the massive amount of biomedical text, there is an urgent need to develop a technology for extracting the drug information automatically.

Biomedical named entity recognition (NER) aims to find entities in biomedical texts, an invaluable function that becomes very important for further processing such as information retrieval, information extraction and knowledge discovery [4]. At present, it has referred to kinds of domains, such as protein [5–9], gene [8–11], RNA [12] or drug [2,13]. Presently, many approaches, including the dictionary-based methods [14], rule-based methods [15], linguistic-based methods [16] and machine-learning-based methods, have been applied to biomedical NER. Early work in the application of machine learning to NER applied hidden Markov models to the MUC6 task [17]. Since then, work on NER has been dominated by the use of discriminative models, such as maximum entropy [18], conditional random fields (CRF) [19], support vector machine (SVM) [20] and semi-supervised learning methods [21]. All these methods have been used for protein or gene name recognition. However, research on drug name recognition is relatively limited. Here, the term drug refers to four general types of entities as defined in DDIExtraction 2013 task 9.1 [22]: (i) drug, all chemical agents used in the treatment, cure, prevention or diagnosis of diseases that have been approved for human use – this type only represents generic drugs; (ii) brand, drugs that were first developed by a pharmaceutical company; (iii) group, a term in text designating a chemical or pharmacological relationship among a group of drugs; (iv) no-human, a chemical agent that affects living organisms – it is an active substance but it has not been approved to be used in humans for a medical purpose.

The aim of drug name recognition is to identify as many drug names as possible, which is the first step in a method for automatic detection of drug interactions from biomedical texts, a specific type of adverse drug event of special interest in patient safety [23]. Kolarik *et al.* developed an approach for the identification of new terms used in unstructured text that provide information about drug properties. It is based on the identification and extraction of phrases corresponding to lexicosyntactic patterns – so-called Hearst patterns that contain drug names and directly related drug annotation terms [24]. Segura-Bedmar *et al.* presented a system for

Corresponding author:. Yang, Z. (yangzh@mail.dlut.edu.cn), (yangzh@dlut.edu.cn), (yangzhih@dlut.edu.cn)

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.10.006

drug name recognition and classification in biomedical texts [2]. The system combines information obtained by the Unified Medical Language System (UMLS) MetaMap Transfer (MMTx) program, and nomenclature rules recommended by WHO's International Nonproprietary Names (INN) program to identify and classify pharmaceutical substances. In their method, the stems recommended by WHOINN are used, which not only allows the classification of the drug names but also helps to find possible new drug candidates that have not been detected by MMTx. However, there are also some disadvantages about their method as follows: (i) some drug names that do not contain the recommended stems cannot be recognized; and (ii) some words that contain the recommended stems, but not drug names, are recognized.

Recently, DDIExtraction 2013 was launched to address the extraction of drug–drug interactions (DDIs) as a whole, but was divided into two subtasks to enable separate evaluation of the performance for different aspects of the problem [22]. The shared task includes a subtask (Task 9.1): recognition and classification of pharmacological substances. This task concerns the named entity extraction of pharmacological substances in text, which is a crucial first step for information extraction of DDIs. In this task, four types of pharmacological substances mentioned above are defined: drug, brand, group and drug-n (active substances not approved for human use).

A total of six teams participated in this task. Their approaches include a dictionary-based approach, ontology-based approach and machine-learning-based approaches such as CRF, decision tree classifier and SVM classifier. As shown in protein or gene name recognition, with suitable features the CRF model usually outperforms the other sequence-tagging models such as HMM (Hidden Markov Model) and MEMM (Maximum Entropy Markov Model) because it is often considered as extensions to them [25]: the best results were achieved by the WBI team with a CRF model. They employed a domain-independent feature set along with features generated from the output of ChemSpot [26] as well as a collection of domain-specific resources. ChemSpot is an existing chemical named entity recognition tool that uses a hybrid approach combining a CRF with a dictionary for identifying mentions of chemicals in texts, including trivial names, drugs, abbreviations, molecular formulas and International Union of Pure and Applied Chemistry (IUPAC) entities [27].

In addition, the usage of an entity dictionary can usually help improve the performance of NER. For example, ChemSpot combines a CRF with a dictionary built by Hettne *et al.* [28] using name lists from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB and ChemIDplus. The second best-performing team in DDIExtraction 2013 task 9.1 also developed a dictionary-based approach combining biomedical resources such as DrugBank, the anatomical therapeutic chemical (ATC) classification system or MeSH, among others [22]. The dictionaries mentioned above were constructed from the existing manually maintained biomedical resources such as UMLS, DrugBank and MeSH. Therefore, the performance of the dictionary look-up is confined to the prompt update of these biomedical resources because it relies mostly on up-to-date dictionary entries [27]. However, the fact that new chemical names appear with high frequency (especially many of them appear in unstructured texts, for example, PubMed into which thousands of new citations are added daily) makes it difficult for these resources to be promptly updated.

In this review, we present a machine-learning-based approach. In this approach, first a drug name dictionary is constructed using the method of context pattern induction. This method starts with a few seed examples, induces context patterns in an unsupervised way and extends the seed list by extracting more instances from the unlabeled data in PubMed. Second, a semi-supervised method called feature coupling generalization (FCG) is used to filter the dictionary. FCG is a semi-supervised method where the goal is to use unlabeled data to enhance the representation of local lexical features and make better use of sparse features. Third, the drug name dictionary look-up is combined with a CRF model to recognize drug names in biomedical texts. The advantage of our approach is that, with the context pattern induction method, a wide range of dictionary entries can be obtained including the new drug names that only appear in newly published PubMed abstracts. At the same time, the FCG method is used to filter the noise introduced by the dictionary expansion. In this way, a large high-quality drug name dictionary can be constructed. In addition, the combination of the CRF with the dictionary is effective in improving the overall performance of drug entity recognition. Experimental results show that our approach achieves an *F*-score of 92.54% on the test set of the DDIExtraction2011 task.

Methodological approach description

Our approach consists of three main steps: the first step is to build a drug name dictionary; the second step is to filter the drug name dictionary with FCG; the last step is to recognize the drug names in biomedical texts using this dictionary combined with a CRF model with lexical features.

Construction of the drug name dictionary

In our approach, first an initial drug name dictionary is constructed with DrugBank (http://www.drugbank.ca/). It only contains 4774 entries (downloaded June 2012) and cannot be used to recognize all the drug names in biomedical texts. But there are orders of magnitude more present in the unlabeled data in PubMed. Therefore, a context pattern induction method [29] is used to extract drug names from the unlabeled data and construct a much larger drug name dictionary. Then a semi-supervised learning method called FCG is used to filter the expanded dictionary.

First, the drug names in DrugBank are used as seeds to extract the context patterns from the unlabeled data from PubMed. The unlabeled dataset used for extracting context patterns and drug names is the PubMed abstracts between 1979 and 2009. Starting with the seed list, the occurrences of seed drug names in PubMed are found. Then, for each such occurrence, the fixed number *WB* and *WA* of tokens immediately preceding and following the matched drug name are extracted (*WB* and *WA* are experimentally optimized and set to 3 and 2, respectively). All drug names are replaced by the single token -ENT- which represents a slot in which an entity can occur. As a result, a collection of contexts is derived. Examples of extracted entity contexts are shown in Table 1 (column 2).

To induce the patterns, we need to determine their starts. It is reasonable to assume that some tokens are more specific to