Domain-independent Model for Chemical Compound and Drug Name Recognition

Utpal Kumar Sikdar, Asif Ekbal, and Sriparna Saha

Department of Computer Science and Engineering
Indian institute of Technology Patna, Bihar, India
E-mail:{utpal.sikdar, asif, sriparna}@iitp.ac.in

Abstract. This paper briefly describes the works that we have carried out as part of our participation in the BioCreative-IV Track-2 shared task on chemical compound and drug name recognition. We submit five runs, all of which are based on the machine learning approaches. As the machine learning techniques we use Conditional Random Field (CRF), Support Vector Machine (SVM) and a simple ensemble technique. Our system is domain-independent in the sense that it does not make use of any domain-specific external resources and/or tools. Here we report the evaluation results for only of those runs where development set is not included as part of the training procedure. We obtain the best performance with a CRF based model that shows the micro average recall, precision and F-score values of 72.80%, 75.82% and 74.28%, respectively. The same model yields the macro average recall, precision and F-core values of 73.96%, 74.22% and 72.47%, respectively.

Key words: Chemical name recognition; CRF; SVM; Domain-independent

1 Introduction

In recent times, information extraction has drawn huge attention to the biological or medical practitioners and researchers. There exists huge amount of un-organized and unstructured web-based data, and every day many documents are being added to it. Therefore organizing, finding and extracting relevant information from such a huge amount of data is an important challenge in our day-to-day life. In life science articles, patents or health agency reports, chemical compounds and drug names like small signal molecules or other biological active chemical substances are the important entity classes. There has been lot of interest to the concerned community to identify mentions of chemical compounds automatically within text as well as to index whole documents with the compounds described in them. The recognition of chemical entities is also crucial for other text mining tasks that include but not limited to the predictions of drug-drug/protein-protein interactions, finding relations to adverse reactions of chemical compounds and their associations to toxicological endpoints or the extraction of pathway and metabolic reaction relations [1].

[1] http://www.biocreative.org/tasks/biocreative-iv/chemdner/

There exists many representations and nomenclatures for chemical names. Some examples are SMILES, InChI and IUPAC, out of which the first two allow a direct structure search, but IUPAC like names are more frequent in biochemical texts. Trivial chemical names can be easily found using a dictionary-based approach and can be subsequently mapped to their corresponding structures. In contrast it is not feasible to enumerate all IUPAC like names. Automatic identification of mentions of chemical compounds in text is of interest for a variety of reasons.

In this paper we report on our participation to the BioCreative-IV Track-2 shared task on chemical compound and drug name recognition. The goal of this task is to promote the implementation of systems that are able to detect mentions of chemical compounds and drugs, in particular those chemical entity mentions that can subsequently be linked to a chemical structure. The task was to provide, for a given document, the start and end indices corresponding to all the chemical entities mentioned in the document.

2 Methods

Our method for drug and chemical name recognition is based on machine learning algorithm. We use CRF [1] and SVM [2] as the learning algorithms. The key focus was to develop systems that could be easily adapted to other domains. We develop five different systems for submission. Three of our models are developed based on CRF, one model is based on SVM, and the rest one is based on a ensemble framework. For ensemble we combine seven models, where apart from the other four submitted runs, we also combine another three CRF based models, generated by varying the feature templates. We identify and implement variety of features, mostly without using any deep domain knowledge or domain-specific external resources and/or tools. We develop our own rule–based technique for sentence splitting and tokenization.

We use the C^{++} based CRF^{++} package [2] for CRF experiments. It models the problem as a sequence learning task. For SVM we use YamCha[3] toolkit, along with TinySVM-0.07[4]. Here, the *pairwise multi-class decision* method and the *polynomial kernel function* are used.

2.1 Features

The success of any machine learning algorithm is based on the feature sets. The features presented here are automatically extracted from the given datasets without using any other source of information.

Context words: These are the words occurring within the context window $w_{i-3}^{i+3} = w_{i-3} \dots w_{i+3}$, $w_{i-2}^{i+2} = w_{i-2} \dots w_{i+2}$ and $w_{i-1}^{i+1} = w_{i-1} \dots w_{i+1}$, where w_i is the current word.

[2] http://crfpp.sourceforge.net
[3] http://chasen-org/ taku/software/yamcha/
[4] http://cl.aist-nara.ac.jp/ taku-ku/software/TinySVM

Word prefix and suffix. These are the word prefix and suffix character sequences of length up to n. We experiment with n=3 (i.e., 6 features) and 4 (i.e., 8 features) both.
Word length. We define a binary-valued feature that fires if the length of w_i is greater than a pre-defined threshold. This filters out very short words.
Infrequent word. A list is compiled from the training data by considering the words that appear less frequently than a predetermined threshold, i.e. 10 in our current experiment. Now, a feature is defined that fires if w_i occurs in the compiled list. This is based on the observation that more frequently occurring words are rarely the chemical names.
Unknown token feature: This is a binary valued feature that checks whether the current token was seen or not in the training data. In the training phase, this feature is set randomly.
Word normalization: This feature indicates how a target word is orthographically constructed. Word shapes refer to the mapping of each word to their equivalence classes. Here each capitalized character of the word is replaced by 'A', small characters are replaced by 'a' and all consecutive digits are replaced by '0'.
Orthographic features: We define a number of orthographic features depending upon the contents of the wordforms. In total, we define 24 features based on the orthographic constructs.
Informative words: Sometimes the words or the sequence of words that precede and follow the chemical names could be useful for mention recognition. From the training set, we extract most frequently occurring words that appear within the context of $w_{i-2}^{i+2} = w_{i-2} \dots w_{i+2}$ of w_i. Thus we create two different lists, one for the informative words that precede the chemical names and the other contains the informative words that follow the chemical names. Thereafter we define two features that fire for the words of these lists.
Chemical prefix and suffix: We extract most frequently occurring prefixes and suffixes of length 2 from the IUPAC entities present in the training data. Thereafter two binary valued features are defined that fire if only if the current token contains any of these prefixes and suffixes.
Dynamic NE information: This is the output label(s) of the previous token(s). The value of this feature is determined dynamically at run time.

3 Experimental Results

Experiments are conducted on the datasets provided for the BioCreative-IV CHEMDNER Track-2 shared task. In order to properly denote the boundaries of multiword chemical names, the class is further divided using the BIO notation, where B, I and O denote the beginning, intermediate or outside tokens of an entity. As in this work our goal was to identify only the chemical and drug names, so we have only three classes[5].

[5] B-Chem: Beginning of a chemical entity, I-Chem: intermediate tokens of chemical entity, and O-other than chemical entity

The system was trained with the training data and evaluated on the development data. We submitted five runs in total. In three of our submitted runs, development set is merged to the training set. Table 1 shows the results on the development set for the rest two runs. It shows that CRF performs better over SVM with few points.

Table 1. Results on the development set (in %)

Model	Micro Average			Macro Average		
	recall	precision	F-score	recall	precision	F-score
CRF	72.80	75.82	74.28	73.96	74.22	72.47
SVM	70.24	77.81	73.83	71.15	76.05	72.08

A close investigation to evaluation results suggest that most of the errors are due to the tokenization problem. In one experiment we kept the boundary information of the tokens (chemical entities) in the development set unaltered, and it showed the micro average F-score of approximately 85%. But, when we tokenize the same development set with our own method, the performance drops significantly. We observe that sampling the training set by removing the sentences that do not contain chemical names could improve the performance. The shallow parsing features such as Part-of-Speech (PoS) and chunk information could be effective for the task.

4 Conclusion

In this paper we report about our submitted system as part of our participation in the BioCreative-IV Track-2 chemical and drug name recognition task. Our system is based on machine learning algorithms such as CRF and SVM. These classifiers are trained with a diverse set of features, which are generated without using any domain-specific external resources and/or tools. There are many scopes for further improvement of the system. At present we have used only the features that are extracted automatically from the training data. In future we would like to (i) implement shallow parsing features such as PoS and chunk, (ii) extract some other domain-specific features from the external resources such as PubChem etc. We would also like to carry out thorough sensitivity analysis to find the most relevant set of features for this particular problem.

References

1. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML. pp. 282–289 (2001)
2. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)