CLRG ChemTMiner @ Biocreative V

Sobha Lalitha Devi., Sindhuja Gopalan., Vijay Sundar Ram R.,
Malarkodi C.S., Lakshmi S., Pattabhi RK Rao

Computational Linguistics Research Group, AU-KBC Research Centre
MIT Campus of Anna University, Chromepet, Chennai, India.

sobha@au-kbc.org

Abstract. This paper describes our system developed for Biocreative V Chemical entity mention in patent (CEMP) task for identifying the chemical mentions from patent data. We have presented two systems which use Support vector machines (SVMs) and Conditional random fields (CRFs) algorithms. In this work we used a rich feature set that includes linguistic, orthographical and lexical clue features. We obtained an F-score of 0.817 using CRFs and F-score of 0.877 using SVMs on the development data. And on test data we obtained an F-score of 0.7532, 0.74523 for SVM and CRFs respectively.

Keywords: Chemical named entity recognition; Support vector machines; Conditional random fields;

1 Introduction

Biocreative V CHEMDNER-patents task aims at the extraction of chemical and biological data from medicinal chemistry patents. This task mainly focuses on the identification of chemical entities, detection of patent titles and abstracts that mention chemical compounds and identification of protein and gene related objects. This task may further help in extraction of useful information from patent documents, as the patents contain important research that is difficult to analyze [1]. This paper describes our NER system developed for Biocreative V "Chemical entity mention in the patent (CEMP) task". The main goal of this task is to detect the chemical named entity mentions from patent abstracts. This task exclusively focuses on identifying any wide definition of chemical terms to facilitate more efficient access to information on chemical compounds and drugs from noisy text data [4]. NLP techniques are useful in excerpting information from the scientific literatures, to build a valuable knowledge resource for the bio scientists, where extraction of named entities (NEs) is the primary step.

Biomedical named entity recognition is a sub task of information extraction that aims to identify and classify the biomedical named entities like genes, proteins, chemicals from the text. Because CEMP task mainly aims at identification of chemical entities, we attempt to develop a named entity recognition (NER) system for identification of chemical terms. Identification of chemical entities is difficult due to diverse naming conventions. Chemical entity annotated corpus has been developed in the past by [2],[3],[5].

We preprocessed the data provided by the CEMP task organizers to the required format to develop our NER system. Subsequently, using rich set of features the entities were identified from the corpus using Machine learning (ML) algorithms. In the following section features and the method used to develop the language models are described. Results are discussed in section 3. The paper ends with the conclusion.

2 Method

In this section we present our systems developed using Condition Random Fields (CRFs) [6] and Support Vector Machines (SVMs) [7]. For our work we use CRF++ and Yamcha tool. CRF++ tool is an open source implementation of CRFs and is a general purpose tool. Yamcha is a generic and customizable tool that is using SVMs algorithm. Our NER system is a sequential pipeline where the data is first preprocessed to required format that is needed to train the system. After training the system the NEs are automatically identified from the test set. Finally to improve the efficiency of the system post-processing rules were employed. Features used for our work are explained in the next section.

2.1 Feature Selection

Feature selection is an important step in the ML approach for NER. Features play an important role in boosting the performance of the system. Features selected must be informative and relevant. We have used word level features, grammatical features, functional terms and lexical clue features that are detailed below:

1. Word level features: Word level features include Orthographical features and Morphological features.

    a. Orthographical features contain Capitalization, combination of digits, symbols and words and Greek words.

b. Prefix/suffix of chemical entities is considered as morphological features.

2. Grammatical features: Grammatical features include word, POS, chunks and combination of word, POS and chunk.

3. Functional term feature: Functional term helps to identify the biological named entities and categorize them to various classes. Example: Alkyl, acid, alkanylene

4. Lexical clue feature: Lexicon of chemical entities is collected from various databases. These lexical clues are used as feature for ML algorithms by giving weight to the entities present in the lexicon

2.2 Preprocessing

The CEMP task training and development data is preprocessed using a sentence splitter and tokenizer and is converted into column format with entities tagged using the file containing detailed chemical mention annotation. The entities are classified into abbreviation, formula, identifier, systematic, trivial, family and multiple classes. Since we do not have to provide the type of chemical mention we used a common tag "chemical" for all chemical compounds. POS and chunks are added using automatic tools. After adding POS and chunk to the data, other features are added as consecutive columns. We created a lexicon of unambiguous chemical mentions extracted from the PubChem, KEGG DRUG and KEGG COMPOUND database. This lexical clue features is added to the data by giving weights to the words present in the lexicon. Since the data format is similar for both the algorithms we trained both the systems with the same preprocessed data.

2.3 Named Entity Identification

Using these created models the chemical named entities were automatically identified. Chemical entity mention in patents requires the detection of the start and end indices corresponding to all chemical entities. Hence we converted the output from the system to the required format for evaluation. First, we developed language models using CRFs. After preparing the training data, the features generated in the template file is given along with the training data to the CRFs to learn NE patterns. Using the NER model the NEs can be automatically identified from the testing corpus. Since the task involves the classification of tokens as chemical and non chemical entities we also used SVMs method. Using the features, we trained the SVM system with training data and models were built. The NER results are stored in the CEMP task prediction format. The CEMP task prediction format consists of tab separating column containing patent identifier, offset string with text type, rank of chemical entity, confidence score and string of chemical entity mention.. The development data is again converted to a similar format as training data and is used to evaluate these language models. After analysing the output we postulated post processing linguistic and heuristic rules to improve the performance of the system. The results are tabulated in Table 1.

2.4 Postprocessing

We devised post-processing rules on the system's output to improve the performance. Using surface clues such as capital letters, synonyms, digits and functional terms such as alkyl, alkoxy etc., and using regular expressions the NEs were extracted by devising linguistic rules. System output had tagging inconsistency, where certain entities don't have a start tag and in some cases end tag may not be correct. There are several challenges in the identification of NEs. The chemical abbreviation like AND (Andrographolide), formulas like ON, OR and name of the chemical like "lead powder" creates ambiguity because they share their lexical representation with English words. In order to overcome these challenges we came up with heuristic rules. In case of abbreviation tagging, if long form mention is tagged by the system and is followed by a token within open and closed brackets, then the token would be tagged. In case of entities like "Lead powder", POS of the word and neighbor words can be used to check the ambiguity. The results obtained are discussed in the following section.

3 Results and Discussion

BioCreative evaluation library script is used to evaluate the system's performance. Micro and macro average precision, recall and F score are calculated. We used training data to train the system and evaluated using development data. And also test results are tabulated in Table 1.

Table 1. Results obtained for development data and Test data

S.no	Methods used	Precision		Recall		F-score	
		Dev Data	Test Data	Dev Data	Test Data	Dev Data	Test Data
1	CRFs with features 1,2,3,4	0.791	0.831	0.71	0.675	0.75	0.745
2	SVMs with features 1,2,3	0.802	0.819	0.832	0.694	0.817	0.752
3	SVM with features 1,2,3,4	0.832	0.810	0.802	0.691	0.817	0.745
4	SVMs with features 1,2,3,4 + post-processing	0.894	0.819	0.861	0.695	0.877	0.752
5	SVMs with all 4 features + dev data	-	0.819	-	0.697	-	0.753

The results for development data show that SVMs outperforms the CRFs method. We obtained an F-score of .877 using SVMs and .817 using CRFs for development data. Then we applied post processing rules to the output obtained from SVMs model and obtained an improved F score of .89. The test data results show that model 5 developed using SVMs performed better than other models and obtained an F-score of .753. The results show that the system performs efficiently on chemical data. The system will be made available as a web service, where the system can be accessed using REST service.

4 Conclusion

We have presented our machine learning based system developed for Biocreative V CEMP task. We used two machine learning algorithms SVMs and CRFs for our work. We have used a rich feature set that includes linguistic, orthographical and lexicon based features. Using heuristic and linguistic rules the system's performance was further improved in the development data, but has not shown much difference in the test data. The results need further, deeper analysis by comparing with the gold standard of the test data. In future, we plan to provide access to our system to all interested users as a web service.

References

1. Aras, Hidir et al., Applications and Challenges of Text Mining with Patents. Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN), Hildesheim, Germany, October 8-10, 2014.
2. Friedrich, Christoph M. et al., Biomedical and Chemical Named Entity Recognition with Conditional Random Fields: The Advantage of Dictionary Features. BMC Bioinformatics (7) 2006: 85–89.
3. Klinger, Roman et al., Detection of IUPAC and IUPAC-like chemical names. Bioinformatics 24(13) 2008: 268-76.
4. Krallinger et al., Overview of the chemical compound and drug name recognition (CHEMDNER) task. Proceedings of the Fourth BioCreative Challenge Evaluation Workshop, Washington, DC, USA, October 8, 2013: 2-33.
5. Krallinger, Martin. et al., The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics 7(1) 2015.
6. Kudo, Taku, CRF++, an open source toolkit for CRF, http://crfpp.sourceforge.net, 2005.
7. Kudo, Taku and Matsumoto, Yuj, Use of ', Support Vector Learning for Chunk Identification, Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000: 142-44.