The System for Recognizing Chemical Names and Detecting Chemical Passages in Patent Documents

Shuo Xu[1] and Weijia Xu[2]

[1] Information Theory and Methodology Research Center,
Institute of Scientific and Technical Information of China,
No. 15 Fuxing Rd., Haidian District, 100038 Beijing, P.R. China
Email: xush@istic.ac.cn
[2] Texas Advanced Computing Center, The University of Texas at Austin,
Austin, TX, USA
Email: xwj@tacc.utexas.edu

Abstract. One of the tasks in the BioCreative V challenge, the CHEMDNER-Patent task, includes three subtasks: CEMP, CPD, and GPRO. We participated in the CEMP and CPD subtasks, and developed a system on the basis of selected open-source NLP, machine learning toolkits. In our system, the CEMP subtask is regarded as a sequence labeling problem, and the CPD subtask is regarded as a text classification problem.

Key words: Chemical Entity Mention; Chemical Passage Detection; Gene and Protein Related object; Conditional Random Fields; Topic Models

1 Introduction

It is very crucial to improve information access on biomedical relevant entities such as chemical compounds, genes and proteins described in patent documents. For example, pharmaceutical patents covering chemical compounds provide information on their therapeutic applications and, in most cases, on their primary biological targets, which can help speed-up the early-stage medicinal chemistry activities [1]. Despite the valuable characterizations of biomedical relevant entities contained in patents, academic research in the area of text mining and information extraction using patent data [2, 3] has been minimal.

The identification and integration of all information contained in these patents (e.g., chemical structures, their synthesis and associated biological data) is currently a very hard task not only for database curators but for life sciences researches and biomedical text mining experts as well. The BioCreative (Critical Assessment of Information Extraction Systems in Biology) challenge is a community-wide effort to build an evaluation framework for assessing text mining systems in biological domains [4]. The CHEMDNER-patents challenge in

BioCreative V was the first time that a biomedical text mining community challenge handles noisy text data (patents) and could result in software that helps to derive annotations from patents.

The CHEMDNER-patents challenge has three subtasks, CEMP (chemical entity in patents) subtask, CPD (chemical passage detection) subtask, and G-PRO (gene and protein related object) subtask. The CEMP subtask is the main task to detect all chemical named entities described within a patent document. The CPD subtask is the text classification task to detect whether a patent document mentions chemical compounds or not. GPRO subtask is the task to identify gene and protein related objects mentioned within a patent document. Here, we present the method and recognition system for CEMP and CPD subtasks.

2 System Description and Methods: CEMP Subtask

As shown in Fig. 1, our system for the CEMP subtask detects sentence boundaries on the title and abstract of each patent document, and then tokenizes each detected sentence during pre-processing steps. Next, our system extracts chemical entity references with a conditional random field (CRF) approach, followed by post-processing steps including a rule-based approach and a format conversion step.

We follow the pre- and post-processing steps in [5, 6], but add several rules for the tokenization. Some examples include: (a) *sulfate* is split into *sulfat* and *e*, (b) *withsinomenine* is split into *with* and *sinomenine*, (c) *ArI* is split into *Ar* and *I*.

2.1 Tagging Scheme

Our system formalize the CEMP subtask as a sequence labeling problem, where each token in a sentence is labeled with a tag that denotes whether a token is part of a chemical name and its position in a chemical name. The CRF (conditional random field) is a typical undirected probabilistic model, and has been demonstrated to be superior to other machine learning methods for NER (named entity recognition). An open source implementation of CRF, CRF++[1], is used in our system.

There are several tagging schemes, such as **BIO**, **BIEO**, **BIEOS** and so on, which label each token as being the beginning of (**B**), the inside of (**I**), the end of (**E**), entirely outside of (**O**) an entity, or a single-token entity (**S**). Here, we used **BIEO** tagging scheme in our system. Although the CHEMDNER-patents challenge classifies the annotated entities into one of seven classes $\mathbb{C} = \{$ SYSTEMATIC, IDENTIFIER, FORMULA, TRIVIAL, ABBREVIATION, FAMILY, MULTIPLE $\}$, we do not consider the entity class information.

[1] CRF++ can be available from http://crfpp.googlecode.com/svn/trunk/doc/index.html.

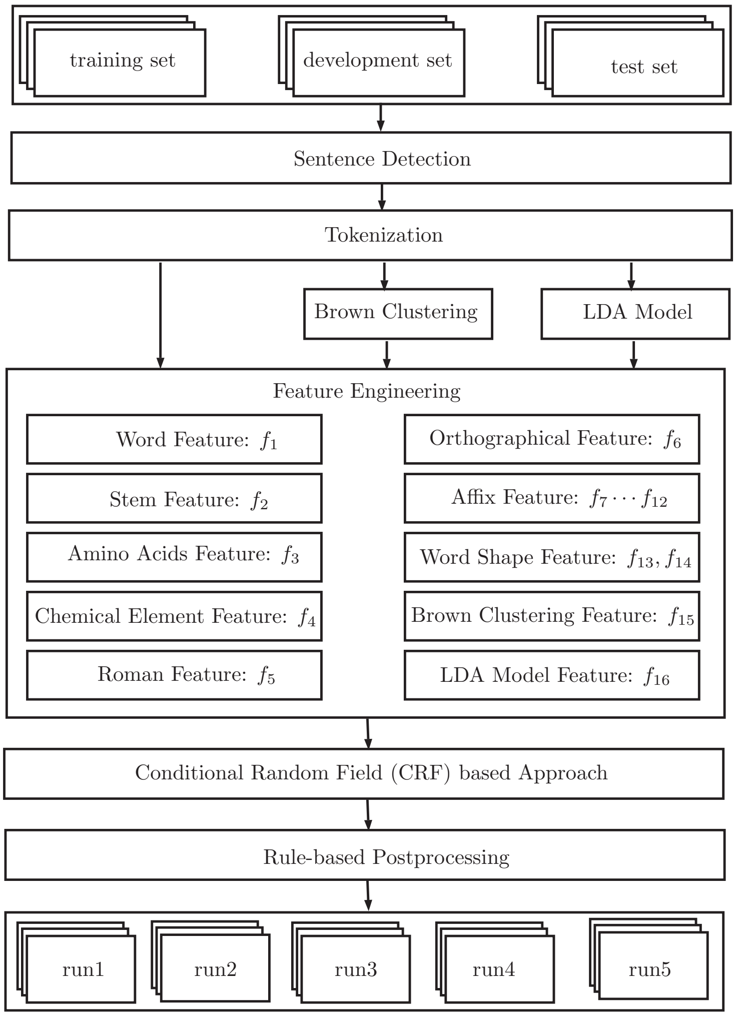
training set
development set
test set
Sentence Detection
Tokenization
Brown Clustering
LDA Model
Feature Engineering
Word Feature: f_1
Stem Feature: f_2
Amino Acids Feature: f_3
Chemical Element Feature: f_4
Roman Feature: f_5
Orthographical Feature: f_6
Affix Feature: $f_7 \cdots f_{12}$
Word Shape Feature: f_{13}, f_{14}
Brown Clustering Feature: f_{15}
LDA Model Feature: f_{16}
Conditional Random Field (CRF) based Approach
Rule-based Postprocessing
run1
run2
run3
run4
run5

Fig. 1. The system processing pipeline for the CEMP subtask.

2.2 Feature Engineering

The following features are utilized for the CEMP subtask.

General linguistic features. Our system includes the original tokens, as well as stemmed tokens, as features using the Porter's stemmer from Stanford CoreNLP[2].

Character features. Since many CEMs contains numbers, Greek letters, Roman numbers, amino acids, chemical elements, and special characters, our system calculates presence or absence of Roman numbers, amino acids (name, 3-char abbreviation, 1-char abbreviation), or chemical elements.

Orthographical feature. Word tokens are classified into six classes {All-capitalized, Is-capitalized, All-lowercase, All-digits, Greek, Others} based on regular expression.

Affix feature. Prefixes and suffixes of the length of 3, 4, 5.

Token shape feature. Similar to [7], two types of token shapes: *generalized token class* and *brief token class* are used. The generalized token class maps any uppercase letter, lowercase letter, digit and other character in a word to 'X', 'x', '0' and 'O' respectively, while the brief token class maps consecutive uppercase letters, lowercase letters, digits and other characters to 'X', 'x', '0' and 'O' respectively.

Word representation features. In order to include unsupervised word representation, we used Brown clustering method [8] on the training, development and testing set to generate 500 clusters. Intuitively, the Brown clustering method can merge the tokens with similar contexts into the same cluster. Following Huffman coding, a particular token can be assigned a binary string representation. Thus, the more similar the prefix of the token's Huffman coding, the more similar the tokens are. The implementation of Brown clustering method by Liang[3] is adopted in our system.

LDA model features. Though the Brown clustering method can assign one Huffman coding for each token, it is difficult to reflect the ambiguity of each token.Therefore, we also run LDA modeling on the training, development and testing set with 200 topics and 2000 iterations. The topic assign is taken as the resulting feature for each token.

Contextual features. For each token, our system includes a combination of the current token and previous token or next token (bigram).

Feature combination. Our system also includes combination features that are generated using multiple individual features listed above, such as the combination of token feature with LDA model feature, the combination of token stem feature with LDA model feature, and so on.

[2] Stanford CoreNLP can be available from http://nlp.stanford.edu/software/corenlp.shtml

[3] Brown clustering can be available from https://github.com/percyliang/brown-cluster

2.3 Experimental Setup

In CRF++, there are 4 major parameters (”-a”, ”-c”, ”-f” and ”-p”) to control the training condition. In our submitted predictions, the parameters ”-a”, ”-f” and ”-p” are set to CRF-L2, 2 and 4, respectively. The option ”-c” trades the balance between over-fitting and under-fitting. The prediction results are heavily affected by the values of those parameters. Ideally, the optimal setting can be identified by cross validation experiments. But due to time constraints, we just set ”-c” option to some fixed value. Table reports the specific setting for our submitted 5 runs. Here, for convenience, the features beyond word representation feature and LDA model feature are collectively referred to as basic features.

Table 1. The specific setting in our system for the CEMP subtask.

	Basic Features	Word Representation	LDA Model	c
Run1	√			1.0
Run2	√	√		1.0
Run3	√	√		2.0
Run4	√	√	√	1.0
Run5	√	√	√	4.0

3 System Description and Methods: CPD Subtask

In our system, the CPD subtask is modeled as a text classification problem. The maximum entropy (MaxEnt) [9] (run1, run4 with Gaussian prior variance = 1.0 and 2.0, repsecitvely), balanced Winnow [10] (run2), naïve Bayes [11] (run3) are used for classification in our system. These three classification methods can output the posterior class probability. The submitted result for run5 is the average from the other 4 runs. The implementation of the above methods in Mallet[4] are used in our system.

4 Conclusions

We participated in the CEMP and CPD subtasks of CHEMDNER-Patent challenge in BioCreative V and developed a system for these two subtasks. In our system, the CEMP and CPD subtasks are regarded as a sequence labeling problem and a text classification problem respectively. Therefore, the famous CRF model is utilized to solve the sequence labeling problem, and the MaxEnt, balanced Winnow and Naïve Bayes are used to solve the text classification problem.

[4] Mallet can be available from http://mallet.cs.umass.edu/index.php

Acknowledgments. This work was supported partially by National Science Foundation of China: Research on Technology Opportunity Detection based on Paper and Patent Information Resources (71403255).

References

1. Muresan, S., Petrov, P., Southan, C., Kjellberg, M.J., Kogej, T., Tyrchan, C., Varkonyi, P., Xie, P.H.: Making every SAR point count: The development of chemistry connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today **16**(23–24) (2011) 1019–1030
2. Jessop, D.M., Adams, S.E., Murray-Rust, P.: Mining chemical information from open patents. Journal of Cheminformatics **3**(1) (2011) 40
3. Lai, H., Xu, S., Zhu, L.: Chemical and biological entity recognition system from patent documents. In: Proceedings of the 2nd International Workshop on Patent Mining and its Application (IPaMin). (2015)
4. Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., Valencia, A.: Evaluation of text-mining systems for biology: Overview of the second BioCreative community challenge. Genome Biology **9**(Suppl 2) (2008) S1
5. Xu, S., An, X., Zhu, L., Zhang, Y., Zhang, H.: A CRF-based system for recognizing chemical entities in biomedical literature. In Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A., eds.: Proceedings of the 4th BioCreative Challenge Evaluation Workshop. Volume 2. (2013) 152–157
6. Xu, S., An, X., Zhu, L., Zhang, Y., Zhang, H.: A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. Journal of Cheminformatics **7**(Suppl 1) (2015) S11
7. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Stroudsburg, PA, USA, Association for Computational Linguistics (2004) 104–107
8. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based *n*-gram models of natural language. Computational Linguistics **18**(4) (1992) 467–479
9. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM (2008) 595–602
10. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine Learning **2**(4) (1988) 285–318
11. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29**(2–3) (1997) 103–130