HTSZ_CEM System for Chemical Entity Mention Recognition in Patents

Junzhao Bu, Buzhou Tang*, Xiaolong Wang, Qingcai Chen, Zengjian Liu, Haodi Li

Intelligent Computing Research Center,
Harbin Institute of Technology Shenzhen Graduate School, China

bujunzhao@outlook.com; tangbuzhou@gmail.com;
wangxl@insun.hit.edu.cn; qingcai.chen@gmail.com;
liuzengjian.hit@gmail.com; haodili.hit@gmail.com

Abstract. In this paper, a machine learning-based system was proposed for the challenge task of chemical entity mention recognition in patents (CEMP) in BioCreative V. The CEMP task was recognized as a sequence labeling problem and conditional random fields (CRF) were employed for it. Evaluation on the CEMP challenge corpus showed that our system (team 293) achieved a micro F-measure of 87.03%.

Keywords. chemical entity mention recognition in patents, sequence labeling problem, conditional random fields;

1 Methods

Dataset

The CEMP task organizers of BioCreative V annotated 21,000 patent abstracts, of which 7,000 records were used for training, 7,000 records for development and the remaining 7,000 records for test.

Overview of system

Our system consisted of three components: preprocessing module, named entity recognition (NER) module and postprocessing module. Given patent records, the preprocessing modules split them into sentences and tokenized the sentences. Then chemical named mentions were recognized by the NER module based on CRF [1]. Finally, postprocessing module adjusted the results according to the abbreviation lists generated by Ab3P [2].

*Corresponding author

In this study, we considered the CEMP task as a sequence labeling problem and employed CRF for it. The features used in the NER module were shown in Table1.

Table 1. Features used in the name entity recognition module.

Feature	Description
Bag of words	Unigrams, bigrams and trigrams of tokens in window of [-2, 2].
POS tags	Unigrams, bigrams and trigrams of POS tags of the tokens in window of [-2, 2]. GENIA Tagger (http://www.nactem.ac.uk/GENIA/tagger/) was used for POS tagging.
Sentence information	Length of the current sentence. Whether there is any bracket unmatched in the current sentence?
Semantic information	Whether the current token contains alkane stems (e.g. “meth,” “eth,” “prop” and “tetracos”), trivial rings (e.g. “benzene,” “pyridine” and “toluene”), and simple multipliers ("di,” “tri” and “tetra”), as mentioned in [3].
Prefix and suffix	Prefixes and suffixes of length from 1 to 5.
Section information	Which section the current token belongs to, title or abstract?
Word Shapes	In the current token, any or consecutive uppercase letter(s) is/are replaced by "A", lowercase letter(s) by "a", digits by "0" and others by "#", respectively.
Orthographical features	Whether the current word is an all Caps word, contains a digit or not, has uppercase characters inside, etc.
Domain knowledge	Whether the current token contains any prefix/suffix of chemical compounds, drugs, proteins, etc?
Character counts feature [4]	Number of characters, number of digits, number of uppercase and lowercase letters and number of lowercase letters.
Word representation features [5]	Brown clustering (https://github.com/percyliang/brown-cluster); Word2vec (https://code. google.com/p/word2vec/).
Results of another system	We developed a CRF-based chemical entity mention recognition (CEM) system on the corpus of the CEM task of BioCreative IV, and used the results of this system as features.
Character n-grams	Character n-grams of length from 2 to 4.

2 Results

In this challenge, each team was allowed to submit five runs. We submitted two runs: 1) a baseline system without using word representa-

tion features (WRs) and results of another system; 2) a system using all features. The system using all features achieved a precision of 87.85%, a recall of 86.23% and an F-measure of 87.03%, which was much better than the baseline system with a precision of 86.15%, a recall of 84.78% and an F-measure of 85.46%, indicating that WRs and results of the system on the corpus of the CEM task of BioCreative IV are beneficial to CEMP.

REFERENCES

1. Okazaki, Naoaki. CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs), 2007.
2. Sohn S, Comeau DC, Kim W, Wilbur WJ: Abbreviation definition identification based on automatic precision estimates. BMC Bioinformatics 2008, 9:402.
3. Lowe DM, Corbett PT, Murray-Rust P, Glen RC: Chemical name to structure: OPSIN, an open source solution. Journal of chemical information and modeling 2011, 51(3):739-753.
4. Leaman R, Wei C H, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization.[J]. J Cheminform, 2015, 7(Suppl 1).
5. Tang B, Cao H, Wang X, et al. Evaluating word representation features in biomedical named entity recognition tasks[J]. BioMed research international, 2014, 2014.