HITextracter System for Chemical and Gene/Protein Entity Mention Recognition in Patents

Zengjian Liu, Xiaolong Wang, Buzhou Tang*, Qingcai Chen, Xue Shi, Jiankang Hou

Key Laboratory of Network Oriented Intelligent Computation,
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China

liuzengjian.hit@gmail.com;wangxl@insun.hit.edu.cn;
*tangbuzhou@gmail.com; qingcai.chen@gmail.com;
603272564@qq.com; 1207005877@qq.com

Abstract. In this paper, a hybrid system was proposed for chemical entity mention recognition (CEMP) and gene/protein related object recognition (GPRO) in BeCalm challenge. Firstly, five individual machine learning-based subsystems were developed to identify chemical and gene/protein related entity mentions, that is, a bidirectional LSTM (long-short term memory, a variant of recurrent neural network)-based subsystem without any manually-crafted feature, a bidirectional LSTM-based subsystem with some manually-crafted features, a bidirectional LSTM-based subsystem with orthographic features learning, a CRF (conditional random field)-based subsystem and a SSVM (structured support vector machine)-based subsystem. Then, an ensemble learning-based classifier was deployed to combine all the results predicted by above individual subsystems. Evaluation on the official test set showed that the best F1-scores achieved by our system are 90.37% on CEMP, 76.34% on CPRO type 1 respectively.

Keywords. Chemical entity mention recognition; gene and protein related object recognition; sequence labeling problem; conditional random fields; recurrent neural network; ensemble learning

1 Introduction

Chemical patents contain a wealth of chemical and biochemical knowledge, such as chemical compounds, genes and proteins. The BioCreative V challenge [1] has aimed to evaluate and encourage the development of tools to extract these information from patents. To enable a more robust evaluation platform (Biomedical Annotation Metaserver, BeCalm), the BioCreative V.5. BeCalm. challenge also was organized

* * Corresponding author

with three tasks: CEMP (Chemical Entity Mention recognition), GPRO (Gene and Protein Related Object recognition), and TIPS (Technical interoperability and performance of annotation servers). We participated in the CEMP and GPRO tasks, and developed a hybrid system based on five individual entity recognition methods.

2 Methods

Dataset

The BeCalm challenge organizers provided total 30,000 manually annotated patents for the CEMP and GPRO tasks, 21,000 out of which are used as a training set and the remaining 9,000 as a test set. In the training set, 99,632 chemical entity mentions and 17,751 gene/protein related objects were annotated. Table 1 shows the numbers of instances of each type in both CEMP and GPRO tasks.

Table 1. Numbers of instances of each type in CEMP and GPRO tasks.

CEMP		**GPRO**	
Type	**Number**	**Type 1**	**Number**
FAMILY	36,238	ABBREVIATION	7,543
SYSTEMATIC	28,580	FAMILY	5,030
TRIVIAL	25,927	FULL_NAME	4,842
FORMULA	6,818	MULTIPLE	178
ABBREVIATION	1,373	NESTED	89
MULTIPLE	418	NO_CLASS	45
IDENTIFIER	278	SEQUENCE	23
		IDENTIFIER	1
Total	99,632	**Total**	17,751

Overview of system

Our system, as shown in Figure 1, consists of seven components: a tokenization module, five individual modules for chemical and gene/protein entity mention recognition respectively, and an ensemble module to combine all results of above individual modules. Given a record with title and abstract, the tokenization module first split each

sentence into tokens. Then, five individual methods were used to identify the chemical and gene/protein entity mentions. Subsequently, the ensemble module used a stacked ensemble learning-based classifier to combine the results of all above individual modules. We will introduce these core modules of our system in the next few sections.

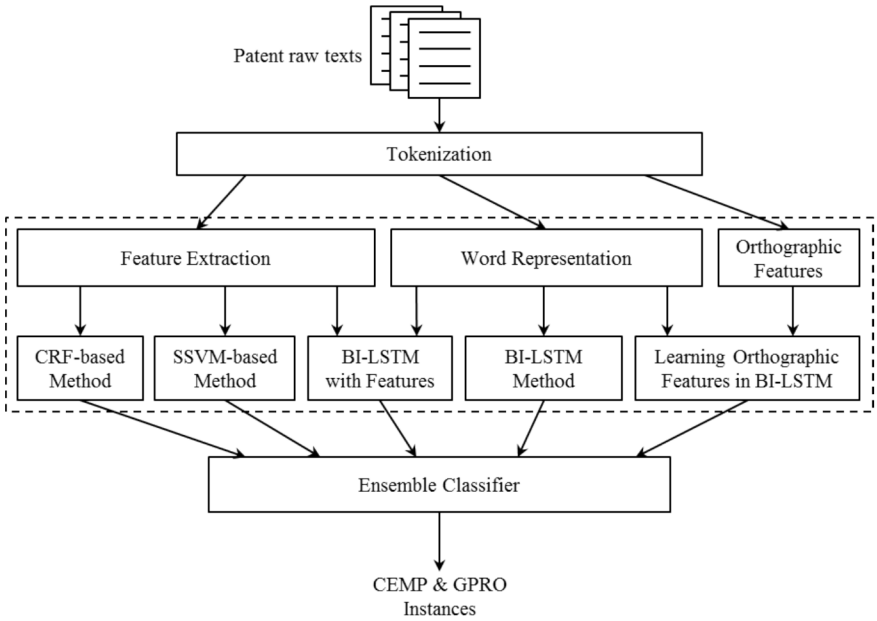

Figure 1. Overview architecture of our system.

Tokenization

After the analysis of raw text, we found that there are lots of unexpected phrases in patents for the existing tokenization tools (e.g. MedEx), for example, the phrase "N-(4-hydrosulfonimidoylphenyl)-5-(trifluorometh yl)pyrimidin-2-amine" is hard to be tokenized. Therefore, in our system for the BeCalm challenge, we employed a new tokenization module. Firstly, we split sentences into tokens by blank spaces, then further separate consecutive numbers, consecutive letters and other characters. For example, above phrase was tokenized into "N - (4 - hy drosulfonimidoylphenyl) - 5 - (trifluorometh yl) pyrimidin - 2 - amine". This method can effectively avoid the boundary errors between predicted instances and gold instances caused by improper tokenization module.

CRF&SSVM-based methods

As our previous work [2], we proposed a CRF-based and a SSVM-based methods for the recognition of chemical and gene/protein entity mentions with above tokenization module. The same set of features were used in both these methods, which are listed in Table 2. We use "CRFsuite" [3] as the implementation of CRFs, and "hmm-svm" [4] as the implementation of SSVM.

Table 2. The features used in our CRF&SSVM-based methods.

Feature	Description
Bag-of-words	The unigrams, bigrams and trigrams of words within a window of [-2, 2].
Part-of-speech (POS) tags	The POS unigrams, bigrams and trigrams within a window of [-2, 2]. We use GENIA tagger for the POS tagging.
Combinations of words and POSs	$\{w_0p_{-1}, w_0p_0, w_0p_1, w_0p_{-1}p_0, w_0p_0p_1, w_0p_{-1}p_1, w_0p_{-1}p_0p_1\}$, where w_0 denotes the current word, and p_{-1}, p_0 and p_1 denote the last, current and next POS tags respectively.
Sentence features	The number of words in the sentence, whether there is an end mark at the end of the sentence such as '.', '?' and '!', whether there is any bracket unmatched in the sentence.
Semantic features	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 []
Affixes	All prefixes and suffixes of length from 1 to 5.
Section features	Which section the token belongs to, title or abstract?
Word Shapes	Any or consecutive uppercase character(s), lowercase character(s), digit (s) and other character(s) in the current word is/are replaced by 'A', 'a', '#' and '-' respectively.
Orthographical features	Whether the word is upper case, has uppercase characters inside, has punctuation marks inside, has digit inside, the word is Roman or Arabic number, etc.
Domain knowledge	Whether the current token contains any prefix/suffix of chemical compounds, drugs, proteins, etc.
Character features	Number of characters, number of digits, number of uppercase and lowercase letters and number of lowercase letters.
Character n-grams	Character n-grams of length from 2 to 4.

Bidirectional LSTM method

A bidirectional LSTM, which has been successfully applied in several sequence labeling tasks [5-7], was deployed for the CEMP and GPRO tasks. It contains three main layers: 1) input layer, which generates the representation of each word in a sentence, and contains two parts: character-level representation and token-level representation; 2) LSTM layer, which includes a forward LSTM and a backward LSTM, takes the word representation sequence of a sentence as input, and outputs a new word representation sequence that captures the context information of each word in this sentence; 3) inference layer, which captures the dependencies between successive labels by keeping a label transition matrix, and predicts the best label sequences with correct structures. The architecture of our BI-LSTM is same as Lample's (2016) [6] for name entity recognition.

Bidirectional LSTM with Feature

In order to incorporate some significant features, we extended the above BI-LSTM model by adding a hidden layer after the LSTM layer [7, 8], which concatenates the word representations generated by LSTM layer and the feature representations together. The features used in the BI-LSTM with feature model are: POS tags, sentence features, semantic features, section information, and domain knowledge features, which are same as the features used in the CRF-based method.

Bidirectional LSTM with Orthographic features learning

To further capture the orthographic information of tokens, we extended the inputs of our BI-LSTM model refer to [9]. Firstly, the orthographic feature of each token was generated by mapping any uppercase character, lowercase character, digit and other character in the word to 'A', 'a', '#' and '-' respectively. For example, the orthographic feature of "1-6C-alkyl" is "#-#A-aaaaa". Then, as the word representations, we also generated the token-level and character-level representations of orthographic features, and concatenated them with the word representations together as the inputs of our BI-LSTM model.

Ensemble Learning Method

To take full advantages of above individual methods, we used an ensemble learning method [10], support vector machine, to merge all results of them. The goal of the ensemble learning method is to determine

whether a predicted CEMP/GPRO instance is a true instance, and the features used in this method includes

- Whether the text spans of a instance exactly match with others?
- Whether the text spans of a instance exactly match with others of the same type?
- Whether the text spans of a instance partially match with others?
- Whether the text spans of a instance partially match with others of the same type?
- Whether a instance contains a conjunction or preposition?
- Which methods have predicted current instance?
- How many times a instance was predicted?
- How many times the span of a instance was predicted?
- The number of tokens in a instances.
- How many times a instance was predicted in same patent?

3 Results

In the BeCalm challenge, we were allowed to submit five runs for CEMP and GPRO tasks respectively. The results of these different runs we submit are listed in Table 3, where "Ensemble-RNN" means combining the results of three RNN-based methods, and "Ensemble-ALL" combines the results of all five methods. The best micro F1-scores achieved by our system are 90.37% on CEMP, 76.34% on CPRO type 1 respectively.

Table 3. The results of our systems for both CEMP and GPRO tasks.

Run	**CEMP**			**GPRO type 1**		
	Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
BI-LSTM	88.97	91.82	90.37	75.23	77.49	76.34
BI-LSTM with feature	88.70	91.28	89.97	79.64	65.89	72.12
BI-LSTM with orthographic	88.91	91.28	90.08	78.25	70.32	74.07
Ensemble-RNN	91.25	88.02	89.60	83.50	60.30	70.03
Ensemble-ALL	91.42	88.56	89.97	83.38	66.38	73.92

4 Acknowledgment

This paper is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundations of China) (61573118, 61402128, 61473101, 61472428), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140508161040764, 20151013161937, JCYJ20140417172417105, JCYJ20140627163809422, JSGG20151015161015297 and JCYJ20160531192358466), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052), Program from the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (93K172016K12) and CCF-Tencent Open Research Fund (RAGR20160102).

REFERENCES

[1] Wei C.-H., Peng Y., Leaman R., Davis A.P., Mattingly C.J., Li J., Wiegers T.C. and Lu Z., Overview of the BioCreative V chemical disease relation (CDR) task, Proc. Proceedings of the fifth BioCreative challenge evaluation workshop, 2015, pp. 154-166.

[2] Liu Z., Chen Y., Tang B., Wang X., Chen Q., Li H., Wang J., Deng Q. and Zhu S., Automatic de-identification of electronic medical records using token-level and character-level conditional random fields, Journal of Biomedical Informatics, vol. 58, 2015, pp. S47-S52.

[3] Okazaki N., CRFsuite: a fast implementation of conditional random fields (CRFs), 2007. URL http://www.chokkan.org/software/crfsuite/.

[4] Joachims T., Finley T. and Yu C.-N.J., Cutting-plane training of structural SVMs, Machine Learning, vol. 77, no. 1, 2009, pp. 27-59.

[5] Ma X. and Hovy E., End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, arXiv preprint arXiv:1603.01354, 2016.

[6] Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C., Neural architectures for named entity recognition, arXiv:1603.01360, 2016.

[7] Huang Z., Xu W. and Yu K., Bidirectional LSTM-CRF models for sequence tagging, arXiv preprint arXiv:1508.01991, 2015.

[8] Chiu J.P. and Nichols E., Named entity recognition with bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics, vol. 4, 2016, pp. 357-370.

[9] Limsopatham N. and Collier N., Learning orthographic features in bi-directional lstm for biomedical named entity recognition, BioTxtM 2016, 2016, pp. 10.

[10] Youngjun Kim E.R., Stacked Generalization for Medical Concept Extraction from Clinical Notes, Proc. Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), Association for Computational Linguistics, 2015, pp. 61-70.