# Recognition of Chemical Entity Mention in Patents Using Feature-rich CRF

Yuankai Guo[1], Shiyi Zhao[2], Chen Qu[3], Lishuang Li[*4]

School of Computer Science and Technology, Dalian University of Technology, China

[1]guoyuankai@mail.dlut.edu.cn;

[2]zhaoshiyi2013@mail.dlut.edu.cn;

[3]quchen0502@gmail.com;

[*4]lilishuang314@163.com

**Abstract.** Chemical named entity recognition is the preliminary groundwork for scientific research and biomedical application. For Chemical Entity Mention in Patents(CEMP), a subtask of the BioCtreative V.5, we implement a CRF++ template trained with a set of features including general linguistic features and chemical characteristics. Our system performs with an F-score of 82.45% on test dataset.

**Keywords.** CEMP, CRF, Named Entity Recognition, Machine Learning

## 1 Introduction

Chemical entity mention recognition is a fundamental step in biomedical research, playing a critical role in the related biomedical researches. For this reason, the BioCreative V.5 Challenge sets Chemical Entity Mention Recognition subtask, which requires the recognition of chemical named entity mentions in text, with a training set of 21,000 patent abstracts and a test set size of 9,000. For this subtask, three main methods are often utilized,which are dictionary-based methods, rule-based methods and statistical machine learning methods. The machine learning methods are increasingly used in Named Entity Recognition (NER) for the good performance and robustness. Combined with the characteristics of Maximum Entropy Model (MaxEnt) and Hidden Markov Model (HMM), conditional random field algorithm (CRF) has achieved good performance on sequence labeling problem. So we treat this subtask as a problem of

*Corresponding author.

sequence labeling and build a CRF based system with rich features to solve it, achieving an F-score of 82.45% on test dataset.

## 2 System Description

### 2.1 System Architecture

Our system consists of four components as Fig.1: preprocessing feature extraction, train & test, and post processing. The preprocessing module utilizes tokenization to produce the labeled train set. The second module extracts the feature. The third module is a training and prediction process using CRF++. At last the post processing module is utilized to refine the results.

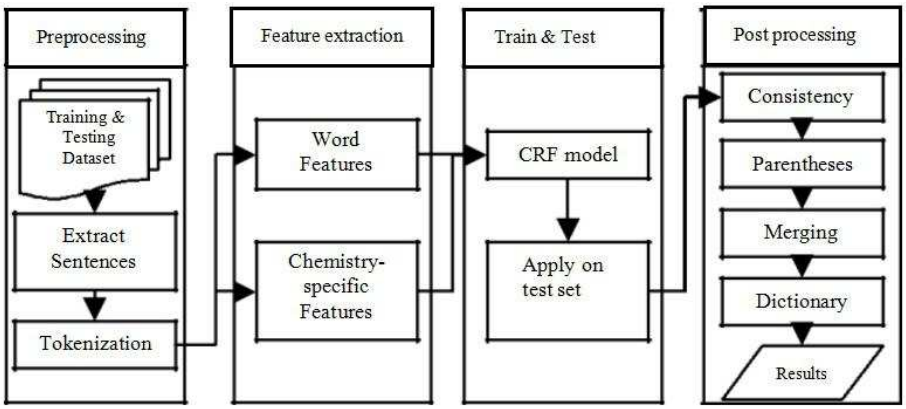


Fig.1 System Architecture

### 2.2 System Features Extraction

The features in our approach are described as Table 1. In feature extraction module, we mainly extract two category features, general linguistic features, shown as the first seven lines, and chemical characteristics, as the rest of the table.

Table1:feature sets

| Feature | Description |
|---|---|
| GENIA features | The original word and stems along with Part-of-speech tag provided by GENIA tagger |
| Affix | Prefixes and suffixes (length: 2 to 4) are extracted as features. |
| Word Shape[1] | Pattern of the word and its brief version. |
| Morphological feature[1] | Number of specific characters: total characters, lower case ones, upper case ones and digits. |
| Word Length | The length of the word (lens:1, |

| | |
|---|---|
| | lens:2, lens:3-5, lens:6+) |
| Vowels | The distribution of vowels. For example, “carbon” is extracted to “-a--o-” |
| Orthographical feature[1] | The classification of the token consists of 31 categories.<br>Word Clustering: Brown Clustering and its prefixes (length: 6 to 8) |
| Element Symbols | We create a lexicon of element symbols for symbol recognition.<br>Chemical Elements: Whether current token is a chemical element |
| Semantic feature[2] | Characristics specific to chemicals, including suffixes (e.g. “-yl”), alkane stems (e.g. “meth”) and trivial rings (e.g. “benzene”) |

### 2.3 Post Processing

- We tag all occurrences of a specific sequence as chemicals if the sequence is tagged by the CRF model at least twice.
- We balance each mention in terms of parentheses and brackets.
- Two mentions will be merged together if they are connected by a single hyphen or chemical bonds in the original text.
- We build a dictionary of chemical identifiers by extracting vocabulary matching specific patterns from CTD (Comparative Toxicoge nomics Database). A token will be recognized as a chemical entity if it can be found in the lexicon.

## 3 Result and Discussion

Our system reports an F-score of 82.45% on test dataset with 83.10% precision and 81.81% recall. Additionally, we also have explored word clusters as a part of features.

## 4 Prospect

Deep learning has been widely applied to tackle NLP related tasks in recent years and has achieved a good performance on various types of tasks. Long-Short Term Memory (LSTM), as a deep learning method, has been widely applied to tackle NLP related tasks in recent years for its excellent ability of learning long-term dependencies. As our earlier work, we adopted the bidirectional recurrent neural network with LSTM unit to identify biomedical entities, achieving an F-score of 88.61% on the BioCreative GM corpus[4]. But for LSTM’s plenty of

parameters which are difficult to tune, we haven't get a better result than the feature-rich CRF solution in CEMP subtask this year. Further experiments will be carried out to achieve satisfactory results.

## 5 Acknowledgments

This work is supported by grant from the National Natural Science Foundation of China (no. 61672126).

## REFERENCES

1. Lishuang Li, Wenting Fan and Degen Huang. "Boosting Performance of Gene Mention Tagging System by Hybrid Methods." Journal of Biomedical Informatics. 45 (2012): 156-164.
2. Leaman, Robert, Chih-Hsuan Wei and Zhiyong Lu. "NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem." Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2 (2014): 34-41.
3. Yuki Tawara， Mai Omura and Mirai MiuraIncorporating. "Incorporating Unsupervised Features into CRF based Named Entity Recognition. " NTCIR 2014.
4. Lishuang Li, Liuke Jin, Yuxin Jiang, Degen Huang. Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. 2016. : 165-176.