FRCRF: A Feature-rich CRF-based Solution for Chemical Entity Mention in Patent Task

Chen Qu\textsuperscript{1}, Tianfu Zheng\textsuperscript{2*}, Liuke Jin\textsuperscript{1}

\textsuperscript{1}School of Computer Science and Technology, Dalian University of Technology, China
\textsuperscript{2}Faculty of Chemical, Environmental and Biological Science and Technology, Dalian University of Technology, China

quchen0502@gmail.com; *lils@dlut.edu.cn

Abstract. Chemical named entity recognition is the preliminary groundwork for scientific research and biomedical application. We build a chemical named entity recognizer to produce our submissions for the BioCreative V CEMP sub-task of the CHEMDNER task. It applies Conditional Random Fields with a rich feature set, including word features and domain specific features. Several post processing modules are also integrated to improve consistency and correct parentheses. Our system performs with an F-score of 82.77\% on test dataset.

1 System Description

1.1 System architecture

Our system consists of four components as Fig.1: (1) A preprocessing module serving as a tokenizer. (2) A feature extraction process to obtain features. (3) A training and prediction process using CRF++. (4) A post processing module to refine the results.

![Fig. 1. System Architecture](image)

1.2 Feature extraction

The features in our approach are described as follows:

- General linguistic features\textsuperscript{[1]}: 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.
• Word Clustering: Brown Clustering and its prefixes (length: 6 to 8).
• Element Symbols: We create a lexicon of element symbols for symbol recognition.
• Chemical Elements: Whether current token is a chemical element.
• Semantic feature\(^{[2]}\): Characteristics specific to chemicals, including suffixes (e.g. “-yl”), alkane stems (e.g. “meth”) and trivial rings (e.g. “benzene”).

1.3 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 Toxicogenomics Database). A token will be recognized as a chemical entity if it can be found in the lexicon.

2 Results and Discussion

Our system reports an F-score of 82.77% on test dataset with 84.31% precision and 80.64% recall. The returned results show that the post processing module helps to improve the F-score by 0.64%.

3 Acknowledgement

The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under No. 61173101, 61173100.

REFERENCES