NCU-IISR System for the CHEMDNER-patents Track at BioCreative V

Yu-Cheng Hsiao\(^1\), Po-Ting Lai\(^2\), Richard Tzong-Han Tsai\(^1\)*

\(^1\) Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C
\(^2\) Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C
104522059@cc.ncu.edu.tw
s102062802@m102.nthu.edu.tw
thtsai@csie.ncu.edu.tw
*Corresponding Author

Abstract. In BioCreative V CHEMDNER-patents track, we propose Conditional Random Fields (CRFs)-based chemical entity mention recognition and chemical passage detection systems for chemical patents. One of the main difficulties in this task is the chemical entity mention is a hierarchy concept which consists of different concepts such as atoms and molecular formula, and different sub-concepts might have different context. We use SOBIE tag set and add an additional S-Atom tag into our tag set to enhance atom recognition. Another is the tokenization problem of the chemical text, and we propose un-tokenized word features which extracted by using un-tokenized sentence. Furthermore, we use retagging approach to collect the chemicals recognized by CRFs-based recognizer to re-annotate whole document. Our best run achieved an F-score of 87.17% on CEMP which ranked 4\(^{th}\), and achieved a sensitivity 98.58% on CPD which ranked 2\(^{rd}\).

Keywords: Named Entity Recognition; Chemical Passage Detection; Conditional Random Fields

1 System Description

Preprocessing: We use the GENIATagger\(^1\) to tokenize sentence, then the regular expression (Regex)-based tokenizer\(^2\) is used to tokenize it again. The twice tokenization approach is used in our previous work [1]. We also used the GENIATagger to generate the Part-of-speech and Chunk tags for extracting features.

Tag set: We use the linear chain Conditional Random Fields model (linear CRFs). We merge all chemical tags into a single tag Chem, and combine the tag with prefix S (Singleton), B (Beginning), I (Inside), E (Ended) or O (Outside) to represent the boundary of named entity. The examples are shown in Fig 1. According to our

\(^1\) http://www.nactem.ac.uk/tsujii/GENIA/tagger/
\(^2\) "\-\-/\/%\*<>\+=~#\"

© Springer-Verlag Berlin Heidelberg 2011
experiments on development set, the atoms are usually missed by the recognizer which use SOBIE tag set. Therefore we add an additional S-Atom tag, which represents the atom, into our tag set to enhance the recognition on the atom.

| Example 1 | ... or/O a/O Ch/B-Chem -I-Chem c/I-Chem alkyl/E-Chem group/O ... |
| Example 2 | ... t/O is/O N/S-Atom /O CH/O or/O CMe/S-Chem ... |

Fig 2. An example of BOUNDARY feature for the tokenized sentence “Application of 1-deoxy-1-veratryl fluorenol in preparing anti-ultraviolet”

**Features Extraction:** We use the same features in our previous work as our baseline. In addition, compare to our baseline’s feature values which are generated from tokenized sentence, we propose un-tokenized word features which are generated from the text that haven’t been tokenized. The un-tokenized word features consist of six orthographical features listed in Table 2 and one boundary feature illustrated in Fig 2. For example, “1,1” is tokenized into “1”, “,” and “1”, and NUM_COMMA feature values are “true”, “true” and “true” and NUM_DASH feature values are “false”, “false” and “false”.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Regular Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQUARE</td>
<td>[.*?]</td>
</tr>
<tr>
<td>PARENTHESES</td>
<td>(.*?)</td>
</tr>
<tr>
<td>TOKEN_COMMA</td>
<td>\S+,\S+</td>
</tr>
<tr>
<td>NUM_COMMA</td>
<td>\d,\d</td>
</tr>
<tr>
<td>NUM_DASH</td>
<td>\d-\d</td>
</tr>
</tbody>
</table>

**Postprocessing:** We merge the training and development set given by BioCreative V CHEMDNER-patents track as our training set. Furthermore, we use retagging approach to collect the chemicals recognized by CRFs-based recognizer and re-annotate the document.

## 2 Results

We participated in both CEMP and CPD of the BioCreative V CHEMDNER-patents track, and five runs were submitted for each stage. Our run1 use our tag set, un-tokenized features and retagging, and it the achieved the best F-score of 87.17% on CEMP which ranked 4th. For CPD task, we return the sentences which contain at least one chemical named entity and achieved the best sensitivity 98.576% on CPD which ranked 2rd.

## References