International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 12 87 – 90

Chinese Chemical Named Entity Recognition and Translation Method Based on Rules

Wen XIONG, and Zi-Hui DING
China Great Wall Computer Shenzhen Company Limited
Beijing, China
Stevens7979@sina.com

Abstract—Aiming at the problem with the recognition of the Chinese chemical named entity (CCNER), such as a long cost for training time, the result relevant to the corpus, and manual corpus annotation. In addition, the noisy words are introduced by the procedure of the Chinese segmentation, which lower the precision of the CCNER. Therefore, we proposed a novel method for CCNER and translation based on rules. First, we employed the Backus-Naur forms (BNF) to represent the rules. Then, we extracted the constituent parts of the BNFs from existing resource semi-automatically. Moreover, we utilized an accumulation method to generate the candidates. Finally, we applied a post-processing for the CCNER and a translation method by the combination the translated texts of the constituent parts using lexicon resources. Experiments on forty abstracts of Chinese's patents indicated: the precision is 78.8%, and the F1 is 78.5%, verifying the effectiveness of the method.

Keywords*-named entity recognition (NER); Backus-Naur forms (BNF); machine translation (MT); Chinese segmentation; text mining (TM); macromolecule compounds*

*****

I. INTRODUCTION

Named entity recognition (NER) is one of the critical tasks of text mining (TM) [1], such as names of human, locations, organizations, time, and events, etc., which reduces the difficulty of the syntactic parsing and semantic analysis. The intact recognition of the special named entities composed of multiple Chinese, English, and digital characters can help the subsequent processing effectively.

Chemical named entity (CNE) is a special named entity (NE) helping researchers to discover new compounds from known CNE. In addition, it contributes to the procedure of the natural language processing (NLP), such as keyword indexing, grammatical analysis, parsing, and machine translation (MT).

The recognition of the English NE, gene expression, and drug name already had a good effectiveness and market prospect, which having many significant mining products appeared in the research. However, the situation of the Chinese's is not as good as that of the English's.

The statistical methods for the CCNER need manually annotated a certain range of corpus covered the majority situation of the CCNER, which is obvious of high cost and difficult to complete. On the other hand, based on the annotation of the corpus, the first step of the recognition method is Chinese segmentation, which introduced the noisy words due to the imprecise algorithm and insufficient word knowledge base. Therefore, this paper proposes a novel CCNER and a translation method based on rules, which was done before the Chinese segmentation, and can avoid the flaw of it. In addition, the method needs not manual annotation for the corpus.

The rest upon the paper is organized as follows. In Section 2, we proposed a novel method for the CCNER. In Section 3, we introduced the experiments. Then, the work of the NER related to the Chemical was described in Section 4. Finally, we gave the summary in the Section 5.

II. METHOD

A. CCNER Based on Rules

We analyzed the forms of the Chinese chemical NE (CCNE) in the Chinese patent, and divided them into three typical representations, such as simple chemical names composed by chemical element, compound chemical names composed by multiple components, and mixed chemical names, including molecular formula.

We extracted some linguistics' components from the existing lexicons, such as bilingual chemistry dictionary for chemical nouns, Chinese characters of chemical elements, general name of Chinese words, number, extended number, punctuation, Grecian characters, and English characters, which can be used for parsing the CCNE according to the rules of the typical representations. In addition, we combined the string matching method with the rule-based method, which scanned the Chinese characters in the sentence from begin to end to generate the components and the candidates. Then, a special post-processing was executed to peel off the illegal fragments in the beginning or on the end of the candidates forming the recognized results.

We proposed the Backus-Naur Form (BNF) for the CCNE by analyzing their composition, which were expressed as follows.

$$
\begin{aligned}
&\text{CCNE ::= < [<N | EN | GC> + [P]] + [CE] + < IMC | SC | CA | CF> + [F] + [N | EN | P] >}\\
&\text{s.t. N ::= <一|二|三|1|2|3| ...>}\\
&\quad \text{EN ::= <对|单|双|全|无| ...>}\\
&\quad \text{GC ::= <}\alpha\,|\,\beta\,|\,\omega\,|\,\text{...>}\\
&\quad \text{P ::= <↓ | 】| ...>}\\
&\quad \text{CE ::= <甲|乙|丙| ...>}\\
&\quad \text{IMC ::= <氟|氢|氧|氨|氮| ...>}\\
&\quad \text{SC ::= <团|粘|系|非|多| ...>}\\
&\quad \text{CA ::= <并用|取代|半熔|可溶|不溶| ...>}\\
&\quad \text{CF ::= <源性|基元|亲水|糖基化|基因座| ...>}\\
&\quad \text{F ::= <基|顺|反|环|异| ...>}
\end{aligned}
\tag{1}
$$

Where: CNE is the Chinese CNE; *N* can be Arabic numerals, or Chinese numerals; EN is the extended Chinese words related to numerals, such as single, double, pair, and whole, etc.; GC is the Grecian characters; P is punctuation and their extension, such as the arrow, and bracket, etc.; CE is the

 ISSN: 2321-8169
87 – 90

Chinese eras; IMC can be the chemical elements; SC is the Chinese characters used in CNEs, such as “group”, “many”, “non”, “series”, “glue”, and “bond”, etc.; CA is the chemical attributes, such as combination, substitution, semi-melt, soluble, and insoluble, etc.; CF is the chemical features, such as source, primitive, hydrophilic, glycosylation, and loci, etc.; F is the chemical fundus, such as “base”, “cis-”, “trans-”, “cyclic”, “neo-”, “iso-”, and “tert-”, etc.

The symbol “<.>” means the inner content will repeat one time or more; “[.]” means the inner content will repeat multiple time or none; “|” means the selection relationship; and “+” means the concatenation of the strings. A sampling CCNE is parsed by using the Eq. (1) as follows.

6－[4［4（2，3－二氯苯基）－1－哌嗪基］丁氧基］－茚满－1－酮

N+P+P+N+P+N+P+N+P+N+P+N+IMC+IMC+F+P+P+N+P+IMC+IMC+F+P+CE+IMC+F+P+P+SC+SC+P+N+P+IMC

Figure 1. A sample of CCNE parsed using the BNF.

On the other hand, we presented another BNF by employing the chemical nouns and the complex substance existed in the lexicon, and combining them as follows, which is a supplement for the Eq. (1).

$$\begin{aligned} &\text{CCNE ::= <CCN | MM | PCN | CEL >} \\ &\text{s.t. CCN ::= <氨基酸|铵盐|臭氧|醋酸钠| ...>} \\ &\quad \text{MM ::= <树脂|醛类聚合物|缩醛树脂|醛醇树脂| ...>} \\ &\quad \text{PCN ::= <玻璃|薄膜|涤纶|尼龙|甘油| ...>} \\ &\quad \text{CEL ::= <对磺酸|多芳烃|酚甲醛树脂|富勒烯衍生物|改性单体| ...>} \end{aligned} \tag{2}$$

Where: CCN is the general chemistry nouns, such as amino acid, ammonium salt, ozone, and natrum acetum, etc.; MM is the macromolecule compounds, such as resin, aldehyde compounds, acetal resin, and aldol resin, etc.; PCN is the informal names of the macromolecule, such as glass, film, terylene, and glycerin, etc.; and CEL is the nouns from chemical engineering, such as on sulfonic acid, arene, fullerene derivative, and Modified monomer, etc., which was accumulated from the open source's lexicons.

B. The Algorithm of the CCNER

The flowchart of the algorithm of the CCNER is as follows.

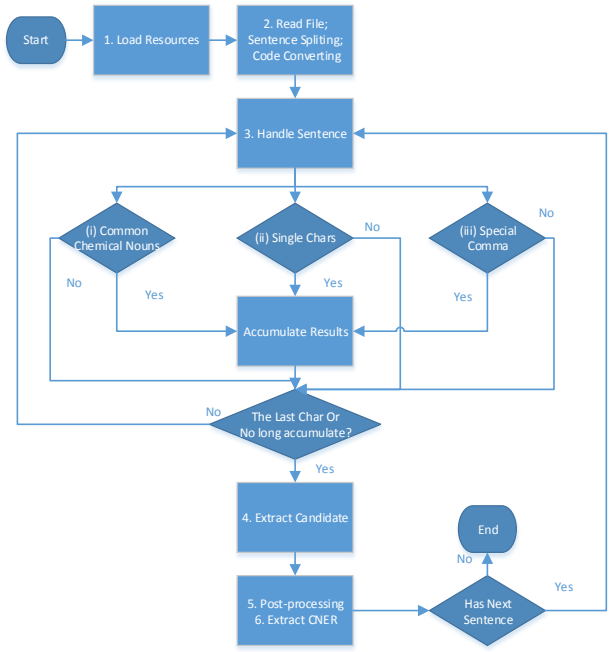

Figure 2. The algorithm of Chinese CCNER.

Where: we describe the steps as follows.

1) Loading resource, such as lexicons of punctuation, Grecian characters, number, extended number, important chemical characters (chemical elements, and transliteration of chemical names), single Chinese characters, Chinese eras, chemical fundus characters, chemical attributes words, chemical feature words, general chemistry nouns, macromolecule compounds, informal names of the macromolecule, and nouns of chemical engineering.

2) Reading file, and splitting the sentence, character code converting.

3) Handling sentence, (I) checking whether the word is a Chinese chemistry word in existing resources or not, and when it is, accumulating it to the recognition results and go to Step 4; (II) checking whether the Chinese word, the number, the punctuation, and the character are in existing resources or not, and when it is, accumulating it to the recognition results and go to Step 4; (III) when it is not in the existing resources, ending the accumulation, and go to Step 4.

4) Extracting candidate, when there is an accumulation, extracting CCNE candidates, and if it is the end to the sentence, go to Step 5; otherwise, go to Step 3, and beginning from the break-point in Step 3.

5) Post-processing, checking whether there is a Chinese character in the IMC set within the accumulation or not, if it is not, the accumulation is an invalid candidate, otherwise, and using the Sub-steps to handling the candidate as follows.

5.1) Head handling, when the character at the beginning is one of the English character, digit, Chinese number, IMC character, or the double characters at the beginning are the same as that at the beginning of words in the CEL, the beginning point is correct.

5.2) Tail handling: when the character at the end is an IMC character, or the double characters at the end are the same as that at the end of the words in the CEL, the end point is correct.

5.3) Peeling off the incorrect point at the beginning, and that at the end of a candidate, where we peel off the first or the last character, and go to 5.1) or 5.2), otherwise, go to the Step 6.

6) Extracting CCNER, when the number of the characters of the result is lower than two, we will abandon it as it is invalid; otherwise, outputting it as a CCNE, and going to Step 3 until the file is ended.

C. The Translation Method of CCNE

We can translate the components of the Eq. 1 as follows.

$$\begin{aligned} &\text{N ::= <一|二|三|1|2|3| ...>} \\ &\text{EN ::= <对(opp)|单(odd)|双(twin)|全(full)|无(non)|...>} \\ &\text{CE ::= <甲(meth)|乙(eth)|丙(prop)| ...>} \\ &\text{IMC ::= <氟(fl)|氢(hydrogen)|氧(oxygen)|氨(ammonia)| ...>} \\ &\text{SC ::= <团(group)|粘(adhering)|系(-based)|非(non-)|...>} \\ &\text{CA ::= <并用(combination)|取代(substituted)|半熔(semi molten)|...>} \\ &\text{CF ::= <源性(genic)|基元(motif)|亲水(hydrophilic)|...>} \\ &\text{F ::= <基(yl)|顺(cis)|反(trans)|环(cyclo)|异(iso)| ...>} \end{aligned} \tag{3}$$

And the components of the Eq. 2 can be translated as follows.

$$\begin{aligned} &\text{CCN ::= <氨基酸(amino acid)|铵盐(ammonium salt)|臭氧(ozone)|...>} \\ &\text{MM ::= <树脂(resin)|醛类聚合物(aldehyde polymer)|缩醛树脂(aidehyde resin)|...>} \\ &\text{PCN ::= <玻璃(glass)|薄膜(film)|涤纶(terylene)|尼龙(nylon)|甘油(glycerin)|...>} \\ &\text{CEL ::= <对磺酸(p-sulfonic acid)|多芳烃(polyaromatic)|酚甲醛树脂(fluosite)|...>} \end{aligned} \tag{4}$$

Since some compounds cannot be accurately translated, we adopted a maximum matching method to translate the CCNE, which formed the readable translation by translating the Chinese characters and Chinese words orderly.

For example, the CCNE in the Fig. 1 can be translated as "6-[4[4(2, 3- dichlorobenzene yl)-1-piperazine yl] butoxy]-indane-1-ketone". Where: Chinese character "Ji" (means "base") is translated as "yl", such as in the "dichlorobenzene yl", and "piperazine yl", etc. The whole translated text is readable and clear.

III. EXPERIMENTS

The experimental data set is forty Chinese patents of chemistry in 2006 selected randomly, which have 7,841 Chinese characters.

We employed the popular evaluation used by information retrieve, such as precision, recall, and F1, which are listed as follows.

$$P = tp/(tp+fp)$$
$$R = tp/(tp+fn) \quad (5)$$
$$F1 = 2\cdot tp/(2\cdot tp+fn+fp).$$

Where: *tp* is the true positive instances if the positive instances are judged as positive categories; *fp* is the false-positive instances if the negative instances are judged as positive categories; *fn* is false-negative instances if the positive instances are judged as negative instances.

For the resources used by above Eq. 1, and Eq. 2, the experiment used some lexicons extracted manually, such as: 11 extended Chinese number for the EN, 24 full-width and half-width punctuations for the P, 22 Chinese eras for the CE, 154 Chinese chemistry characters and transliterations for the IMC, 541 Chinese characters for the SC, 16 Chinese chemistry attributes for the CA, 53 Chinese chemistry features for the CF, 26 Chinese fundus for the F, 151 general Chinese chemistry names for the CCN, 31 Chinese macromolecule compounds for the MM, 55 informal Chinese names of the macromolecule for the PCN, and 26,028 Chinese chemistry nouns with translations for the CEL.

For 5.1) and 5.2), we extracted the first two Chinese characters and the last Chinese characters from the words in the CEL, including 10,463 items for the first, and 8,966 items for the last. We ran the program for the experiment, and obtained the results as follows.

TABLE I. THE RESULTS OF WHOLE MATCH

Total #	CCNE #	Correct #	P %	R %	F1%
133	*132*	*104*	*78.8*	*78.2*	*78.5*

TABLE II. THE RESULTS OF WHOLE AND PARTIAL MATCH

Total #	CCNE #	Correct #	P %	R %	F1%
133	*132*	*112*	*84.9*	*84.2*	*84.5*

From above, the whole correct F1 is 78.5%, and the whole and partial F1 is 84.53%, indicating the effectiveness. We interpreted the reason of the incorrect as follows.

TABLE III. THE SAMPLES OF THE PARTIAL CORRECT AND THE REASON

Original words	Partial correct	Reason
球蛋白	蛋白	"球" (Qiu) is not in the beginning lexicon.
球蛋白基因座	蛋白基因	"因座" (Yin Zuo) is not in the ending lexicon.
淀粉状蛋白源性合成肽	淀粉状蛋白源性	"成" (Cheng) is not included in SC.
硫酸化阴离子	硫酸化	"阴" (Yin) is not included in SC.
寡核苷酸序列	核苷酸序列	"寡" (Gua) is not in the beginning lexicon.
(环)脂族二醇	脂族二醇	"(", "环" (Huan), and ")" is not in the beginning lexicon.
脂质糖基	脂质	"糖基" (Tang Ji) is not in the ending lexicon.

From above, recognition effectiveness can be improved by including more Chinese characters and words in the existing lexicons. In addition, we gave some complicated samples and their translations as follows.

TABLE IV. SOME COMPLICATED SAMPLES RECOGNIZED AND THEIR TRANSLATIONS

Complicated CCNERs recognized	Automatic translations
tetramic 酸型化合物	tetramicsour model chemical compound
直链脂族α, ω- C2 - C12 - 二醇	linear chain aliphatic series α,ω-c2-c12-glycol
NMPRT（烟酰胺磷酸核糖基转移酶）多肽	NMPRT (nicotinic acid amide phosphoribosyl transferase) polypeptide
6，6 - 二甲基 - 3 - 氧杂二环 [3．1．0] 己烷 - 2 - 酮	6,6-dimethyl-3-oxygen hetero dicyclo [3.1.0] hexan-2-ketone

From above, the complicated CCNEs can be recognized by the presented method, and their automatic translations are clear.

IV. RELATE WORK

The algorithm for the CCNER used the conditional random field (CRF) in literature [2], which annotated the chemical corpus by using three kinds of tags, such as beginning tag (B), inner tag (I), and ending tag (E). However, this method needed to annotate on large-scale corpus manually, having long training time, and depending on the fields of the corpus.

In literature [3], they divided the NER systems into four categories as follows.

1) Lexicon-based system, including exact string matching and flexible string matching.

Due to that the capacity of the lexicons is limited, and the composing of the CNEs is complicated; the recall of the exact string matching was low. Therefore, the method of flexible string matching was employed, permitting the editable actions, such as inserting, deleting, and substituting some characters while matching. By using fuzzy matching, the recall of the method is improved, resulting in a popular using in the most NER tasks [4]. However, the system based on the lexicon would have a high precision, but have a low recall due to the spelling mistake and the outdated lexicons. Therefore, it would need costs for the lexicon maintaining to update the lexicons.

2) Rule-based systems, including pattern-based and context-based systems.

These systems utilized a set of hand-made rules, which were used for grammatical (e.g., part of speech (POS)) and syntactic (e.g., word precedence) processing [5]. The rules based on patterns extracted the core terms using surface features, such as capital, number, and special characters. Therefore, the rules of these systems would need updating along with the changing of the domain.

3) ML-based systems, including supervised learning, unsupervised learning, and semi-supervised learning, such as

CRF, support vector machine (SVM), and hidden markov model (HMM) for the former, the clustering for the middle, and bootstrapping for the latter.

The methods based on ML [6] utilized the statistical model, which extracted feature representations using the observed data on the annotated documents. They had two basic steps: annotating step and training step. In [7], CRF method was used for the Chemical NER, and HMM methods were used in [8]. These methods maybe attracted majority research interest during recent years, and data noise, feature extraction, and feature selection would influence the results of them.

Unsupervised learning was not as popular as the others in the NER. Semi-supervised learning used unlabeled and labeled data to start from a trusty seed set, which recognized new contexts according to the similarity of the context, and reused those newly discovered contexts to find newer contexts [9].

4) Hybrid systems, which combined more than one NER methods. For example, the ChemSpot [10] integrated CRF and lexicon-based methods, resulting in better results than that only one method used, due to their superiorities of multiple methods.

V. SUMMARY

The paper proposed a novel CCNER method based on rules, which avoided the influence of the noisy words due to it started before the Chinese segmentation, and utilized the extracted lexicons, such as feature words, and important chemistry characters to parse the rules, where rules were expressed by BNF.

The method has some characteristics, such as non-noisy words introduced, non-manual annotation for the corpus, and extendible, etc. It can be used for NER in TM, and the pre-processing of other NLP, such as Chinese segmentation, MT, to avoid the noise caused by the CCNER.

We will include more nouns to improve the lexicons used during the recognition procedure to enhance the effectiveness, and construct a classifier to handle the invalid characters at the beginning and at the ending instead of the current lexicon matching for the future.

REFERENCES

[1] Ananiadou, Sophia, and J. Mcnaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc. 2005.

[2] Li, Nan, J. M. Ji, and R. T. Zheng. *Recognition of Chemical Names in Chinese Texts. Recent Advances in Computer Science and Information Engineering*. Springer Berlin Heidelberg, 2012:311-318.

[3] Eltyeb, S, and N. Salim. "Chemical named entities recognition: a review on approaches and applications. " *Journal of Cheminformatics* 6.1(2014):17-17.

[4] Klein, and Corinna. "Information Extraction from Text for Improving Research on Small Molecules and Histone Modifications." *Histone Modification* (2011).

[5] Humphreys, R. Gaizauskas, et al. "Description of the LaSIE-II System as Used for MUC7." 1998:127--140.

[6] Chieu, Hai Leong, and H. T. Ng. "Named entity recognition: a maximum entropy approach using global information." *International Conference on Computational Linguistics* Association for Computational Linguistics, 2002:190--196.

[7] Grego, T, et al. "Chemical Entity Recognition and Resolution to ChEBI. " *Isrn Bioinformatics* 2012(2011).

[8] Ponomareva, Natalia, et al. "Conditional random fields vs. hidden markov models in a biomedical named entity recognition task." (2007).

[9] Nadeau, David. "Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision." University of Ottawa, 2007.

[10] Rocktäschel, T, M. Weidlich, and U. Leser. "ChemSpot: a hybrid system for chemical named entity recognition. " *Bioinformatics* 28.12(2012):1633-40.