文章编号：1003-0077（2011）00-0000-00

基于有机物识别和翻译的改进专利机器翻译的方法研究[1]

任智军[1,2]，张威[1]，蒋宏飞[1]，李进[1]，张凯[1]

（1 国家知识产权局中国专利信息中心，北京 100088；2 中国科学技术信息研究所，北京 100038）

摘要：有机化合物系统命名原则的内容很多，显得很繁杂，其表达形式具有多样性的特点，有机物的识别对含有有机物命名实体的专利句子的翻译质量有重要的影响。本文在考虑专利文献中有机物命名实体的特点和构成规律的基础上，将其进行了类别划分并分别制定了相应的识别和翻译规则，最终实现了一个基于有机物识别和翻译的专利文献机器翻译系统。经实验测试，该工具具备较高的有机化合物识别和翻译准确率，提高了含有有机化合物命名实体的机器翻译系统的性能。

关键词：专利文献；机器翻译；有机物识别

中图分类号：TP391　　**文献标识码**：A

An Approach for Recognizing and Translating Organic Compound in Patent Machine Translation

Zhijun Ren[1,2], Wei Zhang[1], Hongfei Jiang[1], Jin Li[1], Kai Zhang[1]

(1.The China Patent Information Center, the State Intellectual Property Office, Beijing 100088; 2.Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: There exist numerous principles in naming organic compound system, which has the characteristic of containing complicated contents and diversified forms of expression. Thus the identification of the organic compound greatly influences the quality of translation in patent sentences which contains named entities of organic compound. Considering the features and formational rules of the named entity of organics in patent document, this paper classifies them into different categories and carries out the corresponding rules to recognize and translate them. Consequently, a patent document machine translation system based on organic identification and translation is implemented. The experiment result shows that the tool has a high identification rate and translation accuracy rate of organic compounds, thus improving the performance of the machine translation system containing named entities of organic compounds.

Key Words: patent document; machine translation; organic compound identification

1 Introduction

Patent document is an important type of technical text in human society. In recent years, as the country strongly advocates for patent applications and a nationwide consciousness of patent protection gradually approaches its maturity, China has witnessed an increasing amount of patent applications. According to statistics, China's patent applications reached 2,051,000 in 2012, increasing 26% compared with the previous year; patent licensing amounted to 1,255,000, growing by 31% compared with last year. Faced with such a huge amount of patent document, China faces a critical task of quickly and effectively spreading internationalized patent around the world. To facilitate the access for foreign users whose native language is not Chinese to retrieve, China is required to translate the patents. Therefore, the State Intellectual Property Office of the People's Republic of China began offering online C-E machine translation service of patent document in 2008.

Patent document involves the most updated technology, methods, devices, etc. According to Item Three in Part II, Chapter X in the Guidelines for Examination, "for full disclosure of chemical invention—invention of the compound, the specification shall indicate the name and structure of the chemical compounds of formula or molecular formula"; Item Eight, "for unity of chemical invention—unity of Markush claims", requires that in patent applications involving organic matter, especially PCT patents, a large part of the description of organic names should be included in its

---

[1]**收稿日期**：　　**定稿日期**：

基金项目：中国博士后科学基金资助项目（2013M530026, 2013M540125）；国家高技术研究发展计划(863 计划)（2012AA011104）

作者简介：任智军（1977——），男，副研究员，机器学习，机器翻译；张威（1981--），男，助理研究员，主要研究方向为自然语言处理、机器翻译、专利信息处理；蒋宏飞（1982——），男，助理研究员，主要研究方向为自然语言处理、机器翻译、专利信息处理；李进（1982——），男，助理研究员，主要研究方向为机器翻译、专利文献研究；张凯（1988——），女，实习研究员，主要研究方向为语料库建设，计算机辅助翻译。

claims and specifications. There exist numerous principles in naming organic compound system, which has the

characteristic of containing complicated contents and the diversity of forms of expression. Take No. CN201180056598 of PCT Application as an example, in its Claim 12, Claim 17 and embodiments, a large number of long terms of organic compound like "4-[2-[4-[(1R,3S)-3-[[(1R)-1-(4-氟-3-甲氧基-苯基)乙基]氨基]-环戊基]-苯氧基]乙酰基]哌嗪-2-酮(化合物 101)" are concluded. Therefore, the identification of the organics greatly influences the quality of translation in patent sentences which contains organic named entities.

This paper aims to research on the identification and translation of named organic entity of machine translation in patent document, and further proposes a method of constructing rules for the identification and translation of named organic entities. Then the methods of the identification and replacement of organic compound during machine translation are discussed. Specific studies include the following aspects: 1. Study and implementation of rule-based organic named entity identification method; 2. study and implementation of rule-based translation methods concerning organic named entity; 3. systematic implementation methods based on identification and translation of organics in patent document machine translation. Experiments suggest that this method has achieved better identification and translation accuracy, and at the same time improves the quality of the translation of sentences containing organic matter.

2 Rules of Identification of Organic Named Entities

Organics refer to the carbonaceous compounds formed by oxygen, hydrogen and carbon elements, but do not include carbon monoxide, carbon dioxide and carbonate materials. In addition to the basic elements of carbon, hydrogen and oxygen, the organic substance may also contain nitrogen, phosphorus, sulfur, halogen and other elements. There exist numerous organic compounds, up to tens of millions due to that the binding of carbon atoms of organic compounds is very strong, and that they can be combined with each other into a carbon chain or a carbon ring.

The number of carbon atoms can be one or two, while it can also be thousands, tens of thousands, or even many organic polymers can have hundreds of thousands of carbon atoms. In addition, isomerism is very common in organic compounds, which partly explains the large quantity of organic compounds. Organics are mainly named according to system, custom and popular name and so on. By different naming methods, multiple names of the same organic matter can be achieved. Moreover, for those organic compounds with complex structure, even under the same naming method they may be given a different name. Thus identification of organic names is difficult. In order to facilitate the identification of organic entity, the fragments of organic named entities are classified into the following seven types:

WA[A] group
WA[B] ring azobenzene structure
WA[C] hydrogenation
WA[D] chemical elements
WA[E] between beside opposite signed position
WA[F] special condition
WA[O] else

And rules of identification are made according to the six types respectively. The strategies of making such rules are as follows:

```
(-1) {WA[A]}+(0) {CHN[,] }+(1) {WA[A]} =>TREE(-1,1)+PUT(fp,WA,B)
(-1) {WA[A]}+(0) {CHN[,] }+(1) {WA[A]} =>TREE(-1,1)+PUT(fp,WA,B)
```

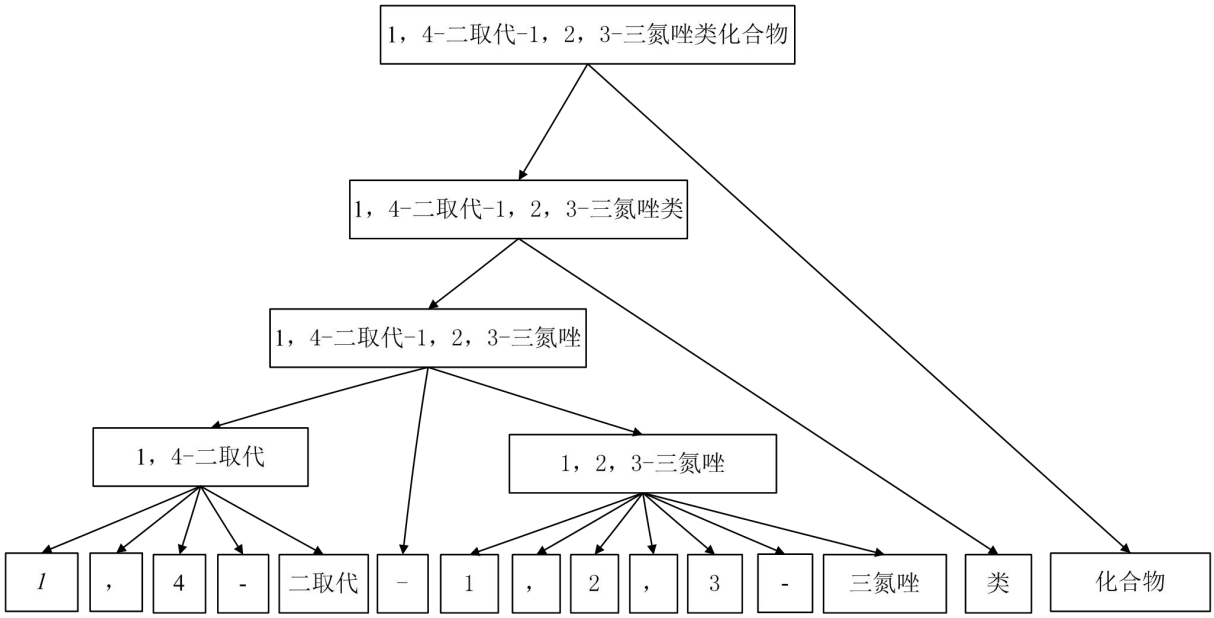
Use the above semantic attributes and identification rules to define the identification process and al gorithm. Specific procedures are as follows:

Input: Chinese sentences
Output: Compound recognization result
Identification process:
(1) Segmentation of Chinese words, annotation of compound attributes.

(2) Analyzing word by word to see if the sentence activities regulation. If so, the sentence will turn to Step 2, otherwise it will turn to Step 5.
(3) Identification according to the regulation. If the sentence complies with the regulation, then co mbine the word, attributing the new word as a compound.
(4) When it comes to the final word, no new compound is combined, and then it will turn to Step 5. Otherwise it will turn to Step 2.
(5) Identification of adjunct word.
(6) Output of the result.

The process of employing the identification rules to identify organic named entities like“1，4-二取代-1，2，3-三氮唑类化合物”is shown in the following figure:

Figure 1: Identification of Organic Named entities Using Rules



3 Rules of Translating Organic Named Entities

Relying on organic syntax tree, root traversal method is adopted in translation; for special terms found in the string affixes, a combined and bottom-up approach is adopted during translation; carrying out hierarchical method, translation convention or common standards are employed to decompose and translate the indentified word strings; translating according to the template, and re-arrange the order of the translation results.

Table 1: Glossary of Basic Terms

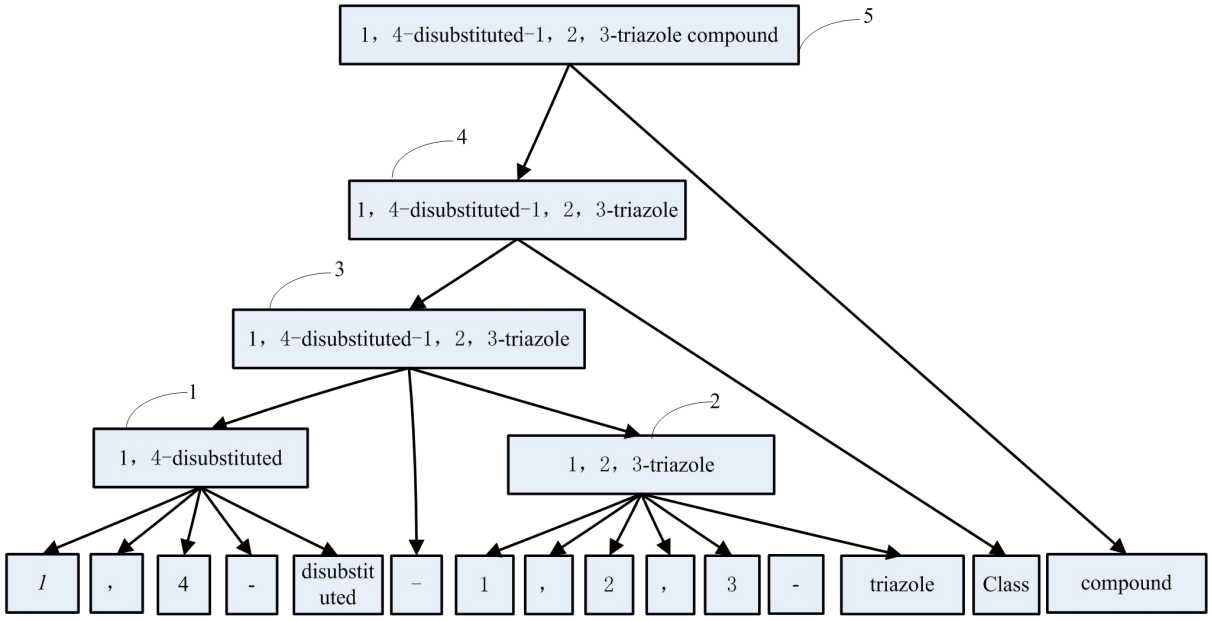
Chinese	English	Chinese	English
烷烃	Alkanes	酸酐	Anhydride
羧酸	Acid	卤烃	Alkylhalide
烯烃	Alkenes	酰胺	Amide
酯	Ester	醇	Alcohol
炔烃	Alkynes	胺	Amine
酰卤	Carbonyl halide	酚	Phenol
芳烃	Aromaties	磺酸醚	Ether
酸酐	Anhydride	醛酮	Carbonyl compounds
卤烃	Alkylhalide	三氮唑	triazole

Rules of Translation：

```
(0) {NUM[A]}+(1) {CHN[,]}+(2){NUM[B]}+(3){WA[F]}=> NUM(0)+STR(,)+NUM(2)+EN(3)
(0) {NUM[A]}+(1) {CHN[,]}+(2) {NUM[B]} +(3){CHN[,] }+(4) {NUM[B]} +(5) {CHN[-]}+(6)
{WA[B]}=> NUM(0)+STR(,)+NUM(2)+STR(,)+NUM(4)+STR(-)+EN(6)
(0) {OC[T]}+(1) {CHN[-]}+(2) {OC[T]}=> OC_STR(0)+STR(-)+OC_STR(2)
```

The process of translation is shown in the following figure. Numbers in the figure refer to the order of translation.

Figure 2: The process of translation



4 Implementation of Patent Document Translation System Based on Organic Identification and Translation

Based on the previous identification and translation rules, we have implemented a rule-based Chinese-English identification and translation system of organics. The system has four basic modules including the recognition module, the translation module, the replacement module and the output module. Segmented Chinese texts, or tokenized English texts, can be accepted by the system as the input. Sentences in the texts are taken as units for procession. The organic identification module extracts expressions containing digital information based on the identification module, then tags them according to the organic identification. The translation module carries out the translation after classifying the sentences according to the identification results of the identification module. The output module can provide three lists of outputs including the identification output, the translation output and the new-word output on the basis of identification and translation. Framework of the system is shown in Figure 3:

Figure 3: Framework of the system

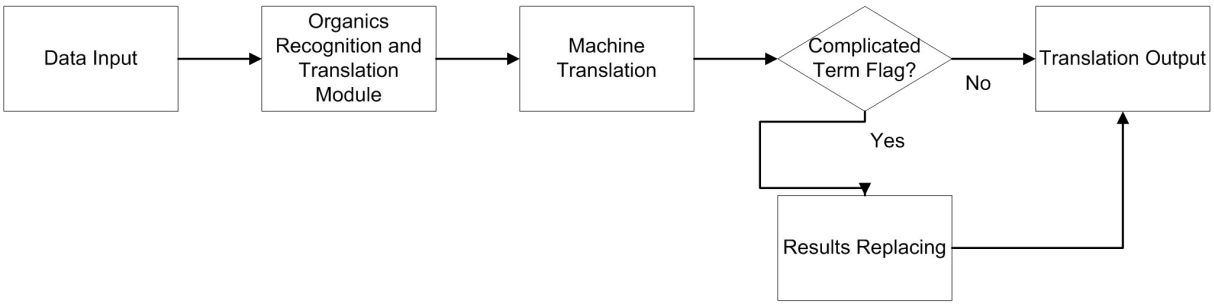
For example, in the following sentence：*“本发明涉及1，4-二取代-1，2，3-三氮唑类化合物的制备及其应用，具体地，提供了1，4-二取代-1，2，3-三氮唑类化合物，具有下列通式(I)所示的结构.*”

According to the identification rule, the organic compound is identified to be: “1，4-二取代-1，2，3-三氮唑类化合物”. Based on the translation rule, “1，4-二取代-1，2，3-三氮唑类化合物” is translated as “1,4-disubstituted-1,2,3-triazole compound”， and the original sentence is replaced by *“本发明涉及 CNNP1 的制备及其应用，具体地，提供了 CNNP1，具有下列通式(I)所示的结构.*”. The following is the final output:

```
<sentence modify=”true”>
<content>本发明涉及CNNP1的制备及其应用，具体地，提供了CNNP1，具有下列通式(I)所示的结构.</content>
<word source=”1，4-二取代-1，2，3-三氮唑类化合物” target=”1,4-disubstituted-1,2,3-triazole compound ” symbol= “CNNP1” />
</sentence>
```

The purpose of the development of translation system lies in that after the complex named entities of organic compounds being examined by the recognition and translation module, the sentences are handed to the machine for translation and the special tags are replaced by the terms in the translation, in order to improve the follow-up accuracy of machine translation. Framework of the system is shown in Figure 4:

Figure 4: Framework of the system



For example, the original sentences, the translation results and the replacement results are compared as below:

Table 2: Input sentences and translation results

Items	***Contents***
Original Sentence	*本发明涉及1，4-二取代-1，2，3-三氮唑类化合物的制备及其应用，具体地，提供了1，4-二取代-1，2，3-三氮唑类化合物，具有下列通式(I)所示的结构.*
Machine Translation Result	It is disubstituted - 1,2 to the invention relates to 1,4-, the preparation and application of 3- triazole class compound, and specifically, it is disubstituted - 1,2 to provide 1,4-, and 3- triazole class compound has the shown structure of following general formula (I)：
Original Sentence	*本发明涉及CNNP1的制备及其应用，具体地，提供了CNNP1，具有下列通式(I)所示的结构.*
Machine Translation Result	The invention relates to preparation and application of *CNNP1*, specifically, provide *CNNP1*, have the shown structure of following general formula (I)：
Translation Result after Replacement	The invention relates to preparation and application of 1,4-disubstituted-1,2,3-triazole compound, specifically, provide 1,4-disubstituted-1,2,3-triazole compound, have the shown structure of following general formula (I)：

Compared with the original translation, the translation has improved after adopting the method of organic replacement.

5 Experimental Results and Analysis

To decide whether the approach is effective and to what extent it can improve the machine translation result, 90 sentences containing organic compound name or names are collected randomly from real patent texts written in Chinese and form a test set. The evaluating strategy is as follows:

(1) Each sentence is processed by two types of machine translation engines, namely an engine with the identification and translation module (*result 1*) and an engine without the identification and translation module (*result 2*).

(2) Each sentence is evaluated and scored by two operators independently, and the average score is used as the final score.

(3) *Result 1* and *result 2* are compared from two aspects, namely the word level and the sentence level and a score is given for each aspect. If *result 1* shows improvement compared with *result 2* on the word level, the sentence will be scored as "+1" on the word level. If *result 1* shows degeneration compared with *result 2* on the word level, the sentence will be scored as "-1" on the word level. And if *result 1* performs the same as *result 2* on the word level, the sentence will be scored as "0" on the word level. The same scoring standard applies to the sentence level as well.

The evaluation result is listed as follows.

Table 3: Experimental evaluation score

	improvement	same	degeneration
Word Level	33 / 90	50 / 90	7 / 90
Sentence Level	43 / 90	43 / 90	4 / 90

It can be seen from the above table that 33.67% (33/90) translation result is improved on the word level by the proposed approach, which is a relatively high rate. And a significant improvement is shown on the sentence level, which is 47.78% (43/90).

6 Conclusion

A rule-based identification and translation system of organic named entities is introduced. Experimental results show that the system can accurately identify and translate organic named entities with the advantages of simplicity and efficiency, which is of great significance for the improvement of performance of machine translation, text information extraction and other natural language processing tasks.

In the next step, we will adopt statistical methods for identification of organic compounds as well as start simultaneously the identification and translation of other complex terminologies.

Acknowledgments

This work was supported by the Hi-Tech Re-search and Development Program of China (2012AA011104), and Postdoctoral Science Foundation of China (Grant No.: 2013M540125, 2013M530026).

References

[1] Bin Sun. 2003. A Summarization of Information Extraction (2) (In Chinese). Terminology Standardization & Information Technology, (1).

[2] Jun Zhao. 2009. A Survey on Named Entity Recognition, Disambiguation and Cross-Lingual Coreference Resolution (In Chinese). Journal of Chinese Information Processing, 23(2).

[3] YuFeng Chen, ChengQing Zong. 2008. A Structure-based Model for Chinese Organization Name Translation. ACM Transactions on Asian Language Information Processing, 7(1): 1-30.

[4] Wenguang Zhang, Zuhao Wang. 2006. English Nomenclature for Organic Compound (In Chinese). Chemical Education, 2006(11).

[5] Nan Li, Rongting Zheng, Jiuming Ji, Qingqing Teng. 2010. Research on Chinese Chemical Name Recognition Based on Heuristic Rules (In Chinese). New Technology of Library and Information Service, 29(9).

[6] Xuebing Wang. 2005. Study on Identifying and Analysis Method of CA Chemical Substance Index Name (In Chinese). Journal of East China Normal University, 2005.

[7] Dan Song, Jiqing Sun. 2009. Automatic Index Chemical Feature Words Based on Rules. Journal of the China Society for Scientific and Technical Information (In Chinese). 2009(5).

[8] Fengsong Xiao. 1997. On Organic Nomenclature (In Chinese). Journal of Huaibei Normal University (Nature Science Edition). 1997(2).

[9] Liang Muliang, Li Wei. 2002. The Identification of Vocabularies about Medicines and Chemicals in Chinese Commodity Text (In Chinese). Journal of Yantai University (Nature Science and Engineering Edition). 2002(4).

任智军 北京市海淀区北太平庄路 25 号 100088
手机：13488845635
电子邮箱: renzhijun@cnpat.com.cn