文章编号：1003-0077（2011）00-0000-00

汉英专利文献机器翻译中的一些关键问题*

蒋宏飞[1,2]，姜涛[2]，张凯[2]，任智军[2]，张威[2]

（1 北京师范大学中文信息处理研究所，北京 100875；2.国家知识产权局中国专利信息中心，北京 100088）

摘要：近年来，专利机器翻译受到了越来越多的关注与重视。和其他翻译任务不同，专利文献机器翻译任务有其特色所在。而汉英专利文献机器翻译具有广泛的实际需求，同时也存在需要特殊考虑的问题。本文针对汉英专利文献机器翻译中的特殊问题，进行了较为全面的总结，同时对一些可行的方法继续了探讨，旨在为相关的研究以及工程实践提供参考。

关键词：专利文献；机器翻译；汉英专利文献机器翻译

中图分类号：TP391　　**文献标识码：**A

Some Key Issues in Chinese-to-English Patent Machine Translation

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

(1. Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875; 2. The China Patent Information Center, the State Intellectual Property Office，Beijing 100088)

Abstract: Patent machine translation draws more and more attention in recent years. It differs from other machine translation tasks in many aspects. There are urgent demands for Chinese-to-English Patent Machine Translation (CEPMT). Meanwhile, many special issues should be considered during the development of a CEPMT system. This paper summarizes some key issues in CEPMT based on practical research and application experiences. And some effective strategies for these issues are discussed. These studies aim to supply useful and practical references for further researches and applications in this task.

Key words: patent; machine translation; Chinese-to-English Patent Machine Translation

1Introduction

A patent is a set of exclusive rights granted by a sovereign state to an inventor or their assignee for a limited period of time, in exchange for the public disclosure of the invention. Patent system is an effective way for intellectual property protection. Patents have played a positive role in technology progress and economic development all over the world.

Since the 1980s, China had established the intellectual property systems, of which the most important is the patent system. In 2008, the State Council of China issued the Outline of National Intellectual Property Strategy. From then on, the intellectual property related issues are becoming more and more important in the whole nation. The number of China’s domestic patent application increase sharply in recent years. And with the continuous development of China's science and technology, the domestic inventions become more and more concerned by the international stakeholders. However, the domestic patent applications of China are usually written in Chinese. Thus, they are not directly usable for the foreign users. Translating the Chinese patent document

* **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：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；姜涛（1980——），男，实习研究员，主要研究方向为机器翻译；张凯（1988——），女，实习研究员，语料库建设，计算机辅助翻译；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。

into foreign language, for example English, is a necessary process in many international affairs. Since the amount of the Chinese patent documents is very huge, translating only by human is not feasible. Thus, patent machine translation systems are considered as potential solutions in many cases.

In recent years, the patent machine translation receives more and more attentions from the intellectual property (IP) organizations and machine translation research communities. From 2008, the five biggest IP offices worldwide, namely the United States Patent and Trademark Office (USPTO), European Patent Office (EPO), State Intellectual Property Office of the P.R.C (SIPO), Korean Intellectual Property Office (KIPO) and Japan Patent Office (JPO), have started a cooperation project Mutual Machine Translation with the aim to improve the quality of the machine translation services and to facilitate the free flow and exchange of prior art documents, search and examination information. The State Intellectual Property Office of China (SIPO) had launched the Chinese-to-English patent machine translation service since 2008 [1]. And from then on, unremitting efforts have been made to enhance this service [2-4]. From NTCIR-9, the Chinese-to-English subtask has been added into the Patent Machine Translation task (PatentMT) [5]. And the PatentMT task of NTCIR-10 received the submissions from 21 groups and among them 12 groups participated in the Chinese-to-English subtask [6]. And the 5th Workshop on Patent Translation had been hold co-located with MT Summit XIV recently.

Su and Chang give a comprehensive and in-depth summarization of designing a machine translation system theoretically and practically [7]. Liu and Yu discuss the key difficulties of Chinese-to-English in general domain [8]. Inspired by them, in this paper, we will first summarize the characteristics of the Chinese patent and then discuss the special issues in Chinese-to-English patent machine translation.

2 The features of Chinese patents

In broad terms, *patent documents* may include the claim, the specification, the office actions and other documents related to an invention.

2. 1 The amount of patent is huge and continue to increase

Patent is one of the largest information sources in the worldwide. As reported in [9], the number of accessible patent documents is greater than 70 million all over the world. And according to the statistics from World Intellectual Property Organization (WIPO), 90%~95% of the inventions are recorded in patents each year in the world.

Patent documents are published in the form of serial reports. Since 2000, the numbers of patent documents published in each country, region or organization are increasing. After 2006, the annual publication number exceeds 3 million in the world. In recent years, with the further implementation of the National Intellectual Property Strategy, the amount of patent applications in China is increasing sharply year by year. In 2011, the total amount of the patent applications is more than 1 million for the first time. And this number is over 1.9 million in 2012.

2. 2 Wide coverage, involving all areas of practical technology

An invention is usually a technical innovation, so it always corresponds to some technology domains. Patent documents cover the vast majority of the technology domains and relate to every aspects of human society. International Patent Classification (IPC) is one of the popular

classification taxonomy for patent in the world wide. Table 1 shows the first level of the IPC system, namely the eight *Sections*.

Table 1: The Sections of the IPC taxonomy

Section	General descriptions
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G	Physics
H	Electricity

Each Section is further divided into different subsections and subclasses etc. There are more than 70 thousands entries in the IPC system according to the IPC Advanced version of 2013.01. And the entry number is still growing along with the emergence of new industries.

2. 3 The contents are detailed, including all kinds of information

Patent documents include the comprehensive descriptions of technical information (such as the *specification*), the related legal information (such as *claims*) and so on. The specialists in the same domain as the said invention involving are expected to be able to carry out the exactly same engineering implementation following the description in the patent. And the stakeholders are able to know the complete legal status and the protected rights by reading the related patent documents. Moreover, the inventor or applicant can realize the deficiencies or the defects contained in their original application from the feedback made by the investigators (e.g. the *Office Actions*). Thus, the contents of the patent are usually detailed, including all kinds of information.

2. 4 Written in consistent and fixed for-mats

The drafting, examination and publication of a patent are generally carried out in accordance with relevant regulations and standards. Thus, the patent documents are always written in consistent and fixed formats.

For example, the specification of an invention comprise the title of invention, the technical field, the background art, the brief summary, the descriptions of drawings and specific embodiments and so on. And each part has to follow some specific written requests and all the parts are arranged in a fixed order. Based on these features, we can design customized translation solutions for different parts.

3 The key issues in Chinese patent analysis and processing

The work of [8] pointed out three kinds of difficulties in Chinese analysis and processing: 1) the Chinese lexical analysis, 2) the Chinese sentence parsing and 3) the layer of Chinese grammar.

Besides them, there are more special issues that should be paid attention to in Chinese patent analysis and processing.

3．1 The long sentence length

The first one is the long sentence length is-sue. Generally speaking, the average length of a Chinese sentence in spoken language is around 10 [10]. And as reported in [11], the average word number is 20 for the sentences in the corpus of *The People's Daily* in the first half year of 1998. Contrastively, the sentences in Chinese patent are much longer. According to our study based on more than 40,000 passages of patents in IPC E section, the average word number of a patent sentence is greater than 50 (around 90 Chinese characters without word segmentation). From this comparison, we can see that the sentence in Chinese patent is much longer than the ones in general domains.

As we know, the sentence length is directly related with the complexity of the processing / analysis algorithms. Most of them are beyond the linear relation. Thus, the longer the sentence is the more time and space will be cost. Indeed, many natural language processing models will become infeasible due to combination explosion if the sentence length beyond a certain threshold. For instance, the GIZA++ [12], one frequent used word alignment tool, restricts the word number of input sentences less than **100**.

Facing to the long sentence issue in Chinese patent, the sentence splitting should be taken. However, the sentence splitting is not a simple task. How to decide the appropriate splitting points is worth for in-depth studies [13-14]. Specifically, for translation task, the sentence splitting should affect the bilingual transformation as little as possible. One possible solution will be the intelligent sentence splitting taking the bilingual alignment relation into considerations.

3．2 Unknown new technical term

The second one is the new technical term detection issue. A patent document is always related with the technology innovation. Thus, the new technical terms are frequently occurs in patent documents.

In the analysis and processing phases, such as word segmentation, POS-tagging and parsing, the unknown new term may bring in serious errors. For example in table 2, in the Chinese sentence (1), a new term "*纵向有肋扭合联轴器*" exists.

This new term tends to be split as several isolate words(*纵向/f 有肋/ad 扭合/v 联轴器/n*), and the Part-of-Speech of some isolate words tends be assigned incorrectly. Furthermore, the parsing is likely to be damaged due to the incorrect verb assignment (e.g. *扭合/v* in (3) of table 2).

Table 2: Possible wrong processing results due to the unknown new term.

(1)Chinese patent sentence*:本发明涉及机械通用零部件领域，具体涉及一种纵向有肋扭合联轴器。*

(2)Correct word segmentation and POS-tagging: *本发明/n 涉及/v 机械通用零部件/nz 领域/n ，/w 具体/d 涉及 /v 一种/q* ***纵向有肋扭合联轴器****/nz*

(3)Possible wrong word segmentation and POS-tagging: *本发明/n 涉及/v 机械通用零部件/nz 领域/n ，/w 具体/d 涉及 /v 一种/q /****纵向/f 有肋/ad 扭合/v 联轴器****/n*

To address these problems, the patent machine translation service providers have to spend lots of manpower to collect and process the new terms persistently. Many works have been done to automatically detect the new terms [15-17]. The more effective and practical solution is the

combining of the automatic new term detection and human verification.

3. 3 Distinctive expressing characteristics

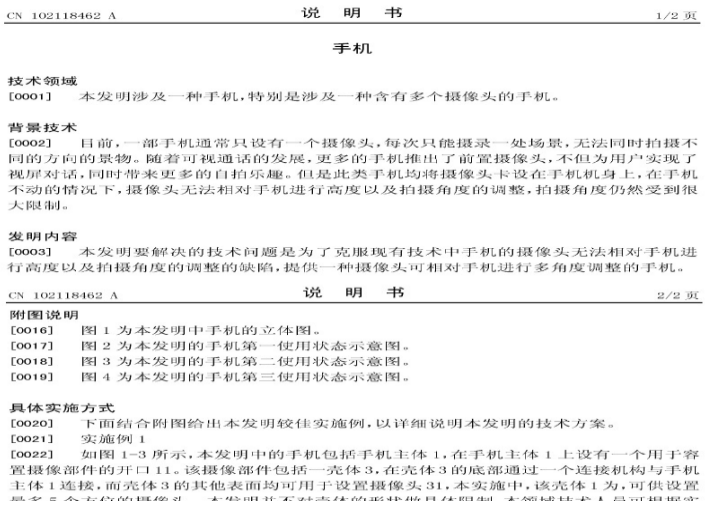
Different kinds of patent documents play different roles. For instance, the claims define, in technical terms, the extent of the protection conferred by a patent, or the protection sought in a patent application. The claims are of the utmost importance both during prosecution and litigation. Thus the claims usually follow the fixed expression formats. Each sentence in claims stands for a claim item. Some claims are independent claims which stand on their own while other are dependent claim which depend on a single claim or on several claims and generally express particular embodiments as fall-back positions. In table 3, Sentence (a) illustrates an independent claim with its English translation; Sentence (b) illustrates a dependent claim with its English translation.

Table 3: The examples of claims.

(a)
1.*一种电极组合物，包含：活性物质，所述活性物质包含锡、钴和碳的合金*
1. *An electrode composition comprising: an active material comprising an alloy of tin, cobalt, and carbon*
(b)
2.*根据权利要求 1 所述的组合物，还包含粘合剂，所述粘合剂包含聚丙烯酸锂。*
2. *The composition of claim 1, further comprising a binder comprising lithium polyacrylate.*

In contrast, the specification, which is also called the disclosure, is a written description of an invention. The patent specification is drafted both to satisfy the written requirements for patentability, as well as to define the scope of the claims. Moreover, each specification can be further divided into several parts: title of invention, technical field, background art, disclosure, description of drawings, and mode for invention. Each part is with distinctive expression characteristics. Figure 1 shows a snapshoot of a patent specification.

Each kind of patent document has distinctive expression characteristics. Even each part in one document has different expression style. In patent machine translation system designing, we can customize specified models or resources for each part or document.



CN 102118462 A 说 明 书 1/2 页

手机

技术领域

[0001] 本发明涉及一种手机，特别是涉及一种含有多个摄像头的手机。

背景技术

[0002] 目前，一部手机通常只设有一个摄像头，每次只能摄录一处场景，无法同时拍摄不同的方向的景物。随着可视通话的发展，更多的手机推出了前置摄像头，不但为用户实现了视屏对话，同时带来更多的自拍乐趣。但是此类手机均将摄像头卡设在手机机身上，在手机不动的情况下，摄像头无法相对手机进行高度以及拍摄角度的调整，拍摄角度仍然受到很大限制。

发明内容

[0003] 本发明要解决的技术问题是为了克服现有技术中手机的摄像头无法相对手机进行高度以及拍摄角度的调整的缺陷，提供一种摄像头可相对手机进行多角度调整的手机。

CN 102118462 A 说 明 书 2/2 页

附图说明

[0016] 图 1 为本发明中手机的立体图。

[0017] 图 2 为本发明的手机第一使用状态示意图。

[0018] 图 3 为本发明的手机第二使用状态示意图。

[0019] 图 4 为本发明的手机第三使用状态示意图。

具体实施方式

[0020] 下面结合附图给出本发明较佳实施例，以详细说明本发明的技术方案。

[0021] 实施例 1

[0022] 如图 1-3 所示，本发明中的手机包括手机主体 1，在手机主体 1 上设有一个用于容置摄像部件的开口 11。该摄像部件包括一壳体 3，在壳体 3 的底部通过一个连接机构与手机主体 1 连接，而壳体 3 的其他表面均可用于设置摄像头 31，本实施中，该壳体 1 为，可供设置

Figure 1: The snapshoot of a patent specification

Based on the above described features of Chinese patent, we know that corresponding considerations should be taken when we implement a Chinese-to-English machine translation service. In the following section, we will discuss the key issues in transformation from Chinese patent to English.

4 The key issues in transform from Chinese patent to English

Five kinds of difficulties are summarized for the transform from Chinese to English in [8]:

1) The decision of singular or plural form for English noun,
2) The decision of the tension of English verbs,
3) The processing of the subjunctive mood,
4) The asymmetry between the Chinese sentence structure and English sentence structure,
5) And the use of articles in English.

All of the five difficulties are very common and critical in Chinese-to-English patent translation. Moreover, there are some special key issues are involved in this task.

4. 1 New term translation

As we discussed in Section 3.2, the new un-known term recognition is a key issue in patent document analysis. Moreover, for the translation task, the system does not only need to recognize the new terms but also need to translate them into target language appropriately. The latter may be harder. In practice, two ways can be attempted. For the first, we can do the recognition and translation in a joint manner. The work of [18] is an inspiring reference. And for the second, some rule-based translation can be carefully designed for some special kinds of terms, e.g. the complex chemical compound.

4. 2 Term ambiguity in different domains

As described in previous sections, the patent documents are usually specified in different technology domains. One key issue in Chinese-to-English patent translation is the term ambiguity in different domains. For example, as shown in Table 4, a Chinese term "*压降*(ya jiang)" can be translated to different meanings in different domains.

Table 4: The examples of different term meanings in different domains.

In electricity domain:
[Chinese sentence]: 本实用新型能够降低回路压降
[English translation]: This utility model can reduce the **voltage drop** of the circuit.
In mechanical domain:
[Chinese sentence]: 该装置利用涡流产生压降.
[English translation]: The device utilizes the vortex to produce **pressure drop**.

One straightforward solution is to utilize do-main specified bilingual resources, e.g., the domain dictionaries. More elegant solutions should be able to model the domain information when making the translation hypothesis selection.

4. 3 Different expression style in different domains and different part

There are different expression styles in different domains. For example, one of the frequent descriptions in mechanical domain is the constitute relations between the parts while the ratios of each elements contributes the main focus of the specification in chemical domain.

Similarly, as discussed in Sections 3.3, there are different expression styles for different parts of a patent. For instance, claims and technical field are in distinct expression styles. The

developers of the patent machine translation systems are better to take this issue into consideration and design the domain-specific or document-category-specific models respectively.

5 Conclusions

In this paper, we summarized some key issues in developing a patent machine translation system. And some effective strategies for these issues are discussed. These studies aim to supply useful and practical references for further researches and applications in this task.

Reference

[1] Wang D. Chinese to English automatic patent machine translation at SIPO. World Patent Information, 2009, 31(2): 137-139.

[2] Jin Y. A hybrid-strategy method combining semantic analysis with rule-based MT for patent machine translation[C]. Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on. IEEE, 2010: 1-4.

[3] Hongfei Jiang. Current status of SIPO's online Chinese-to-English patent machine translation service. The presentation on East meet West 2011.

[4] Hongfei Jiang. Expanding the Applications of MT in Patent Translation. An Invited talk at 4th Workshop on Patent Translation. 2011.

[5] Goto, I., Lu, B., Chow, K. P., Sumita, E., & Tsou, B. K. 2011. Overview of the patent machine translation task at the ntcir-9 workshop. In Proceedings of NTCIR.

[6] Goto, I., Lu, B., Chow, K. P., Sumita, E., & Tsou, B. K. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In Proceedings of NTCIR.

[7] Su, Keh-Yih, and Jing-Shin Chang. "Some key issues in designing MT systems." Machine Translation 5.4 (1990): 265-300.

[8] Qun Liu and Shiwen Yu. Discussion on the Difficulties of Chinese-English Machine Translation. International Conference on Chinese Information Processing, 1998 (In Chinese)

[9] Jianrong Li. 2011. Patent information and its utilization. Intellectual Property Publishing House.

[10] Wei Cheng, Jun Zhao, Bo Xu and Feifan Liu. Bilingual Chunking for Chinese-English Spoken-language Translation. Journal of Chinese information processing, 2003:17(2) (In Chinese)

[11] ZhuangHua Li, Wanxiang Che and Ting Liu. A Study on Constituent to Dependency Conversion. Journal of Chinese information processing, 22(6). 2008 (In Chinese).

[12] Och, Franz Josef, and Hermann Ney. "Improved statistical alignment models." Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000.

[13] Heyan Huang and Zhaoxiong Chen. The Hybrid Strategy Processing Approach of Complex Long Sentence. Journal of Chinese information processing, 2001:16(3) (In Chinese).

[14] Xing Li and Chengqing Zong. A Hierarchical Parsing Approach with Punctuation. Journal of

Chinese information processing, 2005:20(4) (In Chinese).

[15] Chen F, Liu YQ, Wei C, Zhang YL, Zhang M, Ma SP. Open domain new word detection using condition random field method. Journal of Software, 2013,24(5):1051−1060 (in Chinese).

[16] Zheng YB, Liu ZY, Sun MS, Ru LY, Zhang Y. Incorporating user behaviors in new word detection. In: Proc. of the IJCAI 2009. San Francisco: Morgan Kaufmann Publishers, 2009. 2101−2106.

[17] Peng FC, Feng FF, McCallum A. Chinese segmentation and new word detection using conditional random fields. In: Proc. of the 20th Int'l Conf. on Computational Linguistics (COLING 2004). Stroudsburg: Association for Computational Linguistics, 2004.

[18] Yufeng Chen, Chengqing Zong and Keh-Yih Su. 2013. A Joint Model to Simultaneously Identify and Align Bilingual Named Entities. Computational Linguistics, 39(2):229-266

蒋宏飞 北京市海淀区北太平庄路 25 号 100088
手机：15801405738
电子邮箱: hf.jiang@gmail.com