Machine Translation between Uncommon Language Pairs via a Third Common Language: The Case of Patents

Benjamin K. Tsou and Bin Lu
Research Centre on Linguistics and Language Information Sciences,
The Hong Kong Institute of Education
10 Lo Ping Road, Tai Po, New Territories, Hong Kong
{btsou99, lubin2010}@gmail.com

Abstract

This paper proposes to familiarize the MT users with two major areas of development: (1) To improve translation quality between uncommon language pairs, the use of a third language as the pivot. Various techniques have been shown to be promising when parallel corpora for the uncommon language pairs are not readily available. They require the use of two other language pairs involving a common third language pairing with each member of the initial target pair. (2) The surging demands in the field of patent translation and for efforts to bootstrap machine translation in uncommon language pairs (e.g., Japanese and Chinese) via more common language pairs (e.g., Chinese-English and English-Japanese), and the application of the pivot approach to expedite processing.

1. Introduction

Recent success in the application of machine translation (MT) in many multilingual contexts, such as textual translation and cross-lingual information retrieval, has been dependent on the building up of critical bilingual resources such as Translation Memory (TM) to develop and continuously fine tune the translation or search engines. The basis for TM

is parallel sentence or sentence segment pairs drawn from relevant bilingual texts by sophisticated filtering processes. Thus there is the need to draw from quality texts the maximum amount of relevant parallel linguistic structures, which is of critical importance to the cultivation of high quality TM's.

The amount of useful bilingual corpora varies considerably between language pairs, and consequently the amount of useful TM's varies considerably among language pairs with English as a frequent common member in the paired TM's. For example, many parallel corpora with English as one language have been built, such as the French-English Canadian Hansards (Gale and Church, 1991), the Japanese-English parallel patent corpus (Utiyama and Isahara, 2007), the Chinese-English parallel patent corpus (Lu et al., 2010), and the Arabic-English and English-Chinese parallel corpora used in the NIST Open MT Evaluation[1].

However, few parallel corpora exist for language pairs among other languages (e.g. French-Chinese, German-Chinese, Japanese-Arabic or Chinese-Japanese). This is especially so for some domain-specific areas, such as patents, whose use of language embraces both legal and legalistic as well as technical considerations, thus placing limits on the useful application of current MT techniques to meet the needs in the commercial and other sectors.

This paper introduces two major areas of development to the MT users. First, we introduce various approaches with the use of a third common language as the pivot to improve translation quality between uncommon language pairs. When parallel corpora for the uncommon language pairs

[1] http://www.itl.nist.gov/iad/mig/tests/mt/

are not readily available, various techniques have been proposed and shown to be promising.

Second, we discuss the rapidly increasing demands in the field of patent translation and various efforts to bootstrap patent machine translation in uncommon language pairs (e.g., Japanese and Chinese) via more common language pairs (e.g., Chinese-English and English-Japanese), and the application of the pivot approach to expedite processing.

2. Pivoting Approaches for Machine Translation

There are three major approaches introduced below. Suppose the three languages involved are source, pivot and target.

The first is based on *phrase table translation* (Cohn and Lapata 2007; Wu and Wang, 2007). The approach usually first train two translation models: source-pivot and pivot-target; then it induces a new source-target phrase table by using the translation probabilities and lexical weights in source-pivot and pivot-target translation models For example, to translate between Chinese and Japanese, we can first train Chinese-English and English-Japanese translation models based on available bilingual corpora; the two models are then combined together at the phrase level to provide a new Chinese-Japanese phrase translation table with induced translated probability for each new entry.

The second approach is *sentence translation strategy* (Utiyama and Isahara, 2007; Khalilov et al., 2008). The first step of this approach is the same as the first approach: to train source-pivot and pivot-target translation models. But the second step would be quite different: the translation step of this approach would be to first translate the source sentence to the pivot

sentence based on the source-pivot translation model, and then translate the pivot sentence to the target sentence based on the pivot-target translation model.

The third uses existing models to build a *synthetic source-target corpus*, from which a source-target model can be trained (Bertoldi et al., 2008). For example, we can fist build a pivot-target translation model based on the pivot-target parallel corpus. Based on the pivot-target translation model, the pivot sentences in the original source-pivot bilingual corpus can be translated into the target language. We can then obtain a source-target corpus by translating the pivot sentences in the source-pivot corpus into the target language with the pivot-target translation model, and/or obtain a target-source corpus by translating the pivot sentences in the pivot-target corpus into the source language with the source-pivot translation model.

3. Status Quo on the Construction of Patent Parallel Corpora and the Use of Derived TM Resources

3.1 Patents in the Multilingual World

Patents are important indicators of innovation, and patenting increasingly becomes an international activity as the economy is globalized. More firms (especially the multinational ones) are investing large amounts of money on intellectual property (especially patents) to protect their own technologies, and filing patents in foreign countries.

Patent applications are increasing very quickly in recent years as illustrated by Figure 1. The application numbers filed in the leading patent offices from 1996 to 2009 are shown. We can observe that, in about 14 years, China's patent applications have increased by 10 times, and USA and

R. Korea have doubled in their patent applications, but USA and Japan are still the top two with most applications.

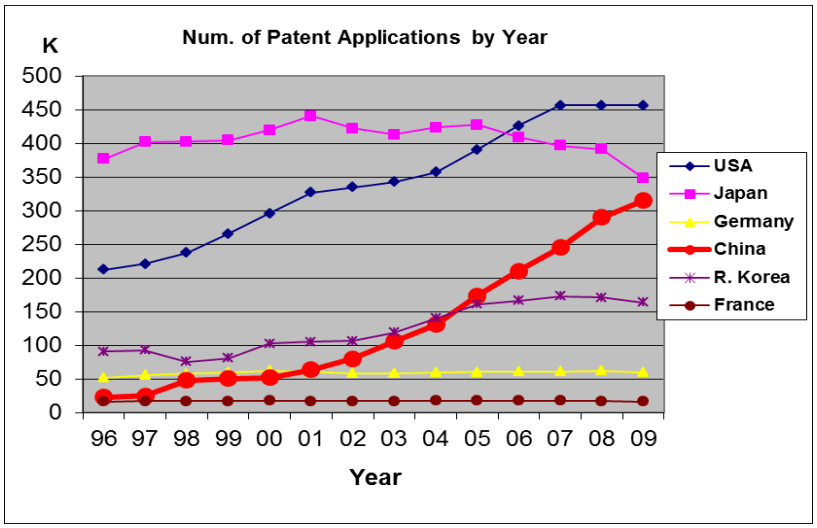

Figure 1. Applications by the leading patent offices[2]

This phenomenal growth in patents and the need to mediate between versions of patents in different languages have provided an important application arena for MT. The traditional practice for monitoring patents filed in foreign languages is usually involving translation companies to manually translate patents into a relevant language, which is slow, time-consuming, high-cost, and often quality-inconsistent.

The number of PCT (Patent Cooperation Treaty) international patent applications has rapidly increased in one decade to 1.8 million in June 2011. [3] A PCT international application may be filed in any language accepted by the relevant receiving office, but must be published in one of the official publication languages (Arabic, Chinese, English, French, German, Japanese, Korean, Russian and Spanish). Other highly used languages for filing include Italian, Dutch, Finnish, Swedish, etc. Table 1

[2] Retrieved March 2010, from http://www.wipo.int/ipstats/en/statistics/patents/csv/wipo_pat_appl_from_1883_list.csv
[3] Retrieved from http://www.wipo.int/pctdb/en/. The data below involving PCT patents comes from the website of WIPO.

shows the number of PCT applications for the most used languages of filing and publication. From Table 1, we can observe that English, Japanese and German are the top 3 languages in terms of PCT applications, and English accounts for over 50% of applications in terms of language of both publication and filing.

	Lang. of Filing	Share (%)	Lang. of Publication	Share (%)
English	895K	52.1	943K	54.9
Japanese	198K	11.5	196K	11.4
German	185K	10.8	184K	10.7
French	55K	3.2	55K	3.2
Korean	24K	1.4	3K[4]	0.2
Chinese	24K	1.4	24K	1.4
Other	336K	19.6	313K	18.2
Total	1.72M	100	1.72M	100

Table 1. PCT Application Numbers for Languages of Publication and Filing

Table 2 shows the total numbers of patent applications filed initially as PCT ones in different countries (column 2), and shows the total numbers of applications filed as PCT ones with English as the language of publication in different countries (column 3).

National Phase Country[5]	ALL	English as Lang. of Publication
Japan	424K	269K
China	307K	188K
Germany	32K	10K
R. Korea	236K	134K
China & Japan	189K	130K
China & R. Korea	154K	91K
Japan & R. Korea	158K	103K
China & Japan & R. Korea	106K	73K

Table 2. Distribution of Multilingual Patents

[4] Korean just became one of the official publication languages for the PCT system since 2009, and thus the number of PCT patents with Korean as language of publication is small.

[5] For the national phase of the PCT System, the statistics are based on data supplied to WIPO by national and regional patent Offices, received at WIPO often 6 months or more after the end of the year concerned, meaning that the numbers are not up-to-date .

Based on Table 2, we can see rough sizes of bilingual corpora, trilingual corpora, and even quadri-lingual corpora we can build from these PCT patents involving different languages.

3.2 Patent Parallel Corpora

A couple of parallel corpora have been introduced in the patent domain. For example, Utiyama and Isahara (2007b) mined about 2 million parallel sentences by using the description section of Japanese-English comparable patents. The corpus was used as the training data for the Japanese-English patent machine translation task at NTCIR PatentMT evaluations (Fujii et al., 2008; Goto et al., 2010).

For the parallel patent corpora involving the Chinese language, we have constructed a large-scale Chinese-English patent parallel corpus containing about 14 million good-quality parallel sentences mined from a large number of comparable patents (Lu et al., 2010). The human evaluation of sampled sentence pairs shows that the mined pairs are of high quality with only 1%-5% wrong pairs. We have chosen one million sentence pairs for the Patent MT evaluation at NTCIR-9 (Goto et al., 2010).

3.3 Patent Machine Translation

Although most MT systems in the patent domain begin with rule-based MT (RBMT) techniques, statistical machine translation (SMT) techniques are increasingly adopted and tremendous strides in SMT have been made in recent decades. However, SMT requires parallel corpora as critical resources, and the unavailability and limited size of parallel patent corpora, especially for uncommon language pairs, are still major limitations for SMT systems to achieve higher performance in the patent domain.

The major patent offices in the world such as Europe, China, Japan, Korea, as well as World Intellectual Property Organization (WIPO), have already employed machine translation-related services on their websites. This provides new opportunities for MT-related practitioners and researchers to examine and check the advances and development of MT technology in terms of new prospects to be explored.

There have been patent translation workshops during MT Summit X, XI, XII and the XIII (in December, 2011), as well as the patent MT competition at NTCIR-7, NTCIR-8 and the ongoing NTCIR-9 which has been an international forum for researchers and practitioners to compare and evaluate MT performance on the Japanese-English and Chinese-English patent translation.

The authors are working with NICT and NII in Japan to co-organizing the NTCIR-9 patent translation evaluation, to which more than 30 participants signed up from all over the world, and finally 130 runs from 21 teams were submitted. Some preliminary observations on the relative success of different approaches based on this large-scale evaluation are as follows (see also Goto et al. (2011)):

- On the Chinese-to-English patent translation task, the state-of-the-art SMT shows much better human evaluation scores (adequacy) than two commercial RBMT systems and the Google online translate system which do not have access to the training data provided by the task organizers.

- On the Japanese-to-English patent translation task, the commercial RBMT systems still show higher adequacies than the state-of-the-art SMT systems.

- On the English-to-Japanese translation task, some SMT systems achieve equal or better human evaluation scores (adequacy) than the top-level commercial RBMT systems. No SMT system did this at NTCIR-7, and this is thought to be the first time that this was achieved.

4. Building Parallel Patent Corpora using English as the Pivot

We have cultivated a trilingual parallel corpus by means of bilingual parallel corpora with English as the pivot (see also Lu et al., (2011)). With the 14 million Chinese-English bilingual sentences introduced in Section 3.2, and 4.2 million Japanese-English bilingual sentences we have in our center, a trilingual sentence-aligned patent corpus has been cultivated. Specifically, we align Chinese-English and English-Japanese sentence pairs by using the English sentences as the pivot, and finally obtain Chinese-English- Japanese sentence triplets. The selectivity in the whole process of this resource building is shown in Table 3.

	Number of Sentences / Sentence Pairs			
	RAW	Filter 1	Filter 2	Filter 3
CH-EN	56.1M(CH)	45.1M	31.5M	**14.3M**
JP-EN	25M(JP)	10.1M	5.5M	**4.2M**
CH-EN-JP	English as Pivot			**2.1M**
CH-JP	Combined Filtering			**1M**

Table 3. Selectivity in Resource Building

The pivoting approach has given us 2.1 million trilingual sentences. The distribution of these trilingual sentences is shown in Table 3. The preliminary manual evaluation of sampled sentences shows that about 70% of the trilingual sentences are correctly aligned. Because this accuracy is not satisfactory for the purpose as MT training data, it shows that the Chinese-Japanese sentence pairs obtained from the pivot approach contain noise because of the propagated errors in both Chinese-English alignment and English-Japanese alignment.

Thus, the trilingual sentence triplets are further filtered by using a Chinese-Japanese bilingual lexicon and capitalizing on some overlapping linguistic features. We finally arrive at about 1 million trilingual sentence triplets. The manual evaluation of 1,000 randomly sampled triplets show that about 93% of sentences are correctly aligned. It has been shown to significantly improve the quality of Chinese-Japanese sentence pairs, thereby opening opportunities to provide TM's for Chinese-Japanese MT, in spite of the scarcity of paired resources.

From the above experiment, we may conclude that the cultivation of large-scale parallel corpora from multilingual patents via a pivot language would alleviate the parallel data acquisition bottleneck in multilingual information processing involving a wide variety of languages, such as English, Chinese, Japanese, Korean, German, etc..

5. Conclusion

Given the 1.8 million PCT patent applications and their corresponding national ones, there is considerable potential to construct large-scale high-quality parallel corpora for a wide variety of languages, and to open

new opportunities for MT practitioners and researchers in the patent domain.

Moreover, with these large-scale patent parallel corpora, MT quality can be enhanced for uncommon language pairs (e.g. between East Asian languages and European languages other than English) by using the common language (e.g. English) as the pivot.

Acknowledgements

We wish to thank our colleagues, Mr. Chow K.P., Dr. Kataoka S. and Mr. Wong B. and others, for their help on this work.

References

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based Statistical Machine Translation with Pivot Languages. In *Proceedings of the International Workshop on Spoken Language Translation*, pp. 143-149.

Tevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 348–355.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the NTCIR-7 Workshop*. pp. 389-400. Tokyo, Japan.

William A. Gale, and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of ACL*. pp.79-85.

Isao Goto**,** Bin Lu**,** Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou**.** 2011. Overview of the Patent Translation Task at the NTCIR-9 Workshop. In *Proceedings of the NTCIR-9 Workshop.* (to appear)

Maxim Khalilov, Marta R. Costa-Juss`a, Carlos A. Henr´ıquez, Jos´e A.R. Fonollosa, Adolfo Hern´andez, Jos´e B. Mari˜no, Rafael E. Banchs, Chen Boxing, Min Zhang, Aiti Aw, and Haizhou Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. In *Proceedings of the International Workshop on Spoken Language Translation*, pp. 116–123.

Lu, Bin, Ka Po Chow and Benjamin K. Tsou. 2011a. The Cultivation of a Trilingual Chinese-English-Japanese Parallel Corpus from Comparable Patents. In *Proceedings of Machine Translation Summit XIII* (*MT Summit*). Xiamen.

Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu. 2010. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In *Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing* (*CLP-2010*), pp. 79-86. Beijing, China.

Benjamin K. Tsou and Bin Lu. 2011. Automotive patents from mainland China and Taiwan: A preliminary exploration of terminological differentiation and content convergence. *World Patent Information*. (under submission)

Masao Utiyama and Hitoshi Isahara. 2007a. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proceedings of human language technology: the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 484–491.

Masao Utiyama, and Hitoshi Isahara. 2007b. A Japanese-English patent parallel corpus. In *Proceeding of MT Summit XI*. pp. 475–482.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pp. 856–863.

Biography

Benjamin K. Tsou

Benjamin K. Tsou received his M.A. and Ph.D from Harvard and U.C. Berkeley respectively. He is the Chiang Chen Chair Professor of Language Sciences and Director of the Research Centre on Linguistics and Language Information Sciences at the Hong Kong Institute of Education, as well as Professor Emeritus of Language Sciences of the City University of Hong Kong. He is also a corresponding member of Académie Royale des Sciences d'Outre-Mer of Belgium and the founding president of Asian Federation of Natural Language Processing (AFNLP). He serves on the Standing Committee of the Executive Board of the Chinese Information Society of China, and on the editorial or advisory bodies of several journals and monograph series.

Tsou has cultivated the largest synchronous corpus of Chinese LIVAC (www.livac.org) on the basis of analysis of more than 400 million characters of Chinese newspaper texts from different Chinese communities in the last 16 years. His research interests have focused on the quantitative and qualitative studies of language to facilitate the curation and cultivation of large resources, including search engines and algorithms to enable meaningful mediation between parallel bilingual linguistic corpora

involving bilingual technical and legal texts in domains such as patents and judicial judgments. He has published widely in these areas.

Bin Lu

Bin Lu received an M.S. in Computer Science from Peking University in 2007. He is currently a PhD candidate at the Department of Chinese, Translation and Linguistics, City University of Hong Kong and a senior research assistant in the Research Centre on Linguistics and Language Information Sciences of the Hong Kong Institute of Education. He has published more than 20 academic papers in the area of Natural Language Processing and Computational Linguistics, and holds two China patents. His research focuses on Statistical Machine Translation (SMT), especially in the patent domain, as well as Sentiment Analysis and Opinion Mining.