Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT

Bin LU
Language Information Sciences Research Centre & Department of Chinese, Language and Linguistics, City University of Hong Kong
lubin2010@gmail.com

Tao JIANG
ChiLin Star Corp., Southern Software Park, Zhuhai, China & Northeastern University, Shenyang, China
jiangtaoster@gmail.com

Kapo CHOW
Language Information Sciences Research Centre, City University of Hong Kong
kapo.chow@gmail.com

Benjamin K. TSOU
Language Information Sciences Research Centre & Department of Chinese, Language and Linguistics,, City University of Hong Kong
rlbtsou@gmail.com

Abstract

The paper provides an account on the augmentation of a Chinese-English patent parallel corpus consisting of about 160K sentence pairs, which has been enlarged by about 45 times to more than 7 million sentence pairs mostly by the means of "harvesting" comparable patents from the Web. First, based on a large corpus of English-Chinese comparable patents, more than 22 million bilingual sentence pair candidates have been mined, of which we extract more than 7 million high-quality parallel sentences, which to our best knowledge is the largest parallel sentence corpus in the patent domain. Based on 1 million parallel sentences extracted from the *abstract* and *claims* sections, some interesting preliminary SMT results are also reported here. Last by not least, the method and approach proposed here should be applicable to other languages, which shows a novel way on how to reduce the data acquisition bottleneck in multilingual language processing.

1. Introduction

Parallel corpora are invaluable resources for NLP applications, including machine translation, multilingual lexicography, and cross-lingual information retrieval. Many parallel corpora have been available, such as the Canadian Hansards (Gale and Church, 1991), the Arabic-English and English-Chinese parallel corpora used in the NIST Open MT Evaluation[1] and Europarl corpus (Koehn, 2005). However, large parallel corpora are still too little.

To overcome this lack of parallel corpora, comparable corpora are also used to mine parallel sentences. For instance, Zhao and Vogel (2002) investigated the mining of parallel sentences for Web bilingual news collections which may contain much noise. Resnik and Smith (2003) introduced the STRAND system for mining parallel text on the web for low-density language pairs. Munteanu and Marcu (2005) presented a method for discovering parallel sentences in large Chinese, Arabic, and English comparable, non-parallel corpora based on a maximum entropy classifier. Wu and Fung (2005) exploited Inversion Transduction Grammar to retrieve truly parallel sentence translations from large collections of highly non-parallel documents.

However, less work has been done in the patent domain, and only the following two are found. The Japanese-English patent parallel corpus (Utiyama and Isahara, 2007) contains more than 2 million parallel sentences, and was provided for the NTCIR-7 patent machine translation task (Fujii et al., 2008). The English-Chinese patent corpus (Lu et al., 2009) contains about 160K parallel sentences which were extracted from more than 6,000 English-Chinese comparable/noisy parallel patents.

In this paper, we enlarge the Chinese-English parallel corpus (Lu et al., 2009) by over 40 times to more than 7 million sentence pairs by mostly harvesting a large corpus of English-Chinese comparable patents from the Web. Compared with the one in Lu et al. (2009), this corpus is not only much larger, but also may have different characteristics because these comparable patents were first filed with English as the original language, and then translated into Chinese and filed in China. On the other hand, the patents in Lu et al. (2009) were filed in the opposite direction (i.e. first Chinese, then English).

With the large number of comparable patents harvested from the Web, we mine parallel sentences based on two publicly available sentence aligners and simple heuristic rules. Currently, more than 22 million bilingual sentence pair candidates are found, of which we extract more than 7 million high-quality parallel sentences, which is the largest parallel sentence corpus in the patent domain to our best knowledge. Based on 1 million parallel sentences extracted from the *abstract* and *claims* sections, a small part of the whole parallel corpus, some preliminary SMT experiments are also reported here. Some sampled parallel sentences are available at http://livac.org/smt/parpat.html. Since patents cover many technical domains (e.g. chemistry, vehicle, electronics, biomedicine, etc.), the large parallel corpus could be a valuable resource for many cross-lingual information access applications not only in the patent domain but also in the related technical domains mentioned above. A rough estimation on the quantity of bilingual and multilingual patents including Chinese, Japanese, Korean, German and English is made. It shows considerable potential for easing the data acquisition bottleneck for these languages in multilingual

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

Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, pages 42–49
Malta, 22 May 2010

language processing.

In the next section we introduce related work, followed by the background in Section 3. Then the process of mining comparable English-Chinese patents from the Web is described in Section 4. The method of extracting parallel sentences from comparable patents and the SMT experiment are presented in Section 5, followed by discussion in Section 6, and we give conclusion and future work in Section 7.

2. Related work

Parallel sentences can be extracted from parallel corpora of documents or from comparable corpora. Since parallel corpora are bilingual text collections consisting of the same content in two or more different languages, it would be easier to find parallel sentences, and different approaches have been proposed: a) the sentence length in bilingual sentences (Brown et al. 1991; Gale and Church, 1991); b) lexical information in bilingual dictionaries (Ma, 2006); c) statistical translation model (Chen, 1993), or the composite of more than one approach (Simard and Plamondon, 1998; Moore, 2002). Comparable corpora raise further challenges for finding parallel sentences since the bilingual contents are not strictly parallel. Related work include Resnik and Smith (2003), Munteanu and Marcu (2005), Wu and Fung (2005), Zhao and Vogel (2002), etc.

For bilingual patent related work, Utiyama and Isahara (2007) used the "*Detailed Description of the Preferred Embodiments*" and "*Background of the Invention*" parts in the *description* section of Japanese-English comparable patents to find parallel sentences because they found these two parts have more literal translations than others. Lu et al. (2009) derives high-quality parallel sentences from English-Chinese comparable patents by aligning sentences and filtering sentence alignments with the combination of different quality measures, followed by the work in (Lu & Tsou, 2009).

The differences between this work with these two above lie in: 1) our comparable patents are mostly harvested from the Web and the parallel sentences mined are much larger compared to 2 million in the former and 160 K in the latter; 2) their comparable patents were both filed in USPTO in English by translating from the original language (namely, Japanese and Chinese) and identified by the priority information in the US patents. However, our comparable patents were first filed in English as a PCT patent, and later translated into Chinese. The different translation process may show different characteristics which will be explored in future.

For SMT, tremendous strides have been made in two decades. Brown et al. (1990; 1993) proposed the groundbreaking IBM approach, and the IBM models are word-based models. Later comes the SMT models called phrase-based models (Och and Ney, 2004; Koehn, 2004) in which translation unit may be any contiguous sequence of words. Phrase-based translation is implemented in the open-source Moses (Koehn et al., 2007), which is widely used in the SMT research community. We also use Moses for the SMT experiments in this paper. Currently, more researchers are taking advantages of syntax-based models (Chiang et al., 2005; Chiang, 2007), in which researchers attempt to incorporate syntax into phrase-based models.

For the evaluation of machine translation, NIST has been organizing MT open evaluations for several years, and the performance of the participants has been improved rapidly. The NTCIR-7 patent machine translation task (Fujii et al., 2008) has tested SMT performance on only the Japanese-English patent translation. Jiang et al. (2010) use Part-of-Speech model for the N-best list Reranking within the phrase-based SMT based on some parallel sentences extracted in this paper.

3. Background

A patent is a legal document representing "*an official document granting the exclusive right to make, use, and sell an invention for a limited period*" (Collins English Dictionary [2]). Patents are important indicators of innovation. As Sun (2003) stated "a*s the economy is globalized, patenting increasingly becomes an international activity*". More firms, especially the multinational ones, are investing more and more money on intellectual property (especially patents) to protect their own technologies, and filing patents in foreign countries. There have been many legal cases involving the claims of patent infringement, such as Nokia vs Apple, Cisco vs. Huawei, Intel vs AMD, and the DVD manufacturers in China vs. the dvd6c licensing group. The companies may be interested in monitoring and analyzing the patents filed in different languages, such as English, Chinese, Japanese, Germany, etc. The traditional practice for monitoring patents filed in foreign languages is usually to involve translation companies to manually translate patents into a relevant language, which is slow, time-consuming, high-cost, and often quality-inconsistent.

Meanwhile, patent applications are increasing very quickly, especially those filed in China (Sun, 2003). The patent application numbers filed in the top leading patent offices including Japan, USA, China and Germany from 1996 to 2008 are shown in Figure 1, from which we can observe that in about 12 years, China's patent applications have increased by 10 times while USA only doubles its patent applications. The increasing trend of patent applications also impose more workload for the manual translation which demands more advanced machine translation engines and more parallel data to help us handle this problem.

Each patent application consists of different sections, namely, *bibliographical data (including title*, *abstract)*, *drawings, claims*, *description*, etc. Since we focus on the text in the patent applications, only *title, abstract, claims*

[2] Retrieved March 18, 2010, from http://www.collinslanguage.com/

and description are used in the experiments discussed below. From the legal perspective, the *claims* section is the most important part in one patent application, because it defines the coverage that the applicant wants to claim. The *description* section gives the technical details of the patent involved, and the descriptions of some patents have further subdivisions, such as *Field of the Invention*, *Background of the Invention*, *Objects of the Invention*, *Summary of the Invention*, etc.

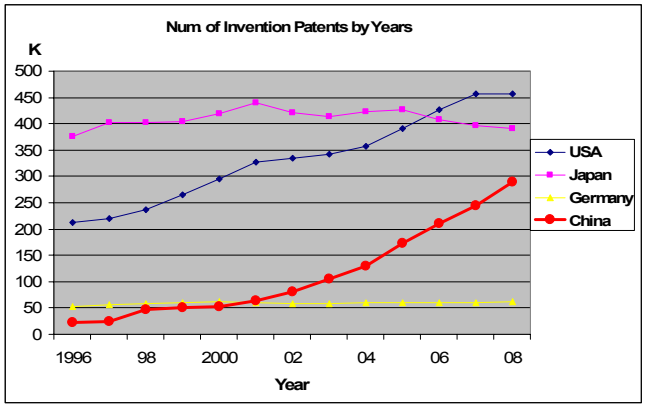

Figure 1: Patent applications by the top leading patent offices
Source[3]: WIPO: Patent Applications by Office

4. Mining comparable patents from the Web

The patents used in Lu et al. (2009) were first filed in China with Chinese as the original language, and we are also interested in patents which were first filed in English, and later filed in Chinese in China.

The intuition here is that from Figure 1 we can see that the number of patents filed in China was quite small in the 1990s compared to that in USA or Japan, and hence the possibility is lower for patents to be first filed in Chinese and then to be filed in English later. The opposite direction is quite different since western companies have accumulated a large amount of patents filed in other languages, and they may file Chinese patents to protect their inventions within China. Therefore, there may have many Chinese patents translated from English. The large amount of mined comparable patents which were first filed in English and later filed in Chinese prove our intuition.

3.1 Mining Chinese patents with English as original language

The official patent office in China is the State Intellectual Property Office (SIPO) of the People's Republic of China. SIPO was established in 1980 and began to accept patent applications since 1985. All Chinese patents are filed through SIPO. About 20 years after its creation, SIPO is regarded "*one of the more vibrant patent offices of the developing world, where an even-increasing number of domestic and non-resident applications are processed each year*" (Landry, 2008).

On the SIPO website[4], Chinese patents can be searched by many fields, such as *application number, publication number, title, International Patent Classification (IPC) code, inventor*, etc., including those patent applications which were originally filed in English with PCT publication numbers.

There were about 200 K Chinese patents both filed in China and previously filed as PCT applications in English up to early 2009. Most of the patents are invention patents. For these Chinese patents, the *bibliographical data, title, abstract* and *the major claim* were first crawled from the Web, and then *other claims* and *description* were also added. Since some contents are in the image format, the images were OCRed and manually verified. Inevitably there are errors in the data, but the quality can be generally acceptable.

3.2 Mining the corresponding English patents

All the PCT patent applications are filed through the World Intellectually Property Organization (WIPO). With the Chinese patents mentioned above, the corresponding English patents may be searched from the website of WIPO[5] to obtain relevant sections of the English PCT applications, including *bibliographical data, title, abstract*, *claims* and *description*. The mined English patents were automatically split into individual sections according to the respective tags inside patents.

However, not all but only about 40% out of the large number of Chinese patents had found their corresponding English ones. Some contents of the English patents were OCRed by WIPO, and hence there may be some errors in the English data.

3.3 Comparable patents mined

Here we give the percentage distribution of the Chinese patents in terms of their primary IPC codes. The IPC consists of 8 sections, ranging from A to H. From the category distribution in Table 1, we can see that *1) H: Electricity* and *C: Chemistry & Metallurgy* are the top two categories in terms of patent number, *2) D: Textiles & Paper* and *E: Fixed Construction* are the two categories with the smallest numbers of patents.

	A	B	**C**	D	E	F	G	**H**	Total
Percent (%)	16.6	11.9	**21.7**	1.7	1.7	4.7	18.0	**23.7**	100

Table 1. Percentage Distribution of Chinese Patents

Meanwhile, we obtain information on the area distribution of the patents, which shows that USA, Europe, Great Britain, Korea and Japan are the top leading areas in terms of the number of the patent priority. The distribution of publication years for the PCT patents filed in China are shown in Figure 2[6], which shows a big growth of the PCT patent applications filed in China in the 21st century.

[3] Retrieved March 20, 2010, from http://www.wipo.int/ipstats/en/statistics/patents/csv/wipo_pat_appl_from_1883_list.csv

[4] http://www.sipo.gov.cn/

[5] http://www.wipo.int/

[6] We only show the numbers within the period of 1996 to 2007, and skip the numbers for other years.

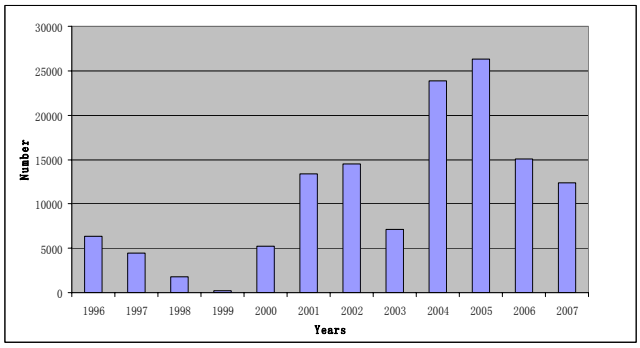

Figure 2. The distribution of publication years

The detailed statistics of each section for both Chinese and English patents are shown in Table 2.

Sections	Chinese		English	
	#Char	#Sent	#Word	#Sent
Title	1.3M	78K	0.8M	78K
Abstract	16M	274K	10M	392K
Claim	183M	3.4M	108M	3.7M
Description	1,233M	24.4M	677M	27.0M
Total	1,435M	28.1 M	795M	31.2M
Avg/Patent	18K	357	10K	394

Table 2. Data Statistics of Comparable Patents

Here we consider the English-Chinese patent pairs as comparable (or noisy parallel) patents because they are not parallel in the strict sense but still closely related in terms of information conveyed. As noted in Lu et al. (2009), loose translations are very common in English-Chinese comparable patents, and the major explanations are:

1) The field of intellectual property is highly regulated in different countries, and the translation may be highly influenced by the stylistic differences in the individual countries;

2) The patent applicants may intentionally change some technical terms or the patent structure to broaden the patent coverage or to avoid potential conflict with other patents in the country when a new version is filed in another language and country;

3) Sometimes, the characteristics of different languages make it difficult to keep the original terminology/structure, and the translator may render it in a target language-specific way.

5. Mining parallel sentences & SMT experiments

The comparable patents are first segmented into sentences according to punctuations, and the Chinese sentences are segmented into words. The sentences in all sections of Chinese patents are aligned with those in the corresponding sections of the corresponding English patents to find parallel sentences.

4.1 Aligning sentences in comparable patents

To find high-quality parallel sentences in comparable patents, we combine two publicly available sentence aligners, namely Champollion (Ma, 2006) and MS aligner (Microsoft Bilingual Sentence Aligner) (Moore, 2002) with simple heuristic rules. Champollion is a sentence aligner based on bilingual dictionaries. We combine three bilingual dictionaries as the dictionary for Champollion: namely, LDC_CE_DIC2.0[7] constructed by LDC, bilingual terms in HowNet [8] and the bilingual lexicon in Champollion. The major steps for mining high-quality parallel sentences in comparable patents are as follows.

1) Champollion is used to preliminarily align the sentences in each section of the comparable patents to generate parallel sentence pair candidates. According to Lu et al. (2009), the generated candidates should have much noise and we will further explore filtering methods to remove misaligned sentences.

2) We remove sentence pairs using length filtering and ratio filtering. For length filtering, if a sentence pair has less than 100 words in the English sentence and less than 333 characters in the Chinese one, it is kept. Otherwise, it is removed. For ratio filtering, we discard the sentence pair candidates with Chinese- English length ratio outside the range of 0.8 to 1.8. The selection of the parameters here is set empirically based on the evaluation on a small sample of the large corpus.

3) MS aligner is utilized to further filter the parallel sentence candidates. MS aligner is a two-phase sentence aligner with high precision as its characteristics, and in the first pass it does alignment by using sentence length information (Gale and Church, 1991), and in the second pass it uses the sentence pairs aligned in the first pass to train an IBM Model-1 (Brown et al., 1993) and realign the sentences with the trained model.

Table 3 shows the statistics of the sentence numbers and the respective percentages of sentences kept with respect to all the sentence candidates in each step above.

Steps		1. CH	2.1 LF	2.2 RF	3. MS (final)
Abstr.	Num.	251K	243K	176K	83K
	Percent	100%	96.5%	70%	33%
Claims	Num.	3.0M	2.9M	2.1M	1.0M
	Percent	100%	96.5%	72.1%	33.4%
Desc.	Num.	19.3M	18.8M	13.4M	6.1M
	Percent	100%	97.2%	69.4%	31.3%
Total[9]	Num.	22.6M	21.9M	15.8M	7.2M
	Percent	100%	97.1%	69.8%	31.7%
Average	Num.	286	277	200	91

Table 3. Statistics of Parallel Sentences during the Aligning Process

In the first row of Table 3, *1.CH* denotes the first step of using the Champollion to align sentences; *2.1 LF* denotes the length filter in the second step; *2.2 RF* refers to the ratio filter in the second step; *3. MS* refers to the third and

[7] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

[8] http://www.keenage.com/html/e_index.html

[9] Here the total number does not include the number of titles. Here we did not use any method to filter the corresponding titles, and just treat them as parallel.

final step of using MS aligner to filter sentence pair candidates.

From Table 3, we can observe that 1) by using Champollion, we obtain about 22 million sentence pair candidates; 2) by filtering in step 2, the number of parallel sentences is reduced by 30%, to 16 million; 3) by using MS aligner, we final arrive at about 7 million parallel sentences.

The final parallel sentences are manually evaluated by randomly sampling 100 sentence pairs for each section of title, abstract, claims and description. The evaluation metric follows the one in Lu et al. (2009), which classifies each sentence pair into *Correct, Partially Correct* or *Wrong*[10]. The results of manual evaluation are shown in Table 4, from which we can see that the percentages of *correct* parallel sentences are quite high, and the wrong percentages are no higher than 5%. Therefore, we could conclude that the mined parallel sentences are high-quality with less than 5% wrong parallel sentences. Meanwhile, the abstract section shows the highest correct percentage, while the description section shows the lowest.

	Correct	*Partially Correct*	*Wrong*
Abstr.	97%	2%	1%
Claims	92%	3%	5%
Desc.	89%	8%	3%

Table 4. Manual evaluation of the final corpus

One may notice that the average number of parallel sentences extracted from one comparable patent in this study is 91, while for the corpus in Lu et al. (2009), it is only about 26 (~160K/6100). Here we recomputed the average numbers of Chinese characters, English words, and Chinese and English sentences for each comparable patent in Lu et al. (2009), which are shown in Table 5.

	Chinese		**English**	
	#Char	**#Sent**	**#Word**	**#Sent**
Avg/Patent	5.8K	119	4.4K	169

Table 5. Data Statistics of Comparable Patents in Lu et al.(2009)

Comparing Table 5 with Table 2, we can see that the comparable patents in Lu et al. (2009) are much smaller than those in this study in terms of numbers of Chinese characters, English words, and Chinese/English sentences. Therefore, the average number of parallel sentences extracted from the patents in this study is much bigger than that in Lu et al. (2009).

The possible explanation is that the patents in Lu et al. (2009) were first filed in China from 1996 to 2006 and later filed in USA from 1996 to 2008, and the applicants were still in their initial stage of learning how to write patent applications which may contain less content than those in this study involving patents filed by more experienced western companies.

[10] *Correct* means the English sentence is exactly the literal translation of the Chinese one, or the content overlap between them are above 80%; *partially correct* means the Chinese sentence and the English one are not the literal translation of each other, but the content of one sentence can cover more than 50% of the other; *wrong* means the contents of the Chinese sentence and the English one are not related, or more than 50% of the content of one sentence is missing in the other.

4.2 SMT experiments

As we have known, few SMT experiments on the English-Chinese patent translation have been reported, especially with a large scale of parallel sentences. We select 101,000 parallel sentences and divide them into three parts: 1 million sentence pairs for training, 500 sentence pairs for development and another 500 sentence pairs for testing. The statistics for the three parts are shown in Table 6.

	Language	**#Sentence pairs**	**#Words**
Training	English	1M	33.4M
	Chinese	1M	32.1M
Development	English	500	17.2K
	Chinese	500	16.1K
Test	English	500	17.2K
	Chinese	500	16.1K

Table 6. Data for SMT Experiments

An SMT system is setup using Moses (Koehn, 2007). We test translation in both directions (namely, Chinese to English and English to Chinese) with/without optimized parameters. The BLEU scores are as shown in Table 7. "*No MERT*" denotes the cases without optimizing parameters using minimal error-rate training (MERT) (Och, 2003) algorithm whereas "*MERT*" denotes the cases with parameter optimization of MERT on development data.

	Chinese->English		**English->Chinese**	
BLEU	No MERT	MERT	No MERT	MERT
	0.273	0.274	0.207	0.240

Table 7. SMT experiment results

The BLEU scores here seem promising, which show that the parallel sentences extracted are of good quality for training the SMT engine. We could expect better results with more training data.

Moreover, we use the 160K parallel sentences in Lu et al. (2009) as the training data to build an SMT system, and the BLEU score for Chinese to English translation is 0.179 on the test data of 500 parallel sentences mentioned above with the MERT optimization on development data. The BLEU score of 0.274 in Table 7 based on 1 million parallel sentences shows a significant 53% relative improvement compared the BLEU score of 0.179, which demonstrates that with more training data we can get better SMT performance.

The BLEU scores for Chinese to English translation in Table 7 seem much better than those for the opposite direction. This is different from the results in NIST SMT evaluation, in which the highest BLEU scores for English to Chinese translation are usually better than those for Chinese to English translation. The possible reasons are: 1) the BLEU scores in this study are calculated without considering recasing or detokenization so we essentially ignore errors caused by them, while in NIST evaluation, recasing and detokenization are essential steps. 2) the evaluation of Chinese sentences is influenced by the boundary of Chinese words. Even when the whole sentence is correct, if the word boundaries are wrong, we

would get a low score. However, the English tokenization is much easier compared to Chinese word segmentation because there is no word boundary problem for English. 3) another relevant factor may be the translation direction of the test data, which is from English to English. Could the direction of human translation have an influence on the BLEU scores? Ozdowska & Way (2009) showed that "*data containing original French and English translated from French is optimal when building a system translating from French into English. Conversely, using data comprising exclusively French and English translated from several other languages is suboptimal regardless of the translation direction.*" Since no such observation seems have been found in Chinese-English translation, we raise the question here and are looking forward to further investigation.

Meanwhile, the MERT algorithm shows better performance on the English to Chinese translation but not on the reverse direction. One possible explanation is that MERT improves the performance with respect to the Chinese word boundary.

The server used for parallel sentence mining and SMT in this study has a 12G memory and 4 two-core 2.67GHz CPUs. Although the server is already much better than common PCs, it is still not powerful enough to do the computing-intensive SMT related tasks. Therefore, our SMT experiments only use a small part of the whole corpus, i.e. only 1 million out of more than 7 million sentence pairs.

6. Discussion

Here we briefly describe the efforts spent for this project. The Chinese and English websites from which the Chinese and English patents were downloaded were quite slow to access, and were occasionally down during access. Meanwhile, some patents are quite large. For example, the Chinese patent with the application number of CN200680029419.3 has 340 pages of description and 40 pages of claims, and its corresponding English patent has 396 pages of description and 46 pages of claims. These large patents would cost much time for the websites to respond and had be specifically handled. To avoid too much workload for the websites, the downloading speed had been limited. It took considerable efforts among different parties to obtain these comparable patents. By comparison, the efforts spent for parallel sentence mining and SMT experiments were much less.

According to recent investigation in 2010, the number of Chinese patent applications with English as the original language has rapidly increased, and we could expect more English-Chinese comparable patents to be filed quickly. This would allow further efforts to enlarge our corpus.

The method and approach proposed here to mine comparable patents should be also applicable to other language pairs, such as English and Japanese, English and Korean, etc. What is more, we could even build trilingual or multilingual parallel corpus by using the PCT patents and their multiple versions in different languages, such as Japanese (JP), Chinese (CH), Korean (KR), English (EN), German (DE), etc. We have searched via the website of WIPO to get an estimate on the quantity of PCT applications which were published in English and later filed in other countries in their corresponding languages, and found that the quantity of bilingual and multilingual patents for CH, KR, JP and EN seems quite considerable, which means that the multilingual patents for these languages could be harvested in remarkable quantities. For example, we have began to build a small trilingual patent corpus by leveraging the PCT patents, i.e. we search for comparable patents filed in simplified Chinese in China, filed in traditional Chinese in Taiwan, and filed in English as a PCT patent (Tsou and Lu, 2010). Although the language varieties found in mainland China and Taiwan are not two distinct languages, there are enormous differences in terms of technical terminology and even syntactic structure, this corpus is still quite useful to compare the two versions of the same PCT patent in China and Taiwan because there are linguistic convergences.

What is of special interest here is the very concept of *"parallel corpus"* in the context of translation. The commonly used BLEU and NIST scores in SMT evaluation just reduce the concept of parallelism to a rather technical mapping of language units. But it is well known that high-quality human translations often do not keep sentence units of the source language. Therefore, we may need more elaborate schemes to better evaluate the quality of machine translation, and translation studies (Munday, 2001) retain its importance.

7. Conclusion and future work

In this paper, we introduce our large parallel corpus which is extracted from a large corpus of English-Chinese comparable patents harvested from the Web. We first preliminarily mine parallel sentence pairs with Champollion, a publicly available sentence aligner, and then further filter the candidates with another sentence aligner, namely MS Aligner. Then, about 7 million high-quality parallel sentences out of more than 22 million bilingual sentence pair candidates are chosen as the final parallel corpus. As we know, this is the largest parallel sentence corpus in the patent domain. Based on the 1 million parallel sentences extracted from the *abstract* and *claims* section, some preliminary SMT results are also reported here.

Meanwhile, with our experimental sentence alignment efforts, only 7 million parallel sentences have been mined from 22 million sentence pair candidates. By exploring more complicated and possibly more accurate approaches such as Munteanu and Marcu (2005) or Lu et al. (2009), we could expect to find more parallel sentences from the comparable patents. More SMT experiments would be done as well since we currently only utilize 1 million parallel sentences in our SMT experiment due to limited time and computer resources.

Since different (sub-)sections (namely, title, abstract, claims, description, and subsections in the description part) in patents have their own writing styles which may influence the word choice and syntactic structure of the sentences, as well as patents cover many technical domains (such as chemistry, biomedicine, electronics, vehicle, etc.), experiments on cross-section and cross-IPC (International Patent Classification) machine translation could be enlightening for further understanding the characteristics of individual sections and technical domains. For example, *claims* have legal effect, and tend to use more relative clauses modifying head words.

Some sampled parallel sentences are available at http://livac.org/smt/parpat.html. We should be able to make some parts of our large parallel corpus available to the research community in the near future. Given the relative paucity of parallel patent data, this large parallel corpus shall be a helpful step towards MT research and other cross-lingual information access applications, in the above mentioned technical domains and especially in the patent domain. Last but not least, our method and approach should be applicable to other languages, which show a novel way on how to reduce the data acquisition bottleneck in multilingual language processing.

8. Acknowledgements

We acknowledge the contributions of Jacky Hui, Andy Chin and other colleagues of the Language Information Sciences Research Centre, City University of Hong Kong and of ChiLin Star Corp.

9. References

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty J.D., Mercer, R.L., & Roossin, P.S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P.F., Lai, J.C., & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp.169-176.

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., & Mercer, R.L. (1993). Mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263-311.

Chen, S.F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp. 9-16.

CHIANG, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*. pp. 263-270.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics,* 33(2), 201–228.

Fujii, A., Utiyama, M., Yamamoto, M., & Utsuro, T. (2008). Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR)*. pp. 389-400. Tokyo, Japan.

Gale, W.A., & Church, K.W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp.79-85.

Jiang, T., Tsou, B.K., & Lu, B. (2010). Part-of-speech model for N-best list reranking in experimental English-Chinese SMT. In *Proceedings of 1st International Workshop on Advances in Patent Information Retrieval*. Milton Keynes, UK.

Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*.

Koehn, P., Hoang H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) Demo Session.* pp. 177-180.

Landry, P.F. (2008). How weak institutions can produce strong regimes: Patents, lawyers, and the improbable creation of an intellectual property regime in China (1985-2007). *Paper presented at Workshop on Rule of Law*, Yale University, March 28-29.

Lu, B., Tsou, B.K., Zhu,J., Jiang, T., & Kwong, O.Y. (2009). The construction of an English-Chinese patent parallel corpus. In *Proceedings of MT Summit XII 3rd Workshop on Patent Translation*. pp. Ottawa, Canada.

Lu, B., & Tsou, B.K. (2009). Towards bilingual term extraction in comparable patents. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*. pp. 755-762. Hong Kong. December, 2009.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*. pp.135-144.

Munday, J. (2001). *Introducing translation studies: theories and applications*. Oxon, UK: Routledge.

Munteanu, D.S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, *31*(4), 477–504.

Och, F.J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pp. 160-167.

Och, F.J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51.

Och, F.J., & Ney, H. (2004). The alignment template approach to machine translation. *Computational Linguistics*, *30*(4), 417-449.

Ozdowska, Sylwia and Way, Andy. (2009) Optimal bilingual data for French-English PB-SMT. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*.

Resnik, P., & Smith, N.A. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Utiyama, M., & Isahara, H. (2007). A Japanese-English patent parallel corpus. In *Proceeding of MT Summit XI*. pp. 475–482.

Simard, M., & Plamondon, P. (1998). Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, *13*(1), 59-80.

Sun, Y. (2003). Determinants of foreign patents in China. *World Patent Information,* 25, pp. 27-37.

Tsou, B.K., & LU, B. (2010). Automotive patents from Mainland China and Taiwan: A preliminary exploration of terminological differentiation and content convergence. *World Patent Information*. (to appear)

Wu, D., & Fung, P. (2005). Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora. In *Proceedings of IJCNLP2005*.

Zhao, B., & Vogel, S. (2002). Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of Second IEEE International Conference on Data Mining (ICDM'02)*.