Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT

Bin Lu[†], **Benjamin K. Tsou**[† ʃ], **Tao Jiang**[$], **Oi Yee Kwong**[†], **and Jingbo Zhu**[£]

[†]Department of Chinese, Translation & Linguistics, City University of Hong Kong

[ʃ]Research Centre on Linguistics and Language Information Sciences,
Hong Kong Institute of Education

[$]ChiLin Star Corp., Southern Software Park, Zhuhai, China

[£]Natural Language Processing Lab, Northeastern University, Shenyang, China

{lubin2010, rlbtsou, jiangtaoster}@gmail.com,
rlolivia@cityu.edu.hk, zhujingbo@mail.neu.edu.cn

Abstract

In this paper, we demonstrate how to mine large-scale parallel corpora with multilingual patents, which have not been thoroughly explored before. We show how a large-scale English-Chinese parallel corpus containing over 14 million sentence pairs with only 1-5% wrong can be mined from a large amount of English-Chinese bilingual patents. To our knowledge, this is the largest single parallel corpus in terms of sentence pairs. Moreover, we estimate the potential for mining multilingual parallel corpora involving English, Chinese, Japanese, Korean, German, etc., which would to some extent reduce the parallel data acquisition bottleneck in multilingual information processing.

1 Introduction

Multilingual data are critical resources for building many applications, such as machine translation (MT) and cross-lingual information retrieval. Many parallel corpora have been built, such as the Canadian Hansards (Gale and Church, 1991), the Europarl corpus (Koehn, 2005), the Arabic-English and English-Chinese parallel corpora used in the NIST Open MT Evaluation.

However, few parallel corpora exist for many language pairs, such as Chinese-Japanese, Japanese-Korean, Chinese- French or Japanese-German. Even for language pairs with several parallel corpora, such as Chinese-English and Arabic-English, the size of parallel corpora is still a major limitation for SMT systems to achieve higher performance.

In this paper, we present a way which could, to some extent, reduce the parallel data acquisition bottleneck in multilingual language processing. Based on multilingual patents, we show how an enlarged English-Chinese parallel corpus containing over 14 million high-quality sentence pairs can be mined from a large number of comparable patents harvested from the Web. To our knowledge, this is the largest single parallel corpus in terms of parallel sentences. Some SMT experiments are also reported. Moreover, we investigate the potential to get large-scale parallel corpora for languages beyond the Canadian Hansards, Europarl and UN news used in NIST MT Evaluation by estimating the quantity of multilingual patents involving English, Chinese, Japanese, Korean, German, etc.

Related work is introduced in Section 2. Patents, PCT patents, multilingual patents are described in Section 3. Then an English-Chinese parallel corpus, its mining process and application to SMT are introduced in Section 4, followed by the quantity estimation of multilingual patents involving other language pairs in Section 5. We discuss the results in Section 6, and conclude in Section 7.

2 Related Work

Parallel sentences could be extracted from parallel documents or comparable corpora. Different approaches have been proposed to align sentences in parallel documents consisting of the same content in different languages based on the following information: 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).

To overcome the lack of parallel documents, comparable corpora are also used to mine parallel sentences, which raises further challenges since the bilingual contents are not strictly parallel. For instance, Zhao and Vogel (2002) investigated the mining of parallel sentences for Web bilingual news. 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. Cao et al., (2007) and Lin et al., (2008) proposed two different methods utilizing the parenthesis pattern to extract term translations from bilingual web pages. Jiang et al. (2009) presented an adaptive pattern-based method which produced Chinese-English bilingual sentences and terms with over 80% accuracy.

Only a few papers were found on the related work in the patent domain. Higuchi et al. (2001) used the titles and abstracts of 32,000 Japanese-English bilingual patents to extract bilingual terms. Utiyama and Isahara (2007) mined about 2 million parallel sentences by using two parts in the *description* section of Japanese-English comparable patents. Lu et al. (2009) derived about 160K parallel sentences from Chinese-English comparable patents by aligning sentences and filtering alignments with the combination of different quality measures. Another closely related work is the English-Chinese parallel corpus (Lu et al., 2010), which is largely extended by this work, in which both the number of patents and that of parallel sentences are augmented by about 100%, and more SMT experiments are given. Moreover, we show the potential for mining parallel corpora from multilingual patents involving other languages.

For statistical machine translation (SMT), tremendous strides have been made in two decades, including Brown et al. (1993), Och and Ney (2004) and Chiang (2007). For the MT evaluation, NIST (Fujii et al., 2008; 2010) has been organizing open evaluations for years, and the performance of the participants has been improved rapidly.

3 Patents and Multilingual Patents

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[1]). A patent application consists of different sections, and we focus on the text, i.e. only *title, abstract, claims and description*.

3.1 PCT Patents

Since the invention in a patent is only protected in the filing countries, a patent applicant who wishes to protect his invention outside the original country should file patents in other countries, which may involve other languages.

The Patent Cooperation Treaty (PCT) system offers inventors and industry an advantageous route for obtaining patent protection internationally. By filing one *"international"* patent application under the PCT via the World Intellectually Property Organization (WIPO), protection of an invention can be sought simultaneously (i.e. the priority date) in each of a large number of countries.

The number of PCT international applications

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

filed is more than 1.7 million[2]. 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[3] shows the number of PCT applications for the most used languages of filing and publication.

	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

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.

3.2 Multilingual Patents

A PCT application does not necessarily mean a multilingual patent. An applicant who has decided to proceed further with his PCT international application must fulfill the requirements for entry into the PCT national phase at the patent offices of countries where he seeks protection. For example, a Chinese company may first file a Chinese patent in China patent office and then file its international application also in Chinese under the PCT. Later on, it may have the patent translated into English and file it in USA patent office, which means the patent becomes bilingual. If the applicant continues to file it in Japan with Japanese, it would be trilingual. Even more, it would be quadrilingual or involve more languages when it is filed in other countries with more languages.

Such multilingual patents are considered comparable (or noisy parallel) because they are not parallel in the strict sense but still closely related in terms of information conveyed (Higuchi et al., 2001; Lu et al., 2009).

4 A Large English-Chinese Parallel Corpus Mined from Bilingual Patents

In this section, we introduce the English-Chinese bilingual patents harvested from the Web and the method to mine parallel sentences from them. SMT experiments on the final parallel corpus are also described.

4.1 Harvesting English-Chinese Bilingual Patents

The official patent office in China is the State Intellectual Property Office (SIPO). In early 2009, by searching on its website, we found about 200K Chinese patents previously filed as PCT applications in English and crawled their *bibliographical data, titles, abstracts* and *the major claim* from the Web, and then *other claims* and *descriptions* were also added. Since some contents are in the image format, the images were OCRed and the texts recognized were manually verified.

All PCT patent applications are filed through WIPO. With the Chinese patents mentioned above, the corresponding English patents were searched from the website of WIPO by the PCT publication numbers to obtain relevant sections of the English PCT applications, including *bibliographical data, title, abstract*, *claims* and *description*. About 80% (160K) out of the Chinese patents found their corresponding English ones. Some contents of the English patents were OCRed by WIPO.

[2] Retrieved Apr., 2010 from http://www.wipo.int/pctdb/en/. The data below involving PCT patents comes from the website of WIPO.

[3] The data in this and other tables in the following sections involving PCT patents comes from the website of WIPO.

[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.

We automatically split the patents into individual sections according to the respective tags inside the patents, and segmented each section sentences according to punctuations. The statistics of each section for Chinese and English patents are shown in Table 2.

Sections	Chinese		English	
	#Char	#Sent	#Word	#Sent
Title	2.7M	157K	1.6M	157K
Abstract	33M	596K	20M	784K
Claim	367M	6.8M	217M	7.4M
Desc.	2,467M	48.8M	1,353M	54.0M
Total	2,870M	56.2M	1,591M	62.3M

Table 2. Statistics of Comparable Patents

4.2 Mining Parallel Sentences from Bilingual Patents

The sentences in each section of Chinese patents were aligned with those in the corresponding section of the corresponding English patents to find parallel sentences after the Chinese sentences were segmented into words.

Since the comparable patents are not strictly parallel, the individual alignment methods mentioned in Section 2 would be not effective: 1) the length-based method is not accurate since it does not consider content similarity; 2) the bilingual dictionary-based method cannot deal with new technical terms in the patents; 3) the translation model-based method would need training data to get a translation model. Thus, in this study we combine these three methods to mine high-quality parallel sentences from comparable patents.

We first use a bilingual dictionary to preliminarily align the sentences in each section of the comparable patents. The dictionary-based similarity score P_d of a sentence pair is computed based on a bilingual dictionary as follows (Utiyama and Isahara, 2003):

$$p_d(S_c,S_e)=\frac{\sum_{w_c\in S_c}\sum_{w_e\in S_e}\frac{\gamma(w_c,w_e)}{\deg(w_c)\deg(w_e)}}{(\ l_e+l_c)\ /2}$$

where w_c and w_e are respectively the word types in Chinese sentence S_c and English sentence S_e; l_c and l_e respectively denote the lengths of S_c and S_e in terms of the number of words; and $\gamma(w_c,w_e)$ = 1 if w_c and w_e is a translation pair in the bilingual dictionary or are the same string, otherwise 0; and

$$\deg(w_c)=\sum_{w_e\in S_e}\gamma(w_c,w_e)$$

$$\deg(w_e)=\sum_{w_e\in S_c}\gamma(w_c,w_e).$$

For the bilingual dictionary, we combine three ones: namely, LDC_CE_DIC2.0[5] constructed by LDC, bilingual terms in HowNet and the bilingual lexicon in Champollion (Ma, 2006).

We then remove sentence pairs using length filtering and ratio filtering: 1) for length filtering, if a sentence pair has more than 100 words in the English sentence or more than 333 characters in the Chinese one, it is removed; 2) for length ratio filtering, we discard the sentence pairs with Chinese-English length ratio outside the range of 0.8 to 1.8. The parameters here are set empirically.

We further filter the parallel sentence candidates by learning an IBM Model-1 on the remaining aligned sentences and compute the translation similarity score P_t of sentence pairs by combining the translation probability value of both directions (i.e. Chinese->English and English->Chinese) based on the trained IBM-1 model (Moore, 2002; Chen, 2003; Lu et al, 2009). It is computed as follows:

$$p_t(S_c,S_e)=\frac{log(P(S_e\mid S_c))+log(P(S_c\mid S_e))}{l_c+l_e}$$

where $P(S_e\mid S_c)$ denotes the probability that a translator will produce S_e in English when presented with S_c in Chinese, and vice versa for $P(\mathrm{S}_c\mid S_e)$. Sentence pairs with

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

similarity score P_t lower than a predefined threshold are filtered out as wrong aligned sentences.

Table 3 shows the sentence numbers and the percentages of sentences kept in each step above with respect to all sentence pairs. In the first row of Table 3, *1.DICT* denotes the first step of using the bilingual dictionary to align sentences; *2. FL* denotes the length and ratio filtering; *3. TM* refers to the third and final step of using translation models to filter sentence pairs.

	1. DICT	2.FL	3. TM (final)
Abstr.	503K	352K (70%)	166K (33%)
Claims	6.0M	4.3M (72.1%)	2.0M (33.4%)
Desc.	38.6M	26.8M (69.4%)	12.1M (31.3%)
Total[6]	45.1M	31.5M (69.8%)	14.3M (31.7%)

Table 3. Numbers of Sentence Pairs

Both the 31.5M parallel sentences after the second step *FL* and the final 14.3M after the third step *TM* are manually evaluated by randomly sampling 100 sentence pairs for each section. The evaluation metric follows the one in Lu et al. (2009), which classifies each sentence pair into *Correct, Partially Correct* or *Wrong*. The results of manual evaluation are shown in Table 4.

	Section	Correct	Partially Correct	Wrong
2. FL	Abstr.	85%	7%	8%
	Claims	83%	10%	7%
	Desc.	69%	15%	15%
3. TM (final)	Abstr.	97%	2%	1%
	Claims	92%	3%	5%
	Desc.	89%	8%	3%

Table 4. Manual Evaluation of the Corpus

From Table 4, we can see that: 1) In the final corpus, the percentages of *correct* parallel sentences are quite high, and the wrong percentages are no higher than 5%; 2) Without the final step of TM, the accuracies of 31.5M sentence pairs are between 69%-85%, and the percentages of wrong pairs are between 7%-15%; 3) The abstract section shows the highest correct percentage, while the description section shows the lowest.

Thus, we could conclude that the mined 14M parallel sentences are of high quality with only 1%-5% wrong pairs, and our combination of bilingual dictionaries and translation models for mining parallel sentences are quite effective.

4.3 Chinese-English Statistical Machine Translation

A Chinese-English SMT system is setup using Moses (Koehn, 2007). We train models based on different numbers of parallel sentences mined above. The test set contains 548 sentence pairs which are randomly selected and different from the training data. The sizes of the training data and BLEU scores for the models are shown in Table 5.

System	BLEU (%)	#Sentence Pairs for training
Model-A	17.94	300K
Model-B	19.96	750K
Model-C	20.09	1.5M
Model-D	20.98	3M
Model-E	22.60	6M

Table 5. SMT Experimental Results

From Table 5, we can see that the BLEU scores are improving steadily when the training data increases. When the training data is enlarged by 20 times from 300K to 6M, the BLEU score increases to 22.60 from 17.94, which is quite a significant improvement. We show the translations of one Chinese sample sentence in Table 6 below.

CN Sent.	电机 主轴 伸入 压缩机 壳体 内 的 工作 腔 中 ，
Ref.	the main shaft of the electric motor extends into the working cavity of the compressor shell ,
Model-A	the motor main shaft into the compressor the chamber

[6] Here the total number does not include the number of titles, which are directly treated as parallel.

Model-B	motor shaft into the compressor housing . the working chamber
Model-C	motor shaft into the compressor housing . the working chamber
Model-D	motor spindle extends into the compressor housing . the working chamber
Model-E	motor spindle extends into the working chamber of the compressor housing ,

Table 6. Translations of One Chinese Sentence

From Table 6, we can see the translations given by Model-A to Model-C are lack of the main verb, the one given by Model-D has an ordering problem for the head noun and the modifier, and the one given by Model-E seems better than the others and its content is already quite similar to the reference despite the lexical difference.

5 Multilingual Corpora for More Languages

In this section, we describe the potential of building large-scale parallel corpora for more languages, especially Asian languages by using the 1.7 million PCT patent applications and their national correspondents. By using PCT applications as the pivot, we can build multilingual parallel corpora from multilingual patents, which would greatly enlarge parallel data we could obtain.

The patent applications filed in one country should be in the official language(s) of the country, e.g. the applications filed in China should be in Chinese, those in Japan be in Japanese, and so on. In Table 7, the second column shows the total numbers of patent applications in different countries which were previously filed as PCT ones; and the third column shows the total numbers of applications in different countries, which were previously filed as PCT ones with English as language of publication.

National Phase Country[7]	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 7. Estimated Numbers of Multilingual Patents

The number of the Chinese-English bilingual patents (CE) in Table 7 is about 188K, which is consistent with the number of 160K found in Section 4.1 since the latter contains only the applications up to early 2009. Based on Table 7, we estimate below the rough sizes of bilingual corpora, trilingual corpora, and even quadrilingual corpora for different languages.

1) Bilingual Corpora with English as one language

Compared to CE (188K), the Japanese-English bilingual corpus (269K) could be 50% larger in terms of bilingual patents, the Korean-English one (134K) could be about 30% smaller, and the German-English one (10K) would be much smaller.

2) Bilingual Corpora for Asian Languages

The Japanese-Chinese bilingual corpus (189K) could be comparable to CE (188K) in terms of bilingual applications, the Chinese-Korean one (154K) could be about 20% smaller, and the Japanese-Korean one (158K) is quite similar to the Chinese-Korean one.

3) Trilingual Corpora

In addition to bilingual corpora, we can also build trilingual corpora from trilingual patents. It is quite interesting to note that the trilingual corpora could be quite large even compared to the bilingual corpora.

The trilingual corpora for Chinese, Japanese and English (130K) could be only 30% smaller than CE in terms of patents. The trilingual corpus

[7] 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, i.e. the numbers are not up-to-date .

for Chinese, Korean and English (91K) and that for Japanese, Korean and English (103K) are also quite large. The number of the trilingual patents for the Asian languages of Chinese, Japanese and Korean (106K) is about 54% of that of CE.

4) Quadrilingual Corpora

The number of the quadrilingual patents for Chinese, Japanese, Korean and English (73K) is about 38% of that of CE. From these figures, we could say that a large proportion of the PCT applications published in English later have been filed in all the three Asian countries: China, Japan, and R. Korea.

6 Discussion

The websites from which the Chinese and English patents were downloaded were quite slow to access, and were occasionally down during access. To avoid too much workload for the websites, the downloading speed had been limited. Some large patents would cost much time for the websites to respond and had be specifically handled. It took considerable efforts to obtain these comparable patents.

In addition our English-Chinese corpus mined in this study is at least one order of magnitude larger, we give some other differences between ours and those introduced in Section 2 (Higuchi et al., 2001; Utiyama and Isahara, 2007; Lu et al, 2009)

1) Their bilingual patents were identified by the priority information in the US patents, and could not be easily extended to language pairs without English; while our method using PCT applications as the pivot could be easily extended to other language pairs as illustrated in Section 5.

2) The translation process is different: their patents were filed in USA Patent Office in English by translating from Japanese or Chinese, while our patents were first filed in English as a PCT application, and later translated into Chinese. The different translation processes may have different characteristics.

Since the PCT and multilingual patent applications increase rapidly in recent years as discussed in Section 3, we could expect more multilingual patents to enlarge the large-scale parallel corpora with the new applications and keep them up-to-date with new technical terms. On the other hand, patents are usually translated by patent agents or professionals, we could expect high quality translations from multilingual patents. We have been planning to build trilingual and quadrilingual corpora from multilingual patents.

One possible limitation of patent corpora is that the sentences are all from technical domains and written in formal style, and thus it is interesting to know if the parallel sentences could improve the performance of SMT systems on NIST MT evaluation corpus containing news sentences and web sentences.

7 Conclusion

In this paper, we show how a large high-quality English-Chinese parallel corpus can be mined from a large amount of comparable patents harvested from the Web, which is the largest single parallel corpus in terms of the number of parallel sentences. Some sampled parallel sentences are available at http://www.livac.org/smt/parpat.html, and more parallel sentences would be publicly available to the research community.

With 1.7 million PCT patent applications and their corresponding national ones, there are considerable potentials of constructing large-scale high-quality parallel corpora for languages. We give an estimation on the sizes of multilingual parallel corpora which could be obtained from multilingual patents involving English, Chinese, Japanese, Korean, German, etc., which would to some extent reduce the parallel data acquisition bottleneck in multilingual information processing.

Acknowledgements

We wish to thank Mr. Long Jiang from Microsoft Research Asia and anonymous reviewers for their valuable comments.

References

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of ACL*. pp.169-176.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. Mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263-311.

Cao, Guihong, Jianfeng Gao and Jianyun Nie. 2007. A System to Mine Large-scale Bilingual Dictionaries from Monolingual Web Pages. In *Proceedings of MT Summit*. pp. 57-64.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL*. pp. 9-16.

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

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.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. In *Proceedings of the NTCIR-8 Workshop*. Tokyo, Japan.

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

Higuchi, Shigeto, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. PRIME: A System for Multi-lingual Patent Retrieval. In *Proceedings of MT Summit VIII*, pp.163-167, 2001.

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL Demo Session.* pp. 177-180.

Lin, Dekang, Shaojun Zhao, Benjamin V. Durme and Marius Pasca. 2008. Mining Parenthetical Translations from the Web by Word Alignment. In *Proceedings of ACL-08*. pp. 994-1002.

Jiang, Long, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In *Proceedings of ACL-IJCNLP*. pp. 870-878.

Lu, Bin, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Olivia Y. Kwong. 2009. The Construction of an English-Chinese Patent Parallel Corpus. *MT Summit XII 3rd Workshop on Patent Translation*.

Lu, Bin, Tao Jiang, Kapo Chow and Benjamin K. Tsou. 2010. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT. *LREC Workshop on Building and Using Comparable Corpora*. Malta. May, 2010.

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy.

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of AMTA*. pp.135-144.

Munteanu, Dragos S., and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-parallel Corpora. *Computational Linguistics*, *31*(4), 477–504.

Och, Franz J., and Hermann Ney. 2004. The Alignment Template Approach to Machine Translation. *Computational Linguistics*, *30*(4), 417-449.

Simard, Michel, and Pierre Plamondon. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation*, *13*(1), 59-80.

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

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