The Construction of a Chinese-English Patent Parallel Corpus

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

[†]Language Information Sciences Research
Centre, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong

[£]Natural Language Processing Lab
Institute of Computer Software and Theory
Northeastern University, ShenYang, China

lubin2@student.cityu.edu.hk rlbtsou@cityu.edu.hk zhujingbo@mail.neu.edu.cn
jingtaoster@gmail.com rlolivia@cityu.edu.hk

Abstract

In this paper, we describe the construction of a parallel Chinese-English patent sentence corpus which is created from noisy parallel patents. First, we use a publicly available sentence aligner to find parallel sentence candidates in the noisy parallel data. Then we compare and evaluate three individual measures and different ensemble techniques to sort the parallel sentence candidates according to the confidence score and filter out those with low scores as the noisy data. The experiment shows that the combination of measures outperforms the individual measures, and that filtering out low-quality sentence pairs is readily justified as it can improve SMT performance. Finally, we arrive at the final corpus consisting of 160K sentence pairs in which about 90% are correct or partially correct alignments.

1 Introduction

Parallel corpora are invaluable resources for many NLP applications, such as 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 Chinese-English parallel corpora used in the NIST Open MT Evaluation [1] and Europarl corpus (Koehn, 2005). However, few parallel corpora exist in the patent domain. The exception is the Japanese-English patent parallel corpus (Utiyama and Isahara, 2007) provided for the NTCIR-7 patent machine translation task (Fujii et al., 2008).

We utilized about 7,000 noisy parallel Chinese-English patents to construct the corpus of parallel sentences. We first compared and evaluated three publicly available sentence aligners and chose one of them to align the sentences in the noisy parallel patents. Because of the loose translation problem in the parallel patents (as will be discussed later), the results include a large proportion of incorrect alignments.

To filter out the incorrect sentences, we compared and evaluated three individual measures and different ensemble techniques. The three measures are the length-based score, the dictionary-based score, and the translation probability score. The experiments showed that the three measures performed differently and that combining all three improved the performance of sentence filtering. Furthermore, we evaluated the effects of sentence filtering on SMT performance. Finally, we set up the final patent parallel corpus consisting of 160K sentence pairs among of which about 90% are correct or partially correct alignments.

In what follows, we present the related work in Sec. 2, describe the noisy parallel Chinese-English patent data in Sec. 3 and the preliminary sentence alignment in Sec. 4., and introduce sentence filtering, including the evaluation of its impact on SMT in Sec. 5 as well as the final parallel corpus in Sec. 6, and conclude this paper.

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

2 Related Work

To get parallel sentences from parallel corpora, different approaches can be used for sentence alignment. The approaches can be based on a) sentence length, b) lexical information in bilingual dictionaries, c) statistical translation model, or d) the composite of more than one approach.

The sentence-length-based approach (Brown et al. 1991; Gale and Church, 1991) aligns sentences based on the number of words or characters in each sentence. Dictionary-based techniques use extensive online bilingual lexicons to match sentences. For instance, Ma (2006) described Champollion, a lexicon-based sentence aligner designed for robust alignment of potential noisy parallel text, and increased the robustness of the alignment by assigning greater weights to less frequent translated words.

Statistical translation model is also used for sentence alignment. Chen (1993) constructed a simple statistical word-to-word translation model on the fly during sentence alignment and then found the alignment that maximizes the probability of generating the corpus. Simard and Plamondon (1998) and Moore (2002) both used a composite method in which the first pass does alignment at the sentence length level and the second pass uses IBM Model-1.

Non-parallel corpora or comparable corpora, in addition to clean, ideal parallel corpora, are also used to mine parallel sentences. For instance, 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 docuements. Utiyama and Isahara (2003) aligned articles and sentences from noisy parallel news articles, then sorted the aligned sentences according to a similarity measure, and selected only the highly ranked aligned sentence alignments.

Although the construction of our Chinese-English patent parallel corpus is similar to that of the Japanese-English one (Utiyama and Isahara, 2007), we have made the following modifications on the basis of our data: 1) all sections of the patents, instead of only two parts in the description section, were used to find sentence alignments; 2) for sentence filtering, we integrated three individual measures, including the dictionary-based one (Utiyama and Isahara, 2007), and the experiments showed the combination of measures can improve the performance of sentence filtering. We also did SMT experiments, showing that filtering out misaligned sentences could improve SMT performance.

3 The Chinese-English Parallel Patents

We use about 7000 Chinese-English parallel patents with same/similar content to construct the parallel sentence corpus. The patents were first filed in the China Patent Office with Chinese as the original language. They were translated into English, and then filed in *USPTO* (United States Patent and Trademark Office). The parallel patents were identified by using the priority information described in the *USPTO* patents.

3.1 Data Description

Each patent has different parts, i.e. *title*, *abstract*, *claim*, *description*, etc, and the description section of some patents also have subdivision. Utiyama and Isahara (2007) used only the "*Detailed Description of the Preferred Embodiments*" and "*Background of the Invention*" part in the description section of each patent to find parallel sentences because they found these two parts have more literal translations than others. However, since our corpus has much less Chinese-English patent pairs, we use all parts of each patent to find parallel sentences. In total, there are about 730K Chinese sentences and 1,080K English sentences in the parallel patents. The detailed statistics for each section are shown in Table 1.

Sections	#Chinese Sentences	#English Sentences
Title	7K	7K
Abstract	29K	32K
Claim	145K	201K
Description	557K	840K
Total	738K	1,080K

Table 1. Statistics for each section

3.2 Problem of Loose Translation

Our observation indicated loose translations in Chinese-English parallel patents to be very common. We consider them as noisy parallel patents, which are not parallel in the strict sense but still closely related because almost the same information is conveyed (Zhao and Vogel, 2002). Higuchi et al. (2001) even considered the noisy parallel patents to be comparable, instead of parallel.

To evaluate the translations, the *abstract* sections of 100 patent pairs were taken from our patent data, and a bilingual annotator was asked to judge whether the abstracts are a) *literally translated*, b) *loosely translated* or c) *rewritten*[2]. The results showed their empirical distribution to be 55%, 26% and 19% respectively. This means that a large proportion of the abstracts are not literally translated.

There may be two major explanations for this phenomenon of common loose translations in these patents: 1) The field of intellectual property is highly regulated and different stylistic convensions may exist for patents in different countries. Thus the translation may be highly influenced by the stylistic differences in the individual countries; 2) For protection of intellectual property, the patent applicants may intentionally change some technical terms or the patent structure to broaden the patent coverage when a new version is produced into another language and country.

4 Preliminary Sentence Alignment

The noisy parallel patents are first segmented into sentences according to punctuations, and the Chinese sentences are segmented into words as was the case in Champollion (Ma, 2006).

To choose a sentence aligner, we first compare three publicly available sentence aligners, namely Champollion, Hunalign (Varga et al., 2007), and MS aligner (Microsoft Bilingual Sentence Aligner) (Moore, 2002), based on the manually aligned Chinese-English parallel corpus included in Champollion. For the bilingual dictionary needed by Champollion and Hunalign, we combine LDC_CE_DIC2.0[3] constructed by LDC, bilingual terms in HowNet[4] and the bilingual lexicon in Champollion. Since the MS aligner only extracts 1-1 sentence matches, we use only the 3,005 manually aligned 1-1 matches in the evaluation corpus so as to compare the three aligners on the same basis. The performance, including precision (P), recall (R) and F-score, is shown in Table 2.

	P (%)	R (%)	F-score (%)
Champollion	98.4	98.3	98.4
Hunalign	82.9	97.1	89.4
MS Aligner	95.4	92.5	93.9

Table 2. Performance of aligners on a small corpus

Because of its better performance than Hunalign and MS aligner, Champollion is chosen as the sentence aligner for our subsequent experiment to extract sentence pair candidates in the relevant sections of the noisy parallel patents. In total, 352K sentence pair candidates are extracted, including 1-1, 2-1, 1-2, 1-3, 3-1, 1-4 or 4-1 alignments. This means more than 48.6% of Chinese sentences or 32.6% of English sentences find their corresponding ones in the other language. The breakdown of sections is shown in Table 3.

Section	Title[5]	Abstract	Claim	Desc.	Total[6]
#Candidate	7K	16K	57K	276K	352K

Table 3. Numbers of sentence pair candidates

To assess the quality of the sentence alignments, we randomly sampled 1,000 pairs from them. Two Chinese-English bilingual annotators were asked to separately classify them into three categories: *correct*, *partially correct*, and *incorrect*[7]. The *correct* ones are the most

[2] *Literally translated* means most components of English sentences are literally translated from the Chinese ones; *loosely translated* means only some components of the supposedly parallel English and Chinese sentences overlap; *rewritten* means the whole patent is rewritten, and it is difficult to find any parallel sentences.

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

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

[5] The pairs of titles are considered parallel sentence candidates.

[6] The repeated sentence pairs in different sections are counted only once for the total numbers. In Table 7 below, the pairs are processed in the similar way.

[7] *Correct* means the English sentence is exactly the literal translation of the Chinese one, or the content overlap between them are above 80% with no need to consider phrasal reordering during the translation; partially *correct* means the Chinese sentence and the English one are not the literal

valuable resources for MT and other NLP applications, but the *partially correct* ones may also be useful for some NLP applications, such as bilingual term extraction or word alignment. Then we compute the inter-annotator agreement among the two annotators, which is 91.5%, showing the high consistency between our annotators and also the task is well-defined. For the 85 disagreed cases, the two annotators discuss and then resolve the final category for each sentence pair. The final numbers for sentence pairs of *correct*, *partially correct*, and *incorrect* are 448 (44.8%), 114 (11.4%) and 438 (43.8%), respectively.

The above evaluation on the sentence alignments from the noisy parallel patents shows that a large proportion of aligned sentences are incorrect because of noise in patents and in the system. To get truly parallel sentence pairs, filtering out the misaligned sentences is quite necessary; otherwise, they may adversely affect the subsequent NLP applications.

5 Filtering of Sentence Pair Candidates

To filter out incorrect alignments, we sort all sentence pairs based on a scoring metric so as to remove those with lower scores as incorrect alignments. Here we compare and evaluate three individual measures and different ensemble techniques for sentence filtering.

5.1 Filtering Measures and Ensemble Methods

Suppose we are given a sentence pair, namely the Chinese sentence S_c and its English counterpart S_e , and l_c and l_e respectively denote the lengths of S_c and S_e in terms of the number of words. Three kinds of measures for scoring aligned sentences are introduced as follows.

1) The ***length-based score*** P_l *(Len)*: we consider the length ratio between S_c and S_e has a normal distribution with mean μ and variance σ^2 (Gale and Church, 1991). The formula for p_l is as follows:

$$p_l(S_c, S_e) = p_l(l_c / l_e) = 2*(1 - \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz)$$

where $\delta = (l_e - l_c \mu) / \sqrt{l_c \sigma^2}$. The parameters μ and σ^2 are estimated on the preliminary sentence pairs obtained in Sec. 4.

2) The ***dictionary-based score*** P_d : the score 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 S_c and S_e; 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

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

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

Here, to alleviate the coverage problem of the bilingual dictionary, we propose a modified version, *the normalized dictionary-based score (DictN)*, in which l_c and l_e denote the numbers of words occurring in the bilingual dictionary in S_c and S_e respectively.

3) The ***bidirectional translation probability score*** P_t *(Tran)*: it combines the translation probability value of both directions (i.e. Chinese->English and English->Chinese), instead of using only one direction (Moore, 2002; Chen, 2003). 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 $P(S_c \mid S_e)$.

A wide variety of ensemble methods have been used in various fields (Polikar, 2006; Wan, 2008).

translation of each other, but the content of each sentence can cover more than 50% of the other; *incorrect* 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 not translated in the other.

We evaluate the following[8]: 1) Average (*Avg*): the average of the individual scores; 2) Multiplication (*Mul*): the product of the individual scores; 3) Linear Combination (*LinC*): the weighted average by associating each individual score with a weight, indicating the relative confidence in the value; 4) *Filter*: use P_t for sorting, but if P_d or P_t of a sentence pair is lower than a predefined threshold, that pair will be moved to the end of the sorting list. The thresholds can be empirically set based on the data.

5.2 Empirical Evaluation of Sentence Filtering

To assess the performance of individual measures and ensemble methods, the randomly selected 1,000 sentence pairs and their final categories mentioned in Sec. 4 are used as the test data and the gold standard. Each method sorts these 1,000 sentence pairs in descending order according to their corresponding scores given by that method. For the evaluation metrics of each sorted list, we use the 11-point interpolated average precision (*P11*) and MAP (Mean Average Precision) which are commonly used in Information Retrieval. The baseline method does not sort sentence pairs, and its precision is 44.8% if only the 448 correct alignments are considered correct (*case 1*); while its precision is 56.2% if we consider the 448 correct pairs plus 114 partially correct ones correct (*case 2*).

For *DictN*, we use the combined bilingual dictionary mentioned in Sec. 4 to compute the scores. For *Tran*, we use the preliminarily aligned sentences mentioned in Sec. 4 as the training data and compute the word alignment probability score given by the default training process of Giza++ (Och and Ney, 2003), which is based on IBM Model 4 (Brown et al., 1993). The performances for *case 1* and *case 2* are shown in Table 4, from which we can observe:

1) *Len* performs the worst among the three measures although it is much better than the baseline method. The reason is that it alone is not reliable for noisy parallel data because of lack of lexical evidence. The performance of *DictN* is worse than that of the translation probability score because it can not fully cover the large amount of technical terms in patents.

2) *Tran* shows much better performance than the other two measures, which may be explained by the fact that the translation model can leverage the probabilistic information of both lexical and length information, and hence generally performs well. However, *TRAN* tends to be error-prone for the highest ranked sentence pairs. The possible explanation is that the training data itself contain some incorrectly aligned sentences, which lead to some bad parameters in the translation model.

3) All ensemble methods outperform individual measures in terms of P11 and MAP, which shows that each individual measure has its own strength in identifying the correct sentence pairs. Thus fusing the evidence together could improve the performance of the sorted list.

4) *LinC* [9] and *Filter* [10] achieve better performance than *Avg* and *Mul*, showing that we can achieve better performance using some delicate fusing strategies than simply using average or multiplication. *Filter* is shown to be the best among all ensemble methods, which can be explained by the good filtering effects of *Len* and *DictN* for misaligned sentences among the highly ranked sentence pairs in the sorted list of *Tran*.

5.3 Impact of Sentence Filtering on SMT

Although the experiment shows that sentence filtering can help identify really parallel sentences, we may wonder whether the sentence filtering actually leads to better SMT performance. Therefore, we evaluated the impact of sentence filtering on SMT. The Moses toolkit (Koehn et al., 2007) was used to conduct Chinese->English SMT experiments and BLEU and NIST scores are used as the evaluation metrics. We followed the instruction of the baseline system for the shared task in the 2008 ACL workshop on SMT.

[8] Before the ensemble of individual scores, we first need to normalize the scores into the range between 0 and 1 according to their distributions: the length-based and dictionary-based scores are already within the range; the translation score roughly follows a linear distribution.

[9] The weights for *Tran*, *Len*, *DictN* are 99, 30 and 16, respectively. They are got by the exhaustive searching of each weight within the integer range of 0-100 for the best performance.

[10] Here we set the un-normalized thresholds of *Len* and *DictN* to 0.25 and 0.0075, respectively.

Measures & Ensemble Methods		Case 1		Case 2	
		P11 (%)	MAP (%)	P11 (%)	MAP (%)
Baseline		44.8	44.8	56.2	56.2
Individual	Len	70	68.5	79.0	77.8
	DictN	73.9	71.8	82.9	83.1
	Tran	85.1	84.3	89.0	88.7
Ensemble	Avg	**89.2**	89.7	92.7	94.7
	Mul	88.0	**89.8**	**92.9**	**95.0**
	LinC	**91.5**	**92.2**	**93.4**	**95.5**
	Filter	**92.0**	**93.4**	**94.7**	**96.6**

Table 4. Performance of sentence filtering

The 352K sentence pair candidates were divided into the training and test data sets following the scenario in (Fujii et al. 2008). Since the most recent English patents in our data were filed in 2008, we used those filed in 2008 in *USPTO* to produce the test data consisting of about 35K sentence pair candidates, and other patents filed before 2008 to produce the training data, which consists of about 320K Chinese-English sentence pair candidates.

All the sentence pairs were sorted using the *Filter* ensemble method combining the three measures mentioned in 5.2. We chose the top ranked 2000 Chinese-English sentence pairs in the test data as the test set, and compared SMT performance by using different percentages of the sorted sentence pair candidates in the training data to get the translation model. The results are shown in Figure 1 and Figure 2. We observed that:

1) The BLEU and NIST scores for the highest ranking 10%-90% of the training data are higher than those of 100%. Even when we only use the highest ranking 10% of the training data, we can get better BLEU and NIST scores than using the highest ranking 80%, 90% or 100%. This shows that sentence filtering can identify really parallel sentences, which in turn improve SMT performance.

2) Performance peaks for the highest ranking 30% and 60% of the training data in terms of BLEU and NIST scores show that filtering out too many or too few sentence pair candidates cannot get the best performance. Performance is worst at 5% of the training data, demonstrating that a training corpus of very small size cannot achieve good performance for SMT.

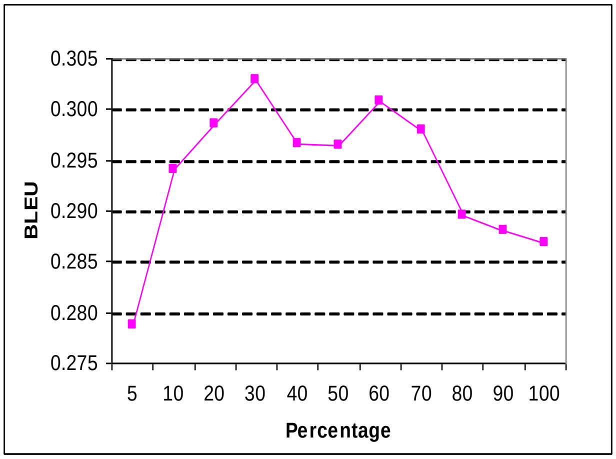

Figure 1. BLEU scores for percentages

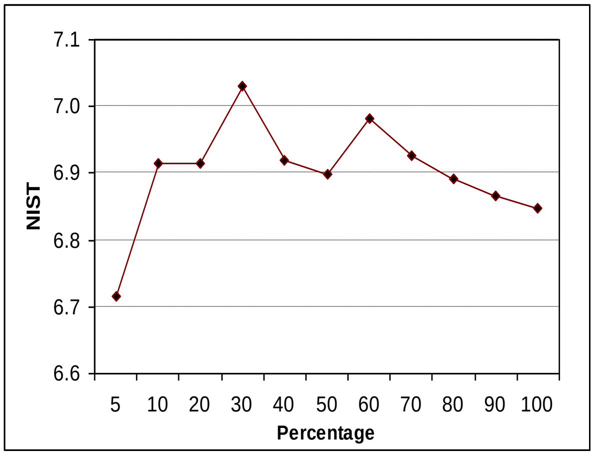

Figure 2. NIST scores for percentages

6 The Final Patent Parallel Corpus

To generate the final corpus of truly parallel sentences, we first evaluated the precision of the 352K sentence pair candidates by sorting them in descending order using the ensemble method *Filter*. We randomly selected 100 samples from each of the 12 blocks ranked at the top 240,000 *Filter*. We randomly selected 100 samples from each of the 12 blocks ranked at the top 240,000 sentence pairs (each block has 20,000 pairs). An annotator classified them into *correct (Cor)*, *partially orrect (PaC)*, and *incorrect (IC)* just as in Sec. 4. The results of evaluation are given in Table 5.

Range	#Cor	#PaC	#IC
1 -	98	1	1
20001 -	98	0	2
40001 -	96	2	2
60001 -	91	5	4
80001 -	92	2	6
100001 -	88	1	11
120001 -	77	6	17
140001 -	73	7	20
160001 -	64	7	29
180001 -	37	7	56
200001 -	34	6	60
220001 -	32	8	60
Total	880	52	268

Table 5. Rank vs judgement

The table shows that the number of IC's increases rapidly as the rank increases. This demonstrates that the ensemble method *Filter* can differentiate the correct alignments from the incorrect ones. Then, we choose the top 160K alignments as the final parallel corpus, in which the average precision of correct and partially correct sentences is about 90.0% based on the samples above. We give some basic statistics of the corpus in Table 6.

#Patents	#Sentence Pairs	#Word Tokens		#Word Types	
		EN	CN	EN	CN
7K	160K	4,168K	4,130K	46K	44K

Table 6. Basic statistics of the final parallel corpus

We also compared the sentence pair candidate numbers among different sections in the final corpus. The result in Table 7 shows that the title and claims sections have two highest precisions: 74.4% and 64.8% respectively; while the abstract and description sections show lower precisions: 45.2% and 40.9% respectively. This shows that it is more difficult to find parallel sentences in the description or abstract section than in the title or claim sections, and that a large proportion of the patent titles are parallel.

Section	Title	Abstr.	Claims	Desc.	Total
#Candidates	7,029	15,755	56,667	275,737	352K
#Final Pairs	5,232	7,119	36,722	112,812	160K
Selected (%)	**74.4**	45.2	**64.8**	40.9	45.4

Table 7. Selected percentages of different sections

7 Conclusion and Future Work

In this paper, we gave an account of the construction of a parallel Chinese-English patent sentence corpus built from noisy parallel Chinese-English patents. We first compared three publicly available sentence aligners and chose one to align sentences in noisy parallel patents. To filter out those incorrect alignments, we compared and examined individual measures and different ensemble methods. The experiments showed that the combinations of measures outperform the individual measures, and filtering out low-quality misaligned sentence pairs can improve SMT performance.

The final Chinese-English patent parallel corpus consists of 160K sentence pairs with the overall precision of about 90%. Given the relative paucity of patent parallel data for SMT, this corpus will be a helpful first step towards MT research and other cross-lingual information access applications in the patent domain. This includes bilingual term extraction and cross-lingual information retrieval, which will be examined in future.

Acknowledgments

We wish to thank our colleagues Mr. Kapo Chow and Mr. Yuwing Chan for their help, Matrixware Information Services GmbH in Vienna for the provision of most of the English patents, Dr. Xiaojun Wan of Peking University for his invaluable comments, as well as the anonymous reviewers for constructive comments. Research reported here is undertaken in the Corpus Linguistics Laboratory of the Language

Information Sciences Research Centre of City University of Hong Kong.

References

Peter F. Brown, Jennifer C. Lai and Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. *Proceedings of ACL*, pp.169-176.

Peter F. Brown, 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.

Stanley F. Chen. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. *Proceedings of ACL*, pp. 9-16. Columbus, OH.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR)*. pp. 389-400. Tokyo, Japan.

William A. Gale and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of ACL*. pp.79-85.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, and et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL, demonstration session, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit X*.

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

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

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

Robi Polikar. 2006. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, 6(3). pp. 21-45.

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

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

Philip Resnik and Noah A. Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Benjamin K. Tsou and Bin LU. Automotive Patents from Mainland China and Taiwan: A Preliminary Exploration of Terminological Differentiation and Content Convergence. *World Patent Information*. (to appear)

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. *Proceedings of ACL*, pp. 72–79.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. *MT Summit XI*, pp. 475–482.

Daniel Varga, Peter Halacsy, and et al. 2005. Parallel Corpora for Medium Density Languages. *RANLP 2005 Conference*.

Xiaojun Wan. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceeding of EMNLP2008. pp. 553-561.

Dekai Wu and Pascale Fung. 2005. Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora. *IJCNLP2005*.

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