2 Mining Patents for Parallel Corpora

Masao Utiyama
Hitoshi Isahara

Large-scale parallel corpora are indispensable language resources for machine translation. However, only a few large-scale parallel corpora are available to the public. We found that a large amount of parallel texts can be obtained by mining comparable patent corpora. This is because patents of the same subject matter are often filed in multiple countries. Such patents are called "patent families." We describe a Japanese-English patent parallel corpus created from patent families filed in Japan and the United States. The parallel corpus contains about 2 million sentence pairs that were aligned automatically. This is the largest Japanese-English parallel corpus and will be available to the public after the NTCIR-7 workshop meeting. We estimated that about 97% of the sentence pairs were correct alignments and about 90% of the alignments were adequate translations whose English sentences reflected almost perfectly the contents of the corresponding Japanese sentences.

2.1 Introduction

The rapid and steady progress in corpus-based machine translation (MT) (Nagao, 1981; Brown et al., 1993) has been supported by large parallel corpora, such as the Arabic-English and Chinese-English parallel corpora distributed by the Linguistic Data Consortium (Ma and Cieri, 2006), the Europarl corpus (Koehn, 2005) consisting of 11 European languages, and the JRC-Acquis corpus consisting of more than 20 European languages (Steinberger et al., 2006). However, large parallel corpora do not exist for many language pairs. For example, there are a few publicly available Japanese-English parallel corpora as listed in the website of the International Workshop on Spoken Language Translation (IWSLT-2007)[1] and these corpora are small compared to the above-mentioned corpora.

1. http://iwslt07.itc.it/menu/resources.html

Much work has been undertaken to overcome this lack of parallel corpora. For example, Resnik and Smith (2003) have proposed mining the web to collect parallel corpora for low-density language pairs. Munteanu and Marcu (2005) have extracted parallel sentences from large Chinese, Arabic, and English nonparallel newspaper corpora. Utiyama and Isahara (2003) have extracted Japanese-English parallel sentences from a noisy-parallel corpus.

In this chapter, we show that a large amount of parallel text can be obtained by mining comparable patent corpora. This is because patents of the same subject matter are often filed in multiple countries. Such patents are called *patent families*. For example, we obtained over 80,000 patent families from patents submitted to the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO), as described in section 2.3. From these patent families, we extracted a high-quality Japanese-English parallel corpus. This corpus and its extension will be used in the NTCIR-7 patent MT task and made available to the public after the NTCIR-7 workshop meeting, which will be held in December 2008.[2] In addition, we believe that multilingual parallel corpora for other languages could be obtained by mining patent families because patents are filed in multiple countries.

Patent translations are required in the society. For example, the JPO provides Japanese-English MT of Japanese patent applications. Consequently, it is important to collect parallel texts in the patent domain to promote corpus-based MT on that domain.

In section 2.2, we review the related work on comparable corpora. In section 2.3, we describe the resources used to develop our patent parallel corpus. In sections 2.4, 2.5, and 2.6, we describe the alignment procedure, the basic statistics of the patent parallel corpus, and the MT experiments conducted on the patent corpus.

2.2 Related Work

Comparable corpora have been important language resources for multilingual natural language processing. They have been used in mining bilingual lexicons (Fung and Yee, 1998; Rapp, 1999; Higuchi et al., 2001), parallel sentences (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Fung and Cheung, 2004a,b), and parallel subsentential fragments (Munteanu and Marcu, 2006; Quirk et al., 2007).

Fung and Yee (1998) and Rapp (1999) have used newspaper corpora to extract bilingual lexicons. Higuchi et al. (2001) have used patent families filed in both Japan and the United States to extract bilingual lexicons. They used only the title and abstract fields from a number of fields (e.g., titles, abstracts, claims, and so on) in patent documents. This is because the title and abstract fields are often parallel in Japanese and English patents, even though the structures of paired patents are not

2. http://research.nii.ac.jp/ntcir/

always the same, e.g., the number of fields claimed in a single patent family often varies depending on the language.

Higuchi et al. (2001) have shown that the title and abstract fields in patent families are useful for mining bilingual lexicons. However, the number of sentences contained in these two fields is small compared to the overall number of sentences in the whole patents. Thus, using only these two fields does not provide enough sentences for a parallel corpus. In this chapter, we show that a large amount of parallel texts can be obtained by mining the "Detailed Description of the Preferred Embodiments" part and the "Background of the Invention" part of patent families.[3] Of the patent families examined, we found that these parts tend to be literal translations of each other, even though they usually contain noisy alignments.

Traditional sentence alignment algorithms (Gale and Church, 1993; Utsuro et al., 1994) are designed to align sentences in clean-parallel corpora and operate on the assumption that there is little noise such as reorderings, insertions, and deletions between the two renderings of a parallel document. However, this assumption does not hold for comparable or noisy-parallel corpora. In our case, for example, some information described in a Japanese patent may not be included when it is submitted to the USPTO. As a result, the patent family consisting of the original Japanese patent and the modified United States patent will contain missing text when compared.

To tackle noise in comparable corpora, Zhao and Vogel (2002) and Utiyama and Isahara (2003) first identify similar parallel texts from two corpora in different languages. They then align sentences in each text pair. Finally, they extract high-scoring sentence alignments assuming that these are cleaner than the other sentence alignments.

Zhao and Vogel (2002) and Utiyama and Isahara (2003) assume that their corpora are noisy-parallel. That is, they assume that document pairs identified by their systems are rough translations of each other. In contrast, Fung and Cheung (2004a,b) and Munteanu and Marcu (2005) do not assume document-level translations. They judge each sentence pair in isolation to decide whether those sentences are translations of each other. Consequently, they do not need document pairs being translations of each other. Munteanu and Marcu (2006) and Quirk et al. (2007) go even further. They do not assume sentence-level translations and try to extract bilingual sentential fragments (e.g., phrases) from nonparallel corpora.

In this chapter, we use Utiyama and Isahara's method (Utiyama and Isahara, 2003) to extract sentence alignments from patent families because we have found that patent families are indeed rough translations of each other.

3. We will provide additional parallel texts obtained from other fields (e.g., claims and abstracts) for the NTCIR-7 patent MT task.

2.3 Resources

Our patent parallel corpus was constructed using patent data provided for the NTCIR-6 patent retrieval task (Fujii et al., 2007). The patent data consists of

- unexamined Japanese patent applications published from 1993 to 2002, and
- USPTO patents published from 1993 to 2000.

The Japanese patent data consists of about 3.5 million documents, and the English data consists of about 1.0 million documents.

We identified 84,677 USPTO patents that originated from Japanese patents. We used the priority information described in the USPTO patents to obtain these patent pairs (families). We examined these patent families and found that the "Detailed Description of the Preferred Embodiments" part (*embodiment part* for short) and the "Background of the Invention" part (*background part* for short) of each application tend to be literal translations of each other. We thus decided to use these parts to construct our patent parallel corpus.

We used simple pattern-matching programs to extract the embodiment and background parts from the whole document pairs and obtained 77,014 embodiment part pairs and 72,589 background part pairs. We then applied the alignment procedure described in section 2.4 to these 149,603 pairs. We call these embodiment and background parts *documents*.

2.4 Alignment Procedure

2.4.1 Score of Sentence Alignment

We used Utiyama and Isahara's method (Utiyama and Isahara, 2003) to score sentence alignments. We first aligned sentences[4] in each document by using a standard dynamic programming (DP) matching method (Gale and Church, 1993; Utsuro et al., 1994). We allowed one-to-n, n-to-one ($0 \leq n \leq 5$), or two-to-two alignments when aligning the sentences. A concise description of the algorithm used is given elsewhere (Utsuro et al., 1994).[5] Here, we only discuss the similarities between Japanese and English sentences used to calculate scores of sentence alignments.

4. We split the Japanese documents into sentences by using simple heuristics and split the English documents into sentences by using a maximum entropy sentence splitter available at `http://www2.nict.go.jp/x/x161/members/mutiyama/maxent-misc.html` We manually prepared about 12,000 English patent sentences to train this sentence splitter. The precision of the splitter was over 99% for our test set.

5. The sentence alignment program we used is available at `http://www2.nict.go.jp/x/x161/members/mutiyama/software.html`

Let J_i and E_i be the word tokens of the Japanese and English sentences for the ith alignment. The similarity between J_i and E_i is[6]

$$\mathrm{SIM}(J_i, E_i) = \frac{2 \times \sum_{j \in J_i} \sum_{e \in E_i} \frac{\delta(j,e)}{\deg(j)\deg(e)}}{|J_i| + |E_i|}, \tag{2.1}$$

where j and e are word tokens and

$|J_i|$ is the number of Japanese word tokens in the ith alignment
$|E_i|$ is the number of English word tokens in the ith alignment
$\delta(j,e) = 1$ if j and e can be a translation pair, 0 otherwise
$\deg(j) = \sum_{e \in E_i} \delta(j,e)$
$\deg(e) = \sum_{j \in J_i} \delta(j,e)$

Note that $\frac{\delta(j,e)}{\deg(j)\deg(e)} = 0$ if $\delta(j,e) = \deg(j) = \deg(e) = 0$.

J_i and E_i were obtained as follows: We used ChaSen[7] to morphologically analyze the Japanese sentences and extract content words, which consisted of J_i. We used a maximum entropy tagger[8] to part-of-speech tag the English sentences and extract content words. We also used WordNet's library[9] to obtain lemmas of the words, which consisted of E_i. To calculate $\delta(j,e)$, we looked up an English-Japanese dictionary that was created by combining entries from the EDR Japanese-English bilingual dictionary, the EDR English-Japanese bilingual dictionary, the EDR Japanese-English bilingual dictionary of technical terms, and the EDR English-Japanese bilingual dictionary of technical terms.[10] The combined dictionary contained over 450,000 entries.

After obtaining the maximum similarity sentence alignments using DP matching, we calculated the similarity between a Japanese document, J, and an English document, E, (AVSIM(J, E)), as defined by Utiyama and Isahara (2003), using

$$\mathrm{AVSIM}(J, E) = \frac{\sum_{i=1}^{m} \mathrm{SIM}(J_i, E_i)}{m}, \tag{2.2}$$

where $(J_1, E_1), (J_2, E_2), \ldots (J_m, E_m)$ are the sentence alignments obtained using DP matching. A high AVSIM(J, E) value occurs when the sentence alignments in J and E take high similarity values. Thus, AVSIM(J, E) measures the similarity between J and E.

---

6. To penalize one-to-0 and 0-to-one alignments, we assigned $\mathrm{SIM}(J_i, E_i) = -1$ to these alignments instead of the similarity obtained by using Eq. (2.1).
7. `http://chasen-legacy.sourceforge.jp/`
8. `http://www2.nict.go.jp/x/x161/members/mutiyama/maxent-misc.html`
9. `http://wordnet.princeton.edu/`
10. `http://www2.nict.go.jp/r/r312/EDR/`

We also calculated the ratio of the number of sentences between J and E ($R(J,E)$) using

$$R(J,E) = \min(\frac{|J|}{|E|}, \frac{|E|}{|J|}), \tag{2.3}$$

where $|J|$ is the number of sentences in J, and $|E|$ is the number of sentences in E. A high $R(J,E)$-value occurs when $|J| \sim |E|$. Consequently, $R(J,E)$ can be used to measure the literalness of translation between J and E in terms of the ratio of the number of sentences.

Finally, we defined the score of alignment J_i and E_i as

$$\text{Score}(J_i, E_i) = \text{SIM}(J_i, E_i) \times \text{AVSIM}(J,E) \times R(J,E). \tag{2.4}$$

A high $\text{Score}(J_i, E_i)$ value occurs when

- sentences J_i and E_i are similar,
- documents J and E are similar,
- numbers of sentences $|J|$ and $|E|$ are similar.

$\text{Score}(J_i, E_i)$ combines both sentence and document similarities to discriminate between correct and incorrect alignments.

We use only high scoring sentence alignments to extract valid sentence alignments from noisy sentence alignments. We use Score in Eq. (2.4) as the score for a sentence alignment as described above because a variant of Score has been shown to be more appropriate than SIM for discriminating between correct and incorrect alignments (Utiyama and Isahara, 2003). When we compare the validity of two sentence alignments in the same document pair, the rank order of sentence alignments obtained by applying Score is the same as that of SIM because these alignments share common AVSIM and R. However, when we compare the validity of two sentence alignments in different document pairs, Score prefers the sentence alignment in the more similar (high AVSIM $\times$ R) document pair even if their SIM has the same value, while SIM cannot discriminate between the validity of two sentence alignments if their SIM has the same value. Therefore, Score is more appropriate than SIM when comparing sentence alignments in different document pairs because, in general, a sentence alignment in a similar document pair is more reliable than one in a dissimilar document pair.[11]

11. However, as pointed out by a reviewer, a document J which is a subset of E (or vice versa) could have a very low similarity score, $\text{AVSIM}(J,E)$. As a result, Eq. (2.4) could get a very low $\text{Score}(J_i, E_i)$, even for a pair of sentences that represent a perfect reciprocal translation. Therefore, if such cases exist in our patent corpus, many good sentence pairs may be lost. We have not yet investigated the amount of such cases in our corpus. We leave it for future work.

2.4.2 Noise Reduction in Sentence Alignments

We used the following procedure to reduce noise in the sentence alignments obtained by using the previously described aligning method on the 149,603 document pairs.

The number of sentence alignments obtained was about 7 million. From these alignments, we extracted only one-to-one sentence alignments because this type of alignment is the most important category for sentence alignment. As a result, about 4.2 million one-to-one sentence alignments were extracted. We sorted these alignments in decreasing order of scores and removed alignments whose Japanese sentences did not end with periods to reduce alignment pairs considered as noise. We also removed all but one of the identical alignments. Two individual alignments were determined to be identical if they contained the same Japanese and English sentences. Consequently, the number of alignments obtained was about 3.9 million.

We examined 20 sentence alignments ranked between 1,999,981 and 2,000,000 from the 3.9 million alignments to determine if they were accurate enough to be included in a parallel corpus. We found that 17 of the 20 alignments were almost literal translations of each other and 2 of the 20 alignments had more than 50% overlap in their contents. We also examined 20 sentence alignments ranked between 2,499,981 and 2,500,000 and found that 13 of the 20 alignments were almost literal translations of each other and 6 of the 20 alignments had more than 50% overlap. Based on these observations, we decided to extract the top 2 million one-to-one sentence alignments. Finally, we removed some sentence pairs from these top 2 million alignments that were too long (more than 100 words in either sentence) or too imbalanced (when the length of the longer sentence is more than five times the length of the shorter sentence). The number of sentence alignments thus obtained was 1,988,732. We call these 1,988,732 sentence alignments the ALL data set (*ALL* for short) in this chapter.

We also asked a translation agency to check the validity of 1000 sentence alignments randomly extracted from ALL. The agency conducted a two-step procedure to verify the data. In the first step, they marked a sentence alignment as

- A if the Japanese and English sentences matched as a whole,
- B if these sentences had more than 50% overlap in their contents,
- C otherwise,

to check if the alignment was correct. The number of alignments marked as A was 973, B was 24, and C was 3. In the second step, they marked an alignment as

- A if the English sentence reflected almost perfectly the contents of the Japanese sentence,
- B if about 80% of the contents were shared,
- C if less than 80% of the contents were shared,
- X if they could not determine the alignment as A, B, or C,

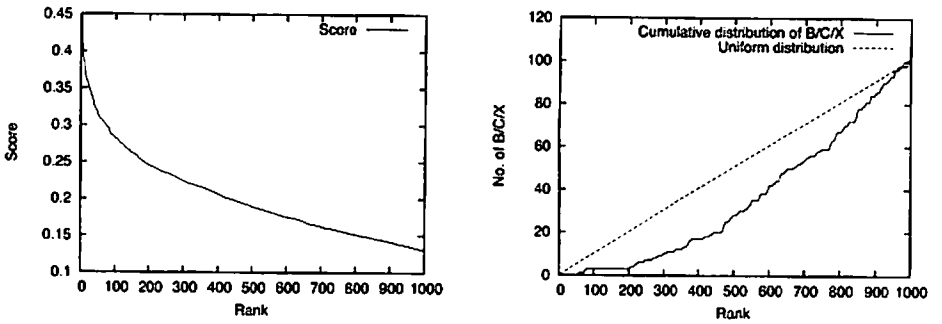

Figure 2.1 Distributions of scores and noisy alignments.

to check if the alignment was an adequate translation pair. The number of alignments marked as A was 899, B was 72, C was 26, and X was 3. Based on these evaluations, we concluded that the sentence alignments in ALL are useful for training and testing MT systems.

Next, we used these 1000 sentence alignments to investigate the relationship between the human judgments and Score given in Eq. (2.4). Figure 2.1 shows the distributions of scores and noisy alignments (marked as B, C, or X in the second step) against the ranks of sentence alignments ordered by using Score. The left figure shows that scores initially decreased rapidly for higher-ranking alignments, and then decreased gradually. The right figure shows the cumulative number of noisy alignments. The solid line indicates that noisy alignments tend to have low ranks. Note that if noisy alignments are spread uniformly among the ranks, then the cumulative number of noisy alignments follows the diagonal line. Based on the results shown in this figure, we concluded that Score ranked the sentence alignments appropriately.

2.5 Statistics of the Patent Parallel Corpus

2.5.1 Comparison of ALL and Source Data Sets

We compared the statistics of ALL with those of the source patents and sentences from which ALL was extracted to see how ALL represented the sources.

To achieve this, we used the primary international patent classification (IPC) code assigned to each USPTO patent. The IPC is a hierarchical patent classification system and consists of eight sections, ranging from A to H. We used sections G (physics), H (electricity), and B (performing operations; transporting) because these sections had larger numbers of patents than other sections. We categorized patents as O (other) if they were not included in these three sections.

As described in section 2.3, 84,677 patent pairs were extracted from the original patent data. These pairs were classified into G, H, B, or O, as listed in the Source

Table 2.1 Number of patents

IPC	ALL (%)	Source (%)
G	19,340 (37.9)	28,849 (34.1)
H	16,145 (31.6)	24,270 (28.7)
B	7,287 (14.3)	13,418 (15.8)
O	8,287 (16.2)	18,140 (21.4)
Total	51,059 (100.0)	84,677 (100.0)

Table 2.2 Number of sentence alignments

IPC	ALL (%)	Source (%)
G	946,872 (47.6)	1,813,078 (43.4)
H	624,406 (31.4)	1,269,608 (30.4)
B	204,846 (10.3)	536,007 (12.8)
O	212,608 (10.7)	559,519 (13.4)
Total	1,988,732 (100.0)	4,178,212 (100.0)

column of table 2.1. We counted the number of patents included in each section of ALL. We regarded a patent to be included in ALL when some sentence pairs in that patent were included in ALL. The number of such patents are listed in the ALL column of table 2.1. Table 2.1 shows that about 60% ($100 \times 51059/84677$) of the source patent pairs were included in ALL. It also shows that the distributions of patents with respect to the IPC code were similar between ALL and Source.

Table 2.2 lists the number of one-to-one sentence alignments in ALL and Source, where Source means the about 4.2 million one-to-one sentence alignments described in section 2.4.2. The results in this table show that about 47.6 % ($100 \times 1988732/4178212$) sentence alignments were included in ALL. The results also show that the distribution of the sentence alignments are similar between ALL and Source.

Based on these observations, we concluded that ALL represented Source well.

In the following, we use G, H, B, and O to denote the data in ALL whose IPCs were G, H, B, and O.

2.5.2 Basic Statistics

We measured the basic statistics of G, H, B, O, and ALL.

We first randomly divided patents from each of G, H, B, and O into training (TRAIN), development (DEV), development test (DEVTEST), and test (TEST) data sets. One unit of sampling was a single patent. That is, G, H, B, and O consisted of 19,340, 16,145, 7287, and 8287 patents (See table 2.1), respectively, and the patents from each group were divided into TRAIN, DEV, DEVTEST, and

Table 2.3 Number of patents

	TRAIN	DEV	DEVTEST	TEST	Total
G	17,524	630	610	576	19,340
H	14,683	487	493	482	16,145
B	6,642	201	226	218	7,287
O	7,515	262	246	264	8,287
ALL	46,364	1,580	1,575	1,540	51,059

Table 2.4 Number of sentence alignments

	TRAIN	DEV	DEVTEST	TEST	Total
G	854,136	33,133	27,505	32,098	946,872
H	566,458	20,125	19,784	18,039	624,406
B	185,778	6,239	6,865	5,964	204,846
O	193,320	6,232	6,437	6,619	212,608
ALL	1,799,692	65,729	6,0591	6,2720	1,988,732

TEST. We assigned 91% of the patents to TRAIN, and 3% of the patents to each of DEV, DEVTEST, and TEST. We merged the TRAIN, DEV, DEVTEST, and TEST of G, H, B, and O to create those of ALL. Table 2.3 lists the number of patents in these data sets and table 2.4 lists the number of sentence alignments.

2.5.3 Statistics Pertaining to MT

We measured some statistics pertaining to MT. We first measured the distribution of sentence length (in words) in ALL. The mode of the length (number of words) was 23 for the English sentences and was 27 for the Japanese sentences. (We used ChaSen to segment Japanese sentences into words.) Figure 2.2 shows the percentage of sentences for English (en) and Japanese (ja) with respect to their lengths. This figure shows that the peaks of the distributions were not sharp and that there were many long sentences in ALL. This suggests that patents contain many long sentences that are generally difficult to translate.

We then measured the vocabulary coverage. Table 2.5 lists the coverage for the types (distinct words) and tokens (running words) in each TEST section using the vocabulary in the corresponding TRAIN section for the English and Japanese data sets. These tables show that the percentages of types in TEST covered by the vocabulary in TRAIN were relatively low for both English and Japanese. However, the coverage of tokens was quite high. This suggests that patents are not so difficult to translate in terms of token coverage.

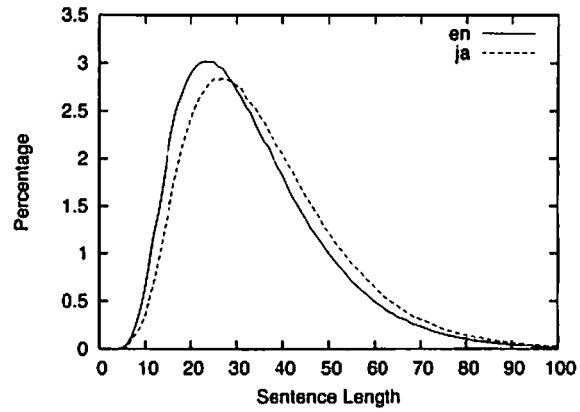

Figure 2.2 Sentence length distribution.

Table 2.5 Percentage of words in test sentences covered by training vocabulary.

(a) Coverage for English

	Type	Token
G	84.37	99.40
H	86.63	99.37
B	90.28	99.38
O	89.19	99.31
ALL	83.36	99.55

(b) Coverage for Japanese

	Type	Token
G	90.27	99.69
H	91.97	99.67
B	94.12	99.65
O	92.50	99.48
ALL	89.85	99.77

2.6 MT Experiments

2.6.1 MT System

We used the baseline system for the shared task of the 2006 NAACL/HLT workshop on statistical machine translation (Koehn and Monz, 2006) to conduct MT experiments on our patent corpus. The baseline system consisted of the Pharaoh decoder (Koehn, 2004a), SRILM (Stolcke, 2002), GIZA++ (Och and Ney, 2003), mkcls (Och, 1999), Carmel,[12] and a phrase model training code.

We followed the instructions of the shared task baseline system to train our MT systems.[13] We used the phrase model training code of the baseline system to extract phrases from TRAIN. We used the trigram language models made from TRAIN. To

12. `http://www.isi.edu/licensed-sw/carmel/`
13. The parameters for the Pharaoh decoder were "-b 0.00001 -s 100." The maximum phrase length was 7. The "grow-diag-final" method was used to extract phrases.

Table 2.6 Comparing reordering limits. Each MT system was trained on each of the G, H, B, O, and ALL TRAIN data sets, tuned for both reordering limits using each DEV data set, and applied to 1000 randomly sampled sentences extracted from each DEVTEST data set to calculate the %BLEU scores listed in these tables. The source and target languages were English-Japanese for (a) and Japanese-English for (b).

(a) English-Japanese

	no limit	limit=4
G	23.56	22.55
H	24.62	24.14
B	22.62	20.88
O	23.87	21.84
ALL	24.98	23.37

(b) Japanese-English

	no limit	limit=4
G	21.82	21.6
H	23.87	22.62
B	21.95	20.79
O	23.41	22.53
ALL	23.15	21.55

tune our MT systems, we did minimum error-rate training[14] (Och, 2003) on 1000 randomly extracted sentences from DEV using BLEU (Papineni et al., 2002) as the objective function. Our evaluation metric was %BLEU scores.[15] We tokenized and lowercased the TRAIN, DEV, DEVTEST, and TEST data sets. We conducted three MT experiments to investigate the characteristics of our patent corpus.

2.6.2 Comparing Reordering Limits

For the first experiment, we translated 1000 randomly sampled sentences in each DEVTEST data set to compare different reordering limits,[16] because Koehn et al. (2005) have reported that large reordering limits provide better performance for Japanese-English translations. We compared a reordering limit of 4 with no limitation. The results of table 2.6 show that the %BLEU scores for no limitation consistently outperformed those for limit=4. These results coincide with those of Koehn et al. (2005). Based on this experiment, we used no reordering limit in the following experiments.

2.6.3 Cross-Section MT Experiments

For the second experiment, we conducted cross-section MT experiments. The results are shown in tables 2.7 and 2.8. For example, as listed in *table 2.7*, when we used section G as TRAIN and used section H as TEST, we got a %BLEU score of 23.51 for English-Japanese translations, whose relative %BLEU score was 0.87 (=23.51/26.88) of the largest %BLEU score obtained when using ALL as TRAIN. In this case, we used all sentences in TRAIN of G to extract phrases and make

14. The minimum error-rate training code we used is available at `http://www2.nict.go.jp/x/x161/members/mutiyama/software.html`
15. %BLEU score is defined as BLEU × 100.
16. The parameter "-dl" for the Pharaoh decoder.

Table 2.7 %BLEU scores (relative %BLEU scores) for cross-section MT experiments (English-Japanese)

TEST: TRAIN	G	H	B	O	ALL
G	**25.89 (0.97)**	23.51 (0.87)	20.19 (0.82)	18.96 (0.76)	23.93 (0.91)
H	22.19 (0.83)	**25.81 (0.96)**	19.16 (0.78)	18.68 (0.75)	22.57 (0.86)
B	18.17 (0.68)	18.92 (0.70)	**22.54 (0.92)**	19.25 (0.77)	18.97 (0.72)
O	16.93 (0.63)	18.45 (0.69)	18.22 (0.74)	**24.15 (0.97)**	18.32 (0.70)
ALL	26.67 (1.00)	26.88 (1.00)	24.56 (1.00)	24.98 (1.00)	26.34 (1.00)

Table 2.8 %BLEU scores (relative %BLEU scores) for cross-section MT experiments (Japanese-English)

TEST: TRAIN	G	H	B	O	ALL
G	**24.06 (0.98)**	22.18 (0.90)	19.40 (0.85)	19.33 (0.80)	22.59(0.93)
H	20.91 (0.85)	**23.74 (0.97)**	18.11 (0.79)	18.60 (0.77)	21.28(0.88)
B	17.64 (0.72)	17.94 (0.73)	**21.92 (0.96)**	19.58 (0.81)	18.39(0.76)
O	17.50 (0.72)	18.43 (0.75)	18.57 (0.81)	**24.27 (1.00)**	18.67(0.77)
ALL	24.47 (1.00)	24.52 (1.00)	22.94 (1.00)	24.04 (0.99)	24.29(1.00)

a trigram language model. We used 1000 randomly sampled sentences in DEV of section G to tune our MT system. We used all sentences in TEST of section H to calculate %BLEU scores (see table 2.4 for the number of sentences in each section of TRAIN and TEST).

The results in these tables show that MT systems performed the best when the training and test sections were the same.[17]

These results suggest that patents in the same section are similar to each other, while patents in different sections are dissimilar. Consequently, we need domain adaptation when we apply our trained MT system to a section that is different from that on which it has been trained. However, as shown in the ALL rows, when we used all available training sentences, we obtained the highest %BLEU scores for all but one case. This suggests that if we have enough data to cover all sections we can achieve good performance for all sections.

Tables 2.7 and 2.8 show that both the domain and quantity of training data affect the performance of MT systems. We conducted additional experiments to see the relationship between these two factors. We trained a Japanese-English MT system

17. The results in these tables indicate that %BLEU scores for English-Japanese translations are higher than those for Japanese-English translations. This is because Japanese words are generally shorter than English words. As described in section 2.5.3, the mode of the length for English sentences was 23 and that for Japanese was 27. This suggests that it is easier to reproduce Japanese n-grams, which leads to higher %BLEU scores.

Table 2.9 %BLEU scores for the additional experiments

	B	G	H	O
Same	21.92	24.06	23.74	24.27
ALL\B	20.72	24.39	24.47	23.69
ALL	22.94	24.47	24.52	24.04

on ALL excluding B (ALL\B). We reused the feature weights of the MT system that was trained and tuned on ALL to save tuning time. We used the system to translate the sentences in TEST of sections B, G, H, and O. The %BLEU scores are shown in the ALL\B row of table 2.9. The figures listed in the row labelled Same were the %BLEU scores obtained by applying MT systems trained on each section to that section. The figures in the Same and ALL rows were taken from table 2.8.

Table 2.9 shows that the system trained on ALL outperformed the system trained on ALL\B for all sections. This suggests that it is more important to have more training data. Next, the system trained on O outperformed the systems trained on ALL and ALL\B. In this case, adding data from other domains reduced performance. The systems trained on G and H were outperformed by those trained on ALL and ALL\B. Thus, additional data helped to improve performance in these cases. Finally, the system trained on B outperformed the system trained on ALL\B despite the fact that the number of sentences in ALL\B (1,613,914) was much larger than that in B (185,778). This suggests that it is the domain that matters more than the quantity of training data.

Table 2.10 lists 15 examples of translation obtained from the Japanese-English MT system trained and tested on TRAIN and TEST of ALL. Reference translations are denoted by an R, and MT outputs are denoted by an M. The vertical bars (|) represent the phrase boundaries given by the Pharaoh decoder. These examples were sampled as follows: We first randomly sampled 1000 sentences from TEST of ALL. The correctness and adequacy of the alignment of these sentences were determined by a translation agency, as described in section 2.4.2. We then selected 899 A alignments whose English translation reflected almost perfectly the contents of the corresponding Japanese sentences. Next, we selected short sentences containing less than 21 words (including periods) because the MT outputs of long sentences are generally difficult to interpret. In the end, we had 212 translations. We sorted these 212 translations in decreasing order of *average n-gram precision*[18] and selected five sentences from the top, middle, and bottom of these sorted sentences.[19]

Table 2.10 shows that top examples (1 to 5) were very good translations. These MT translations consisted of long phrases that contributed to the fluency and

18. Average n-gram precision is defined as $\sum_{n=1}^{4} \frac{p_n}{4}$ where p_n is the modified n-gram precision as defined elsewhere (Papineni et al., 2002).
19. We skipped sentences whose MT outputs contained untranslated Japanese words when selecting these 15 sentences.

adequacy of translations. We think that the reason for these good translations is partly due to the fact that patent documents generally contain many repeated expressions. For example, example 2R is often used in patent documents. We also noticed that lcd61 in example 5M was a very specific expression and was unlikely to be repeated in different patent documents, even though it was successfully reused in our MT system to produce 5M. We found a document that contained lcd61 in TRAIN and found that it was written by the same company who wrote a patent in TEST that contained example 5R, even though these two patents were different. These examples show that even long and/or specific expressions are reused in patent documents. We think that this characteristic of patents contributed to the good translations.

The middle and bottom examples (6 to 15) were generally not good translations. These examples adequately translated individual phrases. However, they failed to adequately reorder phrases. This suggests that we need more accurate models for reordering. Thus, our patent corpus will be a good corpus for comparing various reordering models (Koehn et al., 2005; Nagata et al., 2006; Xiong et al., 2006).

2.6.4 Task-Based Evaluation of the Original Alignment Data

For the third experiment, we assessed the quality of the original alignment data in a task-based setting. In section 2.4.2, we selected the first 2 million sentence alignments based on our observations of the quality of the alignment data. In this section, we increase the size of training data to see how MT performance evolves using more data.

We used the 3.9 million one-to-one sentence alignments obtained in section 2.4.2 as our training data. From this data, we removed the alignments contained in the patents in DEV, DEVTEST, or TEST of ALL. We also removed some sentence pairs that were too long or too imbalanced, as discussed in section 2.4.2. We tokenized and lowercased this data. As a result, we obtained 3,510,846 one-to-one sentence alignments sorted by Score in Eq. (2.4).

We conducted controlled Japanese-English and English-Japanese MT experiments using these 3.5 million sentence alignments. We used the common (a) word alignment data, (b) language models, (c) feature weights, and (d) test data. We changed the size of word alignment data when we built phrase tables in the following experiments.

The common settings were obtained as follows. First, (a) we made word alignment data from all sentence alignments using GIZA++. We randomly divided all the sentence alignment data into two halves, applied GIZA++ separately to each half, and combined them to obtain all word alignment data. (b) We made Japanese and English trigram language models from the first 3.5 million sentence alignments. (c) We reused the feature weights of the English-Japanese and Japanese-English MT systems that were trained and tuned on ALL as described in section 2.6.3. (d) We randomly sampled 2000 sentences from TEST of ALL as the test data.

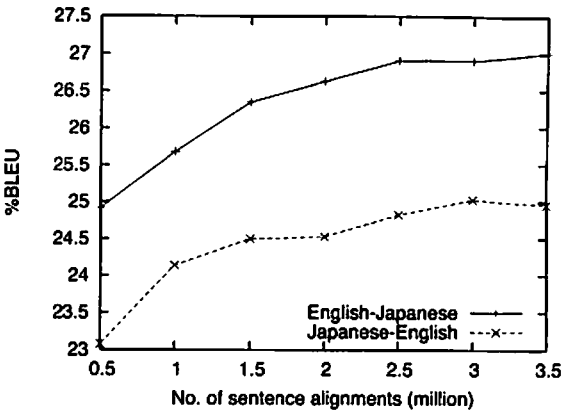

Figure 2.3 Relationship between the %BLEU scores and the number of sentence alignments (in millions).

Finally, we used the first 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5 million sentence alignments to make phrase tables. The %BLEU scores obtained with these phrase tables in the common settings are shown in figure 2.3.

Figure 2.3 shows that the %BLEU scores for the English-Japanese MT experiments reached a plateau around 2.5 million sentence alignments. The %BLEU scores for the Japanese-English experiments increased up to 3.0 million sentence alignments and then dropped when 3.5 million alignments were used as the training data.

These observations indicate that, up to certain points, the increase in the size of the training data offsets the decrease in alignment quality. However, the performance of the MT systems reached a plateau or even decreased after those points due to noise in the alignment data. Therefore, based on the results from these experiments and the results shown in figure 2.1, we conclude that Utiyama and Isahara's method effectively sorted the sentence alignments in decreasing order of their quality.

2.7 Conclusion

Large-scale parallel corpora are indispensable language resources for MT. However, there are only a few publicly available large-scale parallel corpora.

We have developed a Japanese-English patent parallel corpus created from Japanese and U.S. patent data provided for the NTCIR-6 patent retrieval task. We used Utiyama and Isahara's method and extracted about 2 million clean sentence alignments. This is the largest Japanese-English parallel corpus to date. Its size is comparable to other large-scale parallel corpora. This corpus and its extension will

be used in the NTCIR-7 patent MT task and made available to the public after the 7th NTCIR-7 workshop meeting.

We hope that the patent corpus described in this chapter will promote MT research in general and the Japanese-English patent MT research in particular.

Table 2.10 Examples of reference (R) and machine (M) translations

	Top
1R	the printer 200 will now be described .
1M	next , the printer | 200 will now be described . |
2R	preferred embodiments of the present invention will be described hereinbelow with reference to the accompanying drawings .
2M	hereinafter , | preferred embodiments of the | present invention | will be described with reference | to the accompanying drawings | . |
3R	more specifically , variable tr (k) is defined by the following equation .
3M	namely , the variable | tr (| k) | is defined | by the following equation | . |
4R	wd signal is further applied to a command decoder 24 and a data comparator 23 .
4M	further , | signal | wd | is | also applied to a | command decoder 24 | and a data comparator | 23 | . |
5R	at this time , the selected page is displayed on the lcd 61 .
5M	at this time , the | selected | page is | displayed on | the | lcd61 | . |
	Middle
6R	further , reference numbers 201-219 indicate newly-added circuit elements .
6M	further , | reference numerals | 201 to 219 | is newly | added to | the circuit | elements . |
7R	for this purpose , a magnetic head 3 for recording is provided near the disk 1 .
7M	therefore , the | recording | magnetic head | 3 is | provided adjacent to the | disk 1 | . |
8R	accordingly , the energy exerting an influence on the occupant can be reduced .
8M	as a result , the | occupant | on | energy | can be reduced . |
9R	note that nothing is connected to the 1-bit output terminals q0 , q1 of the up-counter 131 .
9M	the | output terminals q0 | , q1 | , the | number | of bits of the | up counter | 131 | is also not connected . |
10R	application program 20 is executed under support of operating system 18 .
10M	an operating system | 20 | of the | support | 18 | under the | application program | is executed . |
	Bottom
11R	numeral 14 denotes a suction surface non-separation streamline , which improves the p-q characteristic and reduces noise .
11M	the | back pressure | , and | no | peeling | surface | 14 | , and | noise | is reduced . | improving characteristics of the | p or q | represents a | stream line |
12R	the use of a robot for deburring work is a known prior art .
12M	deburring | operation | using the | robot | is conventionally | known | technique . |
13R	rdp indicates an address to which a cpu accesses presently .
13M	the cpu | rdp | is currently | being accessed | address | is shown . |
14R	the same is true with regard to the b signal independently of the r signal .
14M	this is | regardless of signals | r | and b signals | similarly . |
15R	the structure of the airbag device 1 will be explained hereinafter .
15M	the air bag apparatus | 1 | are as follows . |