A Japanese-English Patent Parallel Corpus

Masao Utiyama and Hitoshi Isahara

National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun,Kyoto, 619-0289 Japan
E-mail: {mutiyama,isahara}@nict.go.jp

Abstract

We describe a Japanese-English patent parallel corpus created from the Japanese and US patent data provided for the NTCIR-6 patent retrieval task. The corpus contains about 2 million sentence pairs that were aligned automatically. This is the largest Japanese-English parallel corpus, which will be available to the public after the 7th NTCIR 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.

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) and the Europarl corpus (Koehn, 2005) consisting of 11 European languages. However, large parallel corpora do not exist for many language pairs.

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 non-parallel newspaper corpora. Utiyama and Isahara (2003) have extracted Japanese-English parallel sentences from a noisy-parallel corpus.

We have recently aligned Japanese and English sentences in Japanese and US patent data provided for the NTCIR-6 patent retrieval task (Fujii et al., 2007). We used Utiyama and Isahara's method to extract clean sentence alignments. The number of extracted sentence alignments was about 2 million. These sentence pairs and all alignment data that were produced during the alignment procedure are planed to be used in the NTCIR-7 patent MT task.[1] This is the largest Japanese-English parallel corpus, which will be available to the public after the 7th NTCIR workshop meeting.

In Section 2., we describe the resources used to develop the patent parallel corpus. In Sections 3., 4., and 5., we describe the alignment procedure, the basic statistics of the patent parallel corpus, and the MT experiments conducted on the patent corpus.

2. Resources

Our patent parallel corpus was constructed from the patent data provided for the NTCIR-6 patent retrieval task. The patent data consists of

- Unexamined Japanese patent applications published in 1993-2002
- USPTO patent data published in 1993-2000.

The Japanese patent data consists of about 3.5 million documents, and the English data consists of about 1 million documents.

We identified 84,677 USPTO patents that originated from Japanese patents by using the priority information described in the USPTO patents.[2] We examined these 84,677 Japanese and English patent pairs 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 3. to these 149,603 pairs. We call these embodiment and background parts *documents*.

3. Alignment procedure

3.1. Score of sentence alignment

We used Utiyama and Isahara's method (Utiyama and Isahara, 2003) to score sentence alignments. We first aligned sentences[3] in each document by using a standard DP matching method (Gale and Church, 1993; Utsuro et al., 1994). We allowed 1-to-n, n-to-1 ($0 \leq n \leq 5$), or 2-to-2 alignments when aligning the sentences. A concise description of the algorithm used is described elsewhere (Utsuro et

[1] We also plan to extend these alignment data with new patent data.

[2] Some USPTO patents have priority information that identify foreign applications for the same subject matters. Higuchi et al. (2001) have used such corresponding patents filed in both Japan and the United States to extract bilingual lexicons.

[3] 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 testset.

al., 1994).[4] Here, we only discuss the similarities between Japanese and English sentences used to calculate scores of sentence alignments.

Let J_i and E_i be the word tokens of the Japanese and English sentences for i-th alignment. The similarity between J_i and E_i is:[5]

$$\text{SIM}(J_i, E_i) = \frac{2 \times \sum_{j \in J_i} \sum_{e \in E_i} \frac{\delta(j,e)}{\deg(j)\deg(e)}}{|J_i| + |E_i|} \quad (1)$$

where j and e are word tokens and

$|J_i|$ = no. of Japanese word tokens in i-th alignment

$|E_i|$ = no. of English word tokens in i-th alignment

$\delta(j, e) = 1$if j and e can be a translation pair otherwise 0

$\deg(j) = \sum_{e \in E_i} \delta(j, e)$

$\deg(e) = \sum_{j \in J_i} \delta(j, e)$

J_i and E_i were obtained as follows: We used ChaSen[6] to morphologically analyze the Japanese sentences and extract content words, which consisted of J_i. We used a maximum entropy tagger[7] to POS-tag the English sentences and extract content words. We also used WordNet's library[8] 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.[9] The combined dictionary had 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:

$$\text{AVSIM}(J, E) = \frac{\sum_{i=1}^{m} \text{SIM}(J_i, E_i)}{m} \quad (2)$$

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.

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

$$R(J, E) = \min(\frac{|J|}{|E|}, \frac{|E|}{|J|}) \quad (3)$$

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

$$\text{Score}(J_i, E_i) = \text{SIM}(J_i, E_i) \times \text{AVSIM}(J, E) \times R(J, E) \quad (4)$$

A high 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

Score(J_i, E_i) combines both sentence and document similarities to discriminate between correct and incorrect alignments.

3.2. 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 had 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 this top 2 million alignments that were too long (more than 100 words in either sentence) or too imbalanced ($\frac{\text{length of longer sentence}}{\text{length of shorter sentence}} > 5$). 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 paper.

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

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

[5]To penalize 1-to-0 and 0-to-1 alignments, we assigned SIM$(J_i, E_i) = -1$ to these alignments instead of the similarity obtained by using Eq. 1

[6]`http://chasen-legacy.sourceforge.jp/`

[7]`http://www2.nict.go.jp/x/x161/members/mutiyama/maxent-misc.html`

[8]`http://wordnet.princeton.edu/`

[9]`http://www2.nict.go.jp/r/r312/EDR/`

IPC	ALL (%)	Source (%)
G	19340 (37.9)	28849 (34.1)
H	16145 (31.6)	24270 (28.7)
B	7287 (14.3)	13418 (15.8)
O	8287 (16.2)	18140 (21.4)
Total	51059 (100.0)	84677 (100.0)

Table 1: Number of patents

IPC	ALL (%)	Source (%)
G	946872 (47.6)	1813078 (43.4)
H	624406 (31.4)	1269608 (30.4)
B	204846 (10.3)	536007 (12.8)
O	212608 (10.7)	559519 (13.4)
Total	1988732 (100.0)	4178212 (100.0)

Table 2: Number of sentence alignments

	TRAIN	DEV	DEVTEST	TEST	Total
G	17524	630	610	576	19340
H	14683	487	493	482	16145
B	6642	201	226	218	7287
O	7515	262	246	264	8287
ALL	46364	1580	1575	1540	51059

Table 3: Number of patents

	TRAIN	DEV	DEVTEST	TEST	Total
G	854136	33133	27505	32098	946872
H	566458	20125	19784	18039	624406
B	185778	6239	6865	5964	204846
O	193320	6232	6437	6619	212608
ALL	1799692	65729	60591	62720	1988732

Table 4: Number of sentence alignments

- 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.

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.
In the next section, we describe the basic statistics of this patent parallel corpus. In Section 5., we describe the MT experiments conducted on ALL.

4. Statistics of the patent parallel corpus

4.1. 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 consists of eight sections, ranging from A to H. We only used sections G (Physics), H (Electricity) and B (Performing operations; Transporting). We categorized patents as O (Other) if they were not covered by these three sections.

As described in Section 2., 84,677 patent pairs were extracted from the original patent data. These patents were classified into G, H, B or O, as listed in the "Source" column of Table 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 1. Table 1 shows that about 60% ($\frac{51059}{84677} \times 100$) 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 lists the numbers 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 3.2. The results in this table show that about 47.6 % ($\frac{1988732}{4178212} \times 100$) 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 IPC were G, H, B, and O, respectively.

4.2. 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 the sampling was a single patent. That is, G, H, B and O consisted of 19340, 16145, 7287, and 8287 patents (See Table 1), and the patents from each group were divided into TRAIN, DEV, DEVTEST, and TEST. We assigned 91% of the patents to TRAIN, and 3% of the patents to DEV, DEVTEST, and TEST. We merged the TRAIN, DEV, DEVTEST and TEST of G, H, B, and O to create those of ALL. Table 3 lists the number of patents in these data sets and Table 4 lists the number of sentence alignments.
We then counted the number of types (distinct words) and

	TRAIN	DEV	DEVTEST	TEST	Whole
G	124804	18091	16909	17303	132939
H	86127	13915	13149	12975	91620
B	40556	7573	7974	7479	42685
O	47947	7941	7898	8335	50296
ALL	198076	27425	25853	26093	211265

Table 5: Number of types (English)

	TRAIN	DEV	DEVTEST	TEST	Whole
G	77079	14671	13759	13932	81323
H	53307	11077	10740	10602	55899
B	30804	6642	6969	6514	32079
O	36129	6910	6994	7396	37577
ALL	116856	21276	20547	20504	123169

Table 6: Number of types (Japanese)

	TRAIN	DEV	DEVTEST	TEST	Whole
G	27.75	1.08	0.90	1.04	30.76
H	18.47	0.66	0.64	0.59	20.35
B	6.27	0.21	0.23	0.20	6.92
O	6.52	0.21	0.22	0.22	7.17
ALL	59.01	2.15	1.99	2.05	65.20

Table 7: Number of tokens in million (English)

	TRAIN	DEV	DEVTEST	TEST	Whole
G	30.15	1.18	0.97	1.14	33.44
H	20.16	0.72	0.71	0.64	22.22
B	6.76	0.23	0.25	0.22	7.47
O	7.04	0.22	0.23	0.24	7.74
ALL	64.12	2.35	2.16	2.24	70.87

Table 8: Number of tokens in million (Japanese)

tokens (running words) in these datasets. Tables 5 and 6 list the number of types for English and Japanese sentences. Tables 7 and 8 list the number of tokens. To count these numbers, we used ChaSen to segment Japanese sentences into tokens and used a simple tokenizer[10] to tokenize English sentences. All English words were lowercased. In these tables, the figures in columns "TRAIN", "DEV", "DEVTEST" and "TEST" are the number of types and tokens in these datasets and the figures in the "Whole" columns are the number of types and tokens in each G, H, B, O, and ALL.

4.3. 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. Figure 1 shows the percentage of sentences for English (en) and Japanese (ja) with respect to their lengths. This figure shows that the distributions of sentence length were relatively flat and that there were many long sentences in ALL. This suggests that patents contain many long sentences that are generally difficult to translate.

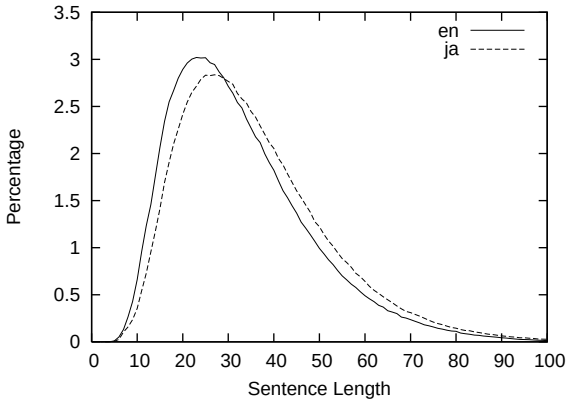

Figure 1: Sentence length distribution

We then measured the coverage of vocabulary. Tables 9 and 10 list the coverage for the types and tokens in each TEST section using the vocabulary in the corresponding TRAIN section for the English and Japanese datasets. 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.

5. MT experiments

5.1. 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, 2004), SRILM (Stolcke, 2002), GIZA++ (Och and Ney, 2003), mkcls (Och, 1999), Carmel,[11] and a phrase model training code.

We followed the instruction of the shared task baseline system to train our MT systems.[12] 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 tune our MT systems, we did minimum error rate training[13] (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.[14] We tokenized and lowercased the TRAIN, DEV, DEVTEST, and TEST data sets as described in Section 4.2. We conducted two MT experiments as described in the following sections.

[10] `http://www2.nict.go.jp/x/x161/members/mutiyama/software/ruby/tokenizer.rb`

[11] `http://www.isi.edu/licensed-sw/carmel/`

[12] 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.

[13] The minimum error rate training code we used is available at `http://www2.nict.go.jp/x/x161/members/mutiyama/software.html`.

[14] %BLEU score is defined as BLEU $\times$ 100.

	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

Table 9: Percentages of words in test sentences covered by training vocabulary (English)

	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

Table 10: Percentages of words in test sentences covered by training vocabulary (Japanese)

5.2. Comparing reordering limits

For the first experiment, we translated 1000 randomly sampled sentences in each DEVTEST data set to compare different reordering limits.[15] We compared a reordering limit of 4 with no limitation. The results of Tables 11 and 12 show that the %BLEU scores for no limitation consistently outperformed those for "limit=4". These results coincide with those of Koehn et al. (2005) who reported that larger reordering limits provide better performance for Japanese-English translations. Based on this experiment, we used no reordering limit in the next experiment.

5.3. Cross section MT experiments

For the second experiment, we conducted cross section MT experiments. The results are shown in Tables 13 and 14. For example, as listed in Table 13, 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 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 4 for the number of sentences in each section of TRAIN and TEST).

	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

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 the table. The source language was English and the target language was Japanese.

Table 11: Comparing reordering limits (English-Japanese)

[15]The parameter "-dl" for the Pharaoh decoder.

	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

Table 12: Comparing reordering limits (Japanese-English)

The results in these tables show that MT systems performed the best when the training and test sections were the same. 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 MT system trained in a section to another section. 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.

Table 15 lists 15 example translations obtained from the Japanese-English MT system trained and tested on TRAIN and TEST of ALL. Reference translations were marked using an R and MT outputs were marked using 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 3.2. 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*[16] and selected five sentences from the top, middle, and bottom of these sorted sentences.[17]

Table 15 shows that top examples (1 to 5) were very good translations. These MT translations consisted of long phrases that contributed to the fluency and adequacy of translations. We think that the reason for this good translation is partly due to the fact that patent documents generally

[16]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).

[17]We skipped sentences whose MT outputs contained untranslated Japanese words when selecting these 15 sentences.

TRAIN \ TEST	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 13: %BLEU scores (relative %BLEU scores) for cross section MT experiments (English-Japanese)

TRAIN \ TEST	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)

Table 14: %BLEU scores (relative %BLEU scores) for cross section MT experiments (Japanese-English)

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).

6. Discussion

We have described the characteristic of our patent parallel corpus and showed that it could be a good corpus for promoting MT research. In this section, we describe three issues about ALL that we found during investigating it as described in Sections 3., 4., and 5. We want to resolve these issues when we extend it for the NTCIR-7 patent MT task.

Issue 1. Our noise reduction procedure described in Section 3.2. reduced the number of sentences from about 4.2 million to about 3.9 million. This reduction could be too aggressive. We want to investigate the effect of noise reduction on the MT performance in our future work.

Issue 2. The English tokenizer we used can not handle some expressions properly. For example, it can not handle character entity references. That is, it tokenizes

&

as

& amp ;

for example. Although the sentences containing character entity references are about 0.3% in ALL, we want to improve our tokenizer to remedy tokenization errors.

Issue 3. We randomly split patents into TRAIN, DEV, DEVTEST, and TEST as described in Section 4.2. This split resulted in a lot of repetitions of long and/or specific expressions as described in the previous section. Consequently, the %BLUE scores obtained in our experiments could be optimistic. We want to try another split in our future work. For example, we can use the patents in 1993 to 1997 as TRAIN, and use those in 1998, 1999, and 2000 as DEV, DEVTEST, and TEST, respectively.

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 developed a Japanese-English patent parallel corpus created from the Japanese and US 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, whose size is comparable to other large-scale parallel corpora. This corpus and its extension are planed to be used in the NTCIR-7 patent MT task.

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

8. References

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

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, pages 359–365.

	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 . |

Table 15: Examples of reference (R) and machine (M) translations

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. 2001. Prime: A system for multi-lingual patent retrieval. In *MT Summit VIII*, pages 163–167.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *IWSLT*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *AMTA*.

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

Xiaoyi Ma and Christopher Cieri. 2006. Corpus support for machine translation at LDC. In *LREC*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Makoto Nagao. 1981. A framework of a mechanical translation between Japanese and English by analogy principle. In *the International NATO Symposium on Artificial and Human Intelligence*. (appeared in Sergei Nirenburg, Harold Somers and Yorick Wilks (eds.) *Readings in Machine Translation* published by the MIT Press in 2003).

Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto, and Kazuteru Ohashi. 2006. A clustered global phrase reordering model for statistical machine translation. In *ACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *EACL*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL*, pages 72–79.

Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. 1994. Bilingual text matching using bilingual dictionary and statistics. In *COLING'94*, pages 1076–1082.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *ACL*.