Mining Parallel Texts from Mixed-Language Web Pages

Masao Utiyama **Daisuke Kawahara** **Keiji Yasuda** **Eiichiro Sumita**
National Institute of Information and Communications Technology (NICT)
Keihanna Science City 619-0288 Kyoto, Japan
{mutiyama, dk, keiji.yasuda, eiichiro.sumita}@nict.go.jp

Abstract

We propose to mine parallel texts from *mixed-language web pages*. We define a mixed-language web page as a web page consisting of (at least) two languages. We mined Japanese-English parallel texts from mixed-language web pages. We presented the statistics for extracted parallel texts and conducted machine translation experiments. These statistics and experiments showed that mixed-language web pages are rich sources of parallel texts.

1 Introduction

Parallel corpora are indispensable language resources for multi-lingual natural language processing, such as corpus-based machine translation (MT) (Nagao, 1981; Brown et al., 1993) and cross-lingual information retrieval.

However, there are relatively few widely available parallel corpora. These include the Arabic-English and Chinese-English parallel corpora distributed by the Linguistic Data Consortium (Ma and Cieri, 2006); the Europarl corpus (Koehn, 2005), which consists of 11 European languages; the JRC-Acquis corpus, which consists of more than 20 European languages (Steinberger et al., 2006); and a Japanese-English patent parallel corpus (Utiyama and Isahara, 2007). Although these parallel corpora are large scale, they are limited in the language registers and language pairs that they cover.

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. Zhao and Vogel (2002), Utiyama and Isahara (2003), Fung and Cheung (2004), and Munteanu and Marcu (2005) have extracted parallel sentences from comparable or non-parallel corpora.

In this paper, we mine parallel texts from the web (Ma and Liberman, 1999; Resnik and Smith, 2003; Shi et al., 2006). The novel contribution of our work compared to previous work is that we propose to mine parallel texts from *mixed-language web pages*. We define a mixed-language web page as a web page consisting of (at least) two languages. We mine Japanese-English parallel texts from mixed-language web pages consisting of Japanese and English texts.

In contrast to our work, previous studies have mined parallel texts from *parallel web pages*. A pair of parallel web pages consists of two monolingual web pages in different languages with almost the same meaning. For example, Shi et al. (2006) have aligned parallel English and Chinese web pages and aligned sentences in these aligned pages.

We verify that mixed-language web pages are rich sources of parallel texts. Our work complements the previous work. By combining our work with the previous work, it will be possible to mine parallel texts from both mixed-language web pages and parallel web pages.

In Section 2, we describe how we mine parallel texts from mixed-language web pages. In Section 3, we show basic statistics for the parallel texts obtained. In Section 4, we use the extracted parallel texts to improve the performance of an SMT system.

2 Mining parallel texts

2.1 A pilot sutdy

The degree of parallelness of mixed-language web pages is wide. First of all, most Japanese web pages that contain both English and Japanese texts do not contain parallel texts.

As a pilot study, we examined 10,000 Japanese web pages to see how many of them contained at least one English sentence. We found that about 20% of them contained English sentences. Twenty percent is a fairly large percentage. However, we found no web pages that contained parallel texts in these 10,000 pages. The English sentences contained in these pages were conventional ones, such as “All rights reserved,” which had no counterparts in Japanese.

From this experiment, we estimated the percentage of mixed-language web pages containing parallel texts to be 0.01% ($\sim \frac{1}{10000+2} \times 100$) according to Laplace’s Law (Manning and Schütze, 1999).

This pilot study shows that we need to impose some constraints on the pages we search for parallel texts. In Section 2.2, we describe how we obtain mixed-language web pages.

2.2 Obtaining mixed-language web pages

We need mixed-language web pages that are suitable for mining parallel texts. We adopt the following procedure to obtain such mixed-language web pages, which consist of Japanese and English texts: (1) Crawl Japanese web pages, (2) process each web page, and (3) extract mixed-language web pages based on several constraints.

2.2.1 Crawl Japanese web pages

We first crawl Japanese web pages. Japanese web pages are efficiently crawled by using a web crawler with a Japanese filter. This Japanese filter uses the meta information of web pages (HTML), and a linguistic characteristic of the Japanese language in the same way as the method used for constructing a Japanese web corpus (Kawahara and Kurohashi, 2006).

(1) Check that the “charset” of the HTML header is one of the Japanese encodings: euc-jp, x-euc-jp, iso-2022-jp, shift_jis, windows-932, x-sjis, shift-jp, shift-jis, or utf-8

(2) If the charset is utf-8, the web page is possibly written in a non-Japanese language. To extract only Japanese pages, check for the existence of a Japanese postposition (*ga*, *wo*, *ni*, *ha*, *no* or *de*) in the HTML body.

We crawled 100 million Japanese pages in this study.

2.2.2 Process each web page

We extract text portions from each web page, and split them into sentences. Then, we judge whether each sentence is Japanese or English.

(1) Split a web page into sentences: A web page is split into sentences using periods and HTML tags, such as “br” and “p”.

(2) Judge whether a sentence is Japanese or English: On Japanese web pages, sentences that are not written in Japanese are mostly English sentences. Therefore, we assume that non-Japanese sentences are written in English when judging whether a sentence is Japanese or English. We consider a sentence to be English if it satisfies all of the following four conditions, otherwise it is judged as Japanese:[1] (A) The sentence does not contain Japanese-specific characters such as HIRAGANA, KATAKANA and KANJI. (B) The sentence contains at least one white space. (C) The sentence ends with “.”, “?” or “!”. (D) More than 90% of the characters in the sentence match [a-zA-Z,.?!].

The results of this process could contain noisy English or Japanese sentences. That is, non-English or non-Japanese sentences could be judged as English or Japanese sentences. However, the alignment method described in Section 2.3 is able to extract clean English-Japanese sentence alignments from these potentially noisy sentences, as verified in Section 3.

2.2.3 Extract mixed-language web pages

If we merely extract mixed-language web pages that contain an English sentence, we erroneously obtain web pages that just contain conventional English sentences, such as “All rights reserved,” which have no counterparts in Japanese. Therefore, we impose the following two constraints to extract mixed-

[1] We impose strong conditions for English judgment, because we use the number of English sentences as a constraint in the subsequent method.

language web pages that possibly contain parallel sentences: (1) The web page contains one of the following 10 Japanese words (in KANJI) that imply the existence of translations, *eigo* (English), *hon'yaku* (translation), *wayaku* (Japanese translation), *eiyaku* (English translation), *eikaiwa* (English conversation), *eibun* (English sentence), *taiyaku* (translation pair), *yakubun* (translation), *nihongoyaku* (Japanese translation), and *houyaku* (Japanese translation). (2) The web page contains more than N English sentences. We conducted a preliminary experiment and empirically determined N to 10.[2]

Finally, we applied the above method to the 100 million Japanese web pages. As a result, 113,420 mixed-language web pages were obtained.

2.3 Alignment procedure

We selected Utiyama and Isahara's alignment method (Utiyama and Isahara, 2007) from various methods for aligning comparable or noisy parallel texts (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2005). This is because their method has been successfully applied in aligning Japanese-English noisy parallel texts (Utiyama and Isahara, 2003; Utiyama and Isahara, 2007) and we could use their tool off-the-shelf.

In order to apply their method to mixed-language web pages, we converted these pages into noisy parallel text files. That is, given a web page containing Japanese and English texts, we made a Japanese text file and an English text file from the web page.[3] We regarded these two text files as a pair of noisy parallel text files and applied Utiyama and Isahara's method to these. In the following, we briefly describe how we applied Utiyama and Isahara's method to these parallel texts. See (Utiyama and Isahara, 2007) for details of their method.

We first aligned the sentences in each pair of noisy parallel text files by using a standard dynamic programming (DP) matching method (Gale and Church, 1993; Utsuro et al., 1994). That is, let J and E be a Japanese text file and an English text file, respectively, we calculated the maximum similarity sentence alignments $(J_1, E_1), (J_2, E_2), \dots (J_m, E_m)$ using DP matching, where J_i and E_i were Japanese and English sentences in J and E. We allowed 1-to-n, n-to-1 ($0 \leq n \leq 5$), or 2-to-2 alignments when aligning the sentences. The similarity between J_i and E_i (SIM(J_i, E_i)) was calculated based on word overlap (i.e., number of word pairs from J_i and E_i that were translations of each other based on a bilingual dictionary with 450,000+ entries).

We next calculated the similarity between J and E (AVSIM(J, E)) as defined by (Utiyama and Isahara, 2003), using:

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

A high AVSIM(J, E) value occurs when the sentence alignments in J and E have high similarity values.

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

$$R(J, E) = \min\left(\frac{|J|}{|E|}, \frac{|E|}{|J|}\right) \qquad (2)$$

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.

Using AVSIM(J, E) and $R(J, E)$, we defined the similarity between J and E (AR(J, E)) as

$$\text{AR}(J, E) = \text{AVSIM}(J, E) \times R(J, E) \qquad (3)$$

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

$$\text{Score}(J_i, E_i) = \text{SIM}(J_i, E_i) \times \text{AR}(J, E) \qquad (4)$$

A high Score(J_i, E_i) value occurs when (1) sentences J_i and E_i are similar, (2) documents J and E are similar, and (3) 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.

[2] Our mining method will not be much affected by N because our method can extract parallel sentences very accurately as shown in Section 3.

[3] We simply extracted Japanese (English) sentences from the web page and put them into a Japanese (English) text file.

3 Statistics for extracted parallel texts

In this section, we evaluate the performance of the alignment method described in the previous section. We applied the alignment method to the 113,420 mixed-language web pages obtained from the 100 million Japanese web pages.

3.1 Basic statistics

We first show summary statistics for the 113,420 mixed-language web pages in Table 1. The "Sentences" row shows the total number of sentences in the 113,420 pages and the "S. mean" row shows the average number of sentences on a page. The "Words" and "W. mean" rows show similar statistics for words. The figures in the "English" and "Japanese" columns are the figures for "English" and "Japanese" texts, respectively.[4]

	English	Japanese
Sentences	21,302,046	66,355,812
S. mean	188	585
Words	87,249,745	1,063,265,797
W. mean	769	9,375

Table 1: Summary statistics

Table 1 shows that the amount of Japanese texts was much larger than that of English. This is because the 100 million web pages were crawled from mainly monolingual Japanese web pages. This table suggests that the 113,420 mixed-language web pages were very noisy parallel texts.

3.2 Parallelness of mined web pages

In order to see the parallelness of the mined web pages, we used the AR values (Equation 3) that were assigned to the 113,420 web pages.

We sorted the web pages in decreasing order of their AR values. We divided these pages into seven ranges according to their ranks, as shown in the "Range" column of Table 2. Each of the first six ranges contained 3000 pages and the last one contained the remaining 95,420 pages.

We extracted 50 web pages from each range[5] and evaluated each of web pages[6] as:

- A_p if that page contained parallel texts and those parallel texts were at least 50% of the texts on that page,
- B_p if that page contained parallel texts and those parallel texts were less than 50% of the texts on that page, or
- X_p if that page did not contain parallel texts.

The evaluation results are shown in Table 2. The "A_p", "B_p", and "X_p" columns show the numbers of samples that were evaluated as the corresponding labels. The figures in the "A_p+B_p (%)" are the percentages of the samples evaluated as A_p or B_p.

This table shows that samples taken from the range "18001 –" did not contain A_p web pages. This means that almost all of the A_p web pages were ranked in the first 18,000 pages (the first 15.9% ($=\frac{18000}{113420} \times 100$) of all pages). This suggests that AR values are very effective for ranking A_p web pages.

Next, the ratio of the A_p or B_p web pages in the first 18,000 pages was about 73%. From this figure, we estimated the number of A_p or B_p web pages in the first 18,000 pages as 13,200 ($= 18000 \times 0.733..$). We also estimated the number of B_p web pages in the remaining 95,420 pages as 20,992 ($= 95420 \times \frac{11}{50}$).

Note that the percentage of mixed-language web pages containing parallel texts was estimated to be 0.01% in Section 2.1. In this section, we have shown that about 13,200 pages are A_p or B_p pages in the first 18,000 pages. These 13,200 pages are 0.0132% of the 100 million web pages. This means that we have extracted a significant subset of parallel mixed-language pages in the first 18,000 pages.

These figures and Table 2 show that we can rank A_p or B_p web pages highly using AR values (the precision is good) but we also miss some B_p web pages if we use only highly ranked web pages (the recall could be improved). Overall, we concluded that we can use AR values for ranking A_p or B_p web pages highly.

[4] We used a Japanese morphological analyzer, ChaSen, to segment Japanese texts into words.

[5] When we randomly sampled web pages, we discarded sample pages if they contained more than 1000 sentences, in order to reduce the evaluation workload. As a result, we discarded 22 pages in the process of sampling 350 pages.

[6] The evaluation conducted in Sections 3.2 and 3.3 were performed by an expert evaluator who has been performing similar jobs more than five years.

Range	A_p	B_p	X_p	A_p+B_p(%)
1 – 3000	39	4	7	86
3001 – 6000	36	10	4	92
6001 – 9000	30	13	7	86
9001 – 12000	28	11	11	78
12001 – 15000	10	15	25	50
15001 – 18000	12	12	26	48
18001 –	0	11	39	22

Table 2: Parallelness of mined web pages

Next, we examined genres of the 231 A_p or B_p web pages in Table 2, in order to examine what kinds of parallel texts were on a single page. We found 13 genres. The top 5 genres were (1) personal opinion (blogs, chat, email etc), (2) computer (software manuals, online games, etc), (3) example sentences, (4) book, and (5) daily conversation. These results show that the genres of mixed language web pages are wide.

3.3 Accuracy of sentence alignments

We show the accuracy of sentence alignments in this section. Our sentence alignments were obtained as follows. First, we obtained 6.3 million one-to-one sentence alignments, as a result of applying the alignment method described in Section 2.3. This was about 30% of the English sentences according to Table 1. Next, we removed the alignments whose English sentences did not end with periods, exclamation marks, or question marks to reduce alignment pairs considered as noise.[7] We also removed some sentence pairs that were too imbalanced ($\frac{\text{length of longer sentence}}{\text{length of shorter sentence}} > 3$). Finally, we 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, 929,011 alignments were obtained.

We sorted these sentence alignments in decreasing order of Score in Equation 4. We divided these alignments into five ranges according to their ranks, as shown in the "Range" column of Table 3. Each of the first four ranges contained 100,000 sentence alignments and the last one contained the remaining 529,011 alignments.

We extracted 100 sentence alignments from each range and evaluated each alignment as:

- A_s if 80% or more of the contents were shared between the English and Japanese sentences,
- B_s if 50% or more and less than 80% of the contents were shared
- C_s if less than 50% of the contents were shared (The English and Japanese sentences should share some contents, but the amount of the shared contents were less than 50%),
- X_s if the meanings of the Japanese and English sentences were totally different.

The evaluation results are shown in Table 3. The figures in the "A_s", "B_s", "C_s", and "X_s" columns are the number of samples that were evaluated as the corresponding labels.

Range	A_s	B_s	C_s	X_s
1- 100,000	88	9	2	1
100,001 – 200,000	71	9	4	16
200,001 – 300,000	37	4	7	52
300,001 – 400,000	7	3	1	89
400,001 –	1	0	0	99

Table 3: Accuracy of sentence alignments

This table shows that most of the A_s alignments were extracted in the first 300,000 alignments. This suggests that Score in Equation 4 ranked good sentence alignments highly.

The percentages of the A_s alignments in the ranges "1 – 100,000" and "100,001–200,000" were 88% and 71%, respectively. This suggests that the first 200,000 sentence alignments were effective for multi-lingual natural language processing.

Based on the statistics presented in Sections 3.2 and 3.3, we concluded that we extracted a clean parallel corpus from the original web corpus.

4 Machine translation experiments

We verify the usefulness of the extracted sentence alignments for SMT in this section.

We used a state-of-the-art phrase-based SMT system (Finch and Sumita, 2008), which is comparable

[7] The tool we used for alignment automatically re-segmented Japanese and English texts. As a result, the sentence segmentations obtained in Section 2.2.2 were changed. Consequently, we again needed to remove noisy English sentences.

in performance to the MOSES system (Koehn et al., 2007). To train SMT models, we used a training toolkit adapted from the MOSES system. We used GIZA++ (Och and Ney, 2003) for word alignment and SRILM (Stolcke, 2002) for language modeling. We used 5-gram language models trained with modified Kneser–Ney smoothing. Minimum error rate training was used to tune the decoder's parameters on the basis of the bilingual evaluation understudy (BLEU) score (Papineni et al., 2002), and tuning was performed using the standard technique developed by Och (Och, 2003).

We used the development data for the IWSLT-2007 Japanese-English translation task (Fordyce, 2007) to verify the usefulness of the extracted sentence alignments. The development data consisted of five sets, devset1, devset2, devset3, devset4, and devset5. Each of these data sets had about 500 sentences. The numbers of reference translations were 16 for devset1, devset2, and devset3 and 7 for devset4 and devset5.

We used devset1 to tune the SMT system and used devset2, devset3, devset4, and devset5 as the testsets to evaluate the performance of the SMT system in terms of BLEU scores. Hereafter, we refer to devset2, ..., devset5 as set2, ..., set5, respectively.

In the following experiments, we only change the training data that were used for making our language and translation models, in order to compare various parallel texts.

4.1 Relationship between alignment data size and BLEU scores

As described in Section 3.3, we have extracted about 900,000 sentence alignments. In this section, we increase the size of training data to see how MT performance evolves using more data.

First, we made an English 5-gram language model from the first 900,000 sentence alignments. Next, we used the first 100,000, 200,000, ..., 900,000 sentence alignments to make our translation models (a phrase-table and reordering table). Each of these models was used with the same 5-gram language model.

The BLEU scores for these settings are shown in Figure 1. The lines entitled “set2,” “set3,” “set4,” and “set5” indicate the BLEU scores for these testsets. The middle line, which is entitled “mean,” indicates the mean value of the BLEU scores for these 4 testsets.

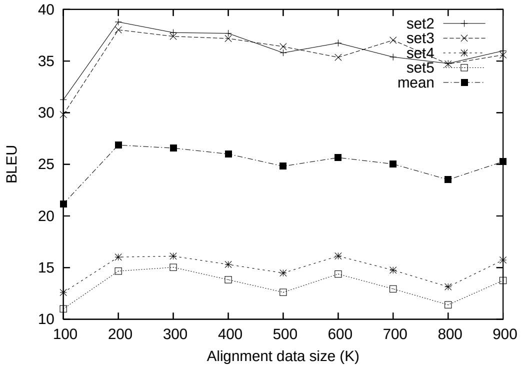

Figure 1: Relationship between alignment data size and BLEU scores

Figure 1 shows that the highest mean BLEU score was obtained when using the first 200,000 sentence alignments. The BLEU scores decreased when we used more than 200,000 alignments.

It is interesting that this observation is consistent with that in Section 3.3. In that section, we observed that the first 200,000 sentences were relatively clean parallel sentences. In this section, we showed that using 200,000 sentences resulted in the highest BLEU score. This suggests that adding noisy parallel sentences is detrimental to improving BLEU scores.

4.2 IWSLT training data

Next, we used the training data for the IWSLT-2007 workshop. It consisted of about 40,000 English-Japanese parallel sentences. We used this training data to make our language and translation models. The BLEU scores are shown in Table 4. The values in the “Mean” column are the averages of the BLEU scores for set2, set3, set4 and set5.

set2	set3	set4	set5	Mean
58.70	58.15	24.55	20.48	40.47

Table 4: Results for the IWSLT training data

Figure 1 and Table 4 show that the performance of the SMT systems trained with our extracted sentence alignments are inferior to that of the SMT system trained with the IWSLT training data. A likely reason is that the extracted alignments are out-ofdomain data with respect to the IWSLT testsets. In the following, we show that the extracted alignments are useful for improving the performance of the SMT system trained with the IWSLT training data, even though these alignments are not best suited to the testsets.

4.3 Interpolation of models

We linearly interpolated[8] language and translation models (Foster and Kuhn, 2007) to improve the performance of the SMT system.

4.3.1 Interpolation of language models

We first interpolated language models (LMs). We interpolated the language model made from 900,000 sentences in Section 4.1 (hereafter *LM(900k)*) and that made from the IWSLT training data in Section 4.2 (hereafter *LM(IWSLT)*). The weight of LM(IWSLT) was 0.1, 0.2, ..., 0.9. In addition to these interpolated language models, we used the translation model made from the IWSLT training data in Section 4.2 for all of the weights.

The figures in the 0.1, ..., 0.9 rows in Table 5 show the BLEU scores for set2, ..., set5, along with the mean values. The figures in the "IWSLT" row were taken from Table 4. This table shows that the interpolation of the language models improved BLEU scores for a wide range of interpolation weights.

weight	set2	set3	set4	set5	mean
0.1	58.35	58.45	26.10	21.46	41.09
0.2	57.67	58.89	**26.52**	21.16	41.06
0.3	**59.25**	58.90	25.99	21.63	41.44
0.4	58.70	58.24	25.05	21.77	40.94
0.5	58.87	**59.43**	26.12	**21.95**	**41.59**
0.6	58.12	58.63	25.14	20.95	40.71
0.7	58.73	58.71	24.85	21.11	40.85
0.8	56.94	57.04	22.93	18.71	38.91
0.9	58.73	58.69	24.75	20.75	40.73
IWSLT	58.70	58.15	24.55	20.48	40.47

Table 5: Results for interpolation of LMs

4.3.2 Interpolation of translation models

We next interpolated translation models (TMs). We interpolated the translation model (a phrase-table and reordering-table) made from the first 200,000 sentences in Section 4.1 (hereafter *TM(200k)*) and that made from the IWSLT training data in Section 4.2 (hereafter *TM(IWSLT)*). The weight of TM(IWSLT) was 0.1, 0.2, ..., 0.9. We used LM(IWSLT) for all of the weights.

The figures in Table 6 shows the BLEU scores for this setting. These shows that the interpolation of the translation models improved BLEU scores for a wide range of interpolation weights.

weight	set2	set3	set4	set5	mean
0.1	59.00	55.62	22.72	19.18	39.13
0.2	61.62	58.61	24.79	20.86	41.47
0.3	62.24	59.89	25.10	21.09	42.08
0.4	61.51	**60.53**	**26.24**	**21.89**	42.54
0.5	**62.77**	**60.53**	25.72	21.42	**42.61**
0.6	60.39	58.24	24.47	20.08	40.80
0.7	61.05	58.92	24.75	21.22	41.49
0.8	59.75	58.20	23.09	19.09	40.03
0.9	59.05	59.22	24.41	20.86	40.89
IWSLT	58.70	58.15	24.55	20.48	40.47

Table 6: Results for interpolation of TMs

4.3.3 Interpolation of both models

Finally, we interpolated both language and translation models. We interpolated LM(900k) and LM(IWSLT) with equal weights and interpolated TM(200k) and TM(IWSLT) with equal weights. The "LM&TM" row shows the BLEU scores for this setting. The "IWSLT", "LM", and "TM" rows were taken from Tables 4, 5, and 6, respectively. This table shows that "LM&TM" were not always better than "LM" or "TM". However, "LM&TM" were always better than "IWSLT."

	set2	set3	set4	set5	mean
IWSLT	58.70	58.15	24.55	20.48	40.47
LM	58.87	59.43	**26.12**	**21.95**	41.59
TM	**62.77**	60.53	25.72	21.42	42.61
LM&TM	61.93	**61.44**	25.79	21.50	**42.67**

Table 7: Results for interpolation of both models

Based on the experiments in Sections 4.3.1, 4.3.2, and 4.3.3, we concluded that our mined sentence alignments are useful for improving the performance of a state-of-the-art SMT system.

[8] Let p_1 and p_2 be two probabilities and w be the weight of p_1, the linear interpolation of these probabilities is $wp_1 + (1 - w)p_2$, where $0 \leq w \leq 1$.

5 Conclusion

The web is a rich source of parallel texts. Indeed, researchers have already been mining parallel texts from *parallel web pages*. In contrast, we have proposed to mine parallel texts from *mixed-language web pages*. We have extracted parallel texts from mixed-language web pages containing Japanese and English. We have presented the statistics for these extracted parallel texts and conducted MT experiments. These statistics and experiments have shown that mixed-language web pages are rich sources of parallel texts.

Our work complements the previous work. By combining our work and the previous work, it will be possible to mine parallel texts from both mixed-language web pages and parallel web pages.

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.

Andrew Finch and Eiichiro Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *the Third Workshop on Statistical Machine Translation*, pages 208–215.

Cameron S. Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *IWSLT*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *the Second Workshop on Statistical Machine Translation*, pages 128–135.

Pascale Fung and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *EMNLP*, pages 57–63.

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

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the Web using high-performance computing. In *LREC*, pages 1344–1347.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, pages 79–86.

Xiaoyi Ma and Christopher Cieri. 2006. Corpus support for machine translation at LDC. In *LREC*, pages 859–864.

Xiaoyi Ma and Mark Y. Liberman. 1999. BITS: A method for bilingual text search over the Web. In *MT Summit*.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

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).

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

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.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM tree alignment model for mining parallel data from the web. In *COLING/ACL*, pages 489–496.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC*, pages 24–26.

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.

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

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

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of 2002 IEEE International Conference on Data Mining*, pages 745–748.