

# Preprocessing Methods for Word Alignment

Anonymous Author

## Abstract

This paper compares four preprocessing approaches for word alignment: 1) sentence removal approach, 2) good points approach, 3) sentence duplication approach, and 4) removal of doubtful alignments approach. Two are statistically motivated and the other two are heuristics. We focus on the ability of a word aligner of IBM Model 4 that it should often face with troubles when handling paraphrase, multi-words and non-literal translation. We assume that IBM Model 4 works 90 % correct, while only around 5 % wrong.

## 1 Introduction

A phrase-based approach (Koehn et al., 03) has become the main stream in SMT (Statistical Machine Translation) despite its origin as a word-based approach (Brown et al., 93). While the phrase alignment (Marcu and Wong, 02) has recently attracted researchers in its theory but still in infancy in its practice, the word alignment has been used quite widely combined with the phrase extraction strategy (Koehn et al., 05) to provide phrase tables to the decoder. In this context, this paper aims at improving the quality of word alignments by preprocessing the parallel corpus.

For a given sentence aligned parallel corpus, a word alignment task is to obtain lexical translation probabilities between bilingual pair of words. The approach based on Bayesian generative models, such as IBM models (Brown et al., 93), HMM alignment models (Vogel et al., 96) and IBM Model 6 (Och and Ney, 03), has been dominant although several other approaches have appeared as well, such as discriminative approaches (Moore, 05) and posterior-based approaches (Liang, 06). This paper restricts ourselves to consider only on

Bayesian generative models which use EM (Expectation Maximization) algorithm (Dempster et al., 97) due to its dominance in practice. The software in this line includes GIZA++ (Och and Ney, 03), mtk toolkit (Deng and Byrne, 05), and mgiza (Gao and Vogel, 08).

Our concern is on the ability of EM algorithms due to the fact that parallel corpus is real life data: Can EM algorithm correctly handle paraphrase (Callison-Burch, 07; Lin and Pantel, 01), non-literal translation (Imamura et al., 03), and multi-words expressions (Lambert and Banchs, 05)? (From now on, we call these *outliers*, meaning that *outliers* are more than the normal systematic *noise*.) Due to the definition of paraphrase, non-literal translation and multi-words expressions, we can align them in phrase level as a phrase pair in general, but not in word level as a word pair if we exclude all the special cases of  $1 : n$  correspondence. However, the alignment which can be resolved by a word aligner is restricted to  $1 : n$  alignments. Hence, the  $n : m$  alignments are out of reach. So the answer to this question becomes rather negative. Under the incorrect word alignments, can we get a correct  $n : m$  phrase pairs after we do phrase extraction? Is there some mechanism to remove incorrect alignments in order not to provide such incorrect alignments to the next phrase extraction step? These tend to be negative although there are some possibilities. These outliers may be a potential danger in terms of quality of word alignment and translation quality as a whole due to its harmful wrong matching of words with wrong probabilities.

Lambert and Banchs (Lambert and Banchs, 05) extract multiword expressions and make them into one token before supplying to a word aligner. They mentioned that if the size of corpus is small, they may not be extracted due to its sparsity in the corpus. Callison-Burch et al. (Callison-

Burch et al., 06) shows the approach to supply the unknown phrases simply by externally extracted paraphrases.

This paper approaches the problem of *outliers* purely by the given parallel corpus as in the line of Lambert and Banchs. In statistics, there are several common practices to handle outliers: to search for the *good* points or to change the models that the distribution has heavier tails. Alternatively, we also mentioned some heuristic approaches where we check the alignment results. It is noted that our analysis only applies to word alignment tools which use EM algorithm. It will need another analysis for discriminative word alignment.

This paper is organized as follows. Section 2 outlines the  $1 : n$  characteristics of word alignment task and mentions why paraphrase, non-literal translation, and multi-word expression are difficult for a word aligner. Section 3 explains four algorithms. Experimental results are presented in Section 4. Section 5 concludes and provides avenues for further research.

## 2 $1 : n$ Word Alignment and Outliers

This section explains uni-directional alignments of word alignment in an empirical manner. Our discussion is limited ourselves to IBM Model 4.

**Definition 1 (Word alignment task)** Let  $e_i$  be the  $i$ -th sentence in target language,  $\bar{e}_{i,j}$  be the  $j$ -th word in  $i$ -th sentence, and  $\bar{e}_i$  be the  $i$ -th word in parallel corpus (Similarly for  $f_i$ ,  $\bar{f}_{i,j}$ , and  $\bar{f}_i$ ). Let  $|e_i|$  be a word length of  $e_i$ , and similarly for  $|f_i|$ . We are given a pair of sentence aligned bilingual texts  $(f_1, e_1), \dots, (f_n, e_n) \in \mathcal{X} \times \mathcal{Y}$ , where  $f_i = (\bar{f}_{i,1}, \dots, \bar{f}_{i,|f_i|})$  and  $e_i = (\bar{e}_{i,1}, \dots, \bar{e}_{i,|e_i|})$ . It is noted that  $e_i$  and  $f_i$  may include more than one sentence. The task of word alignment is to find a lexical translation probability  $p_{\bar{f}_i} : \bar{e}_i \rightarrow p_{\bar{f}_i}(\bar{e}_i)$  such that  $\sum p_{\bar{f}_i}(\bar{e}_i) = 1$  and  $\forall \bar{e}_i : 0 \leq p_{\bar{f}_i}(\bar{e}_i) \leq 1$  (It is noted that some models such as IBM Model 3 and 4 have deficiency problems). It is noted that there may be several words in source language and target language which do not map to any words, which are called unaligned (or null aligned) words. Triples  $(\bar{f}_i, \bar{e}_i, p_{\bar{f}_i}(\bar{e}_1))$  (or  $(\bar{f}_i, \bar{e}_i, -\log_{10} p_{\bar{f}_i}(\bar{e}_1))$ ) are called *T-tables*.

For example, we can observe from the upper figure in Figure 4 that most of the alignments, which are the results of a word aligner for DE-EN News Commentary corpus, result in either  $1:1$  mapping or NULL insertions; small numbers are

it is a pity . i cannot go today . NULL ( { } ) i ( { } ) am ( { 5 } ) sorry ( { 1 2 4 } ) that ( { 3 } ) i ( { 10 } ) cannot ( { 9 } ) go ( { 8 } ) today ( { 7 } ) . ( { 6 } )
---

Figure 1: The A3 file for a toy example. A word ‘sorry’ is mapped into 3 length word ‘it is pity’ (1, 2, 4), while ‘am’ is mapped into ‘.’, ‘that’ is into ‘a’, the second ‘i’ is into ‘.’, ‘cannot’ is into ‘today’, and ‘go’ is into ‘go’.

	words	#	fer=0	1	2	3
2	i	7	0.3	0.7	0	0
3	cannot	4	0	1	0	0
4	go	4	0	0.96	0.04	0
5	today	4	0	0.8	0	0.2
6	,	1	0.3	0.7	0	0
10	sorry	1	0	0.8	0	0.2

Figure 2: A part of fertility table for a toy example. The second column shows the words used in source language. The columns from 4 to 7 correspond to the fertility 0 to 3. This table can be read in this way: the word ‘i’ becomes zero length word in probability 0.3 and 1 length word in probability 0.7 (2nd row). Similarly, the word ‘today’ can be converted into 1 length word in 0.8 and 3 length word in 0.2 (5th row).

Source Language	Target Language
to my regret i cannot go today . i am sorry that i cannot visit today . it is a pity that i cannot go today . sorry . today i will not be available	i am sorry that i cannot visit today . it is a pity that i cannot go today . sorry . today i will not be available to my regret i cannot go today .
GIZA++ alignment results for IBM Model 4	
i NULL 0.667	available pity 1
cannot available 0.272	cannot sorry 0.55
it am 1	go sorry 0.667
is am 1	am to 1
sorry go 0.667	sorry to 0.33
go 1	to . 1
that regret 0.25	my . 1
cannot regret 0.18	will is 1
visit regret 1	not is 1
regret not 1	a that 1
be pity 1	pity that 1
	today . 1
	.. 1
	i cannot 0.33
	that cannot 0.75

Figure 3: Example shows an alignment of paraphrase in a monolingual case. Source and target use the same set of sentences. Results show that only the matching between colon is correct. It is noted that there might be a criticism that this is not a fair comparison because we do not have sufficient data. Under a transductive setting (where we can access the test data), we believe that our statement is valid. Considering the nature of  $1 : n$  mapping, it would be quite lucky if we obtain  $n : m$  mapping after phrase extraction (Our focus is not on the incorrect probability, but the incorrect matching.)

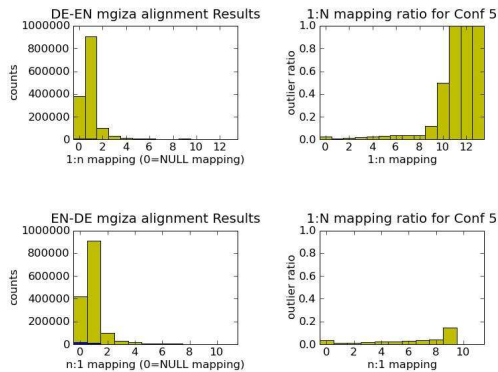


Figure 4: Two figures in the left show the results of word alignment with (main figures in yellow) and without Algorithm 2 (miniscule amounts at the bottom in blue) for DE-EN. We check all the alignment cept pairs in training corpus inspecting so-called A3 final files whether they fall in which types of alignments from 1:1 to 1:13 (or NULL alignment). It is noted that the results of our algorithm are miniscule in the left figure because all the counts are only 3 percents. Most of them are NULL alignment or 1:1 alignment, while there are small numbers of alignments take 1:3 and 1:4 (up to 1:13 in the DE-EN direction). In EN-DE direction in the figure below, 1:11 is the greatest. Two figures in the right shows the ratio of outliers over all the counts. Upper right figure shows that in the case of 1:10 alignments, 1/2 of alignments are considered to be outliers by our Algorithm 2, while 100 percents of alignment from 1:11 to 1:13 are considered to be outliers. Lower right figure shows that in the case of EN-DE, most of the outlier ratio are less than 20 percents.

1:2 mappings and miniscule numbers are from 1:3 to 1:13.

The aim of IBM Model 4 is to make a  $1 : n$  uni-directional word alignment. The modification of length  $n$  in  $1 : n$  alignments are done by a fertility and a NULL insertion. A fertility is a mechanism to augment one word into several words or none as in Figure 2, while a NULL insertion is a mechanism to create several words from blank words. A fertility is a conditional probability depending only on the lexicons. For example, the length of ‘today’ can be conditioned only on the lexicon ‘today’. So we know from Figure 2 that ‘today’ becomes 1 length word or 3 length word. It will not be converted into 2 length word or more than 4 length word.

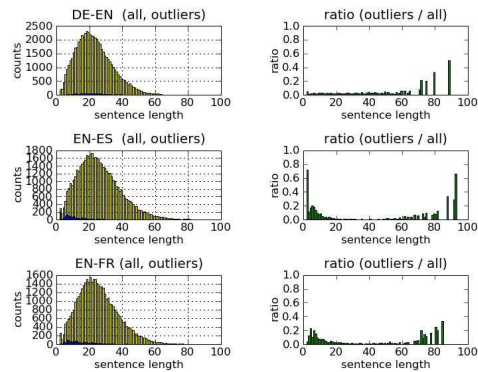


Figure 5: Three figures in the left show the histogram of sentence length (main figures) and histogram of sentence length of outliers. (As the numbers of outliers are less than 5 percents in each case, outliers are miniscule. In the case of EN-ES, we can observe the black small distributions at the bottom from 2 to 16 words length.) Three figures in the right show that if we see this by ratio of outliers over all the counts, all of three figures tend to be more than 20 to 30 percents from 80 to 100 words length. The lower two figures show that from 1 to 4 words length also tend to be more than 10 percents.

Paraphrase, non-literal translation, and multi-word expression are quite naturally appeared in parallel corpus. They are basically  $n : m$  mappings between source language and target language. If we consider the  $1 : n$  mapping nature of word alignment by IBM Model 4, they may be one potential source of outliers. Table 3 shows one example of difficulties in such a case. It is noted that we show a monolingual paraphrase for convenience in Table 3, but without loss of generality this can be easily extended for bilingual paraphrase. In this case, results of word alignment are completely wrong except colon. Although these paraphrase, non-literal translation, and multi-word expression do not always become an outlier, they may have the potential danger in producing the incorrect word alignments with incorrect probabilities.

### 3 Our Approach

We describe here four approaches under the assumption that IBM Model 4 is very close to the reality except a few wild *outliers*, i.e. paraphrase, non-literal translation and multi-word expression.

The reason behind this assumption comes from the experimental trial and errors. Especially Algorithm 2 suggests us that this figure might be around 5 percents if parallel corpus is News Commentary English-Spanish or German-English. Based on this assumption, we take an approach either to reduce parallel corpus (Algorithm 1, 2, 4) or to augment some part of parallel corpus (Algorithm 3). The first approach reduces parallel corpus by sentence length. The second approach reduces it by literalness of pair of sentences. The third approach augments the parallel corpus by frequencies conditioned on the sentence length pair. The fourth approach reduces it by the suspicious alignments after the trial alignment by a word aligner. In sum, we take an intuitive approach in all of these: even if we do not know which pair of sentences are *outliers*, if we collect some portion of pair of sentences by some measure, we may be able to avoid such *outliers*.

### 3.1 Sentence Removal Approach

This approach aims at removing *outliers* by sentence length. The algorithm is very simple which is shown in Algorithm 1. It is noted that this algorithm is a well known practice in SMT community except that it is not known how to determine  $X$  for a given parallel corpus. It is noted that while this algorithm is a well known heuristics, the following three approaches are our originals.

---

#### Algorithm 1 Sentence Removal Algorithm

---

Remove sentences whose lengths are greater than  $X$ .

---

However, the reason why this approach works is not well known. Our explanation is in Figure 5. It is noted that outliers shown in a figure in the bottom (which are almost invisible) are extracted by Algorithm 2. The region that Algorithm 1 removes is the region where the ratio of outliers are possibly high. Even if there are considerably a lot of numbers of outliers in the region that a lot of inlier reside, e.g. between 10 and 30 word length, the outlier ratio might not be big compared to the outlier ratio in the both ends, e.g. more than 60 word length or less than 5 word length. Hence if we could remove such uncertain areas whose outlier ratio are possibly high, we could make a success in removing real outliers. Hence we could interpret this approach as the approach which aims

at removing the possible high regions in terms of outlier ratio.

### 3.2 Good Points Approach

This approach aims at removing *outliers* by the literalness score between a pair of sentences. The literalness score is defined as the degree of literalness where non-literal translation is defined as the translation which is not a word-to-word translation, while literal translation is defined as a word-to-word translation. Hence, the low literalness score is the pair of sentences which should be removed.

Following two propositions are the theory behind this. Let a word-based MT system be  $M_{WB}$  and a phrase-based MT system be  $M_{PB}$ . Then,

**Proposition 1** *Under a MT system  $M_{PB}$ , a paraphrase is an inlier (or realizable), and*

**Proposition 2** *Under a MT system  $M_{WB}$ , a paraphrase is an outlier (or not realizable).*

Based on these propositions, we could assume that if we measure the literalness score under a word-based MT  $M_{WB}$  we will be able to determine the degree of *outlier*-ness whatever the measure we use for it. Hence, we score it under a word-based MT  $M_{WB}$  by Bleu for the moment (Later we replace it with the variant of Bleu, i.e. cumulative n-gram score). Hence, the summary of our approach becomes as follows: 1) employing the mechanism of word-based MT trained on the same parallel corpus, we measure the literalness between a pair of sentences. 2) we use the variants of Bleu score as the measure of literalness, and 3) based on this score, we reduce sentences. And our algorithm becomes as follows:

---

#### Algorithm 2 Good Points Algorithm

---

Step 1: Train word-based MT.

Step 2: Translate all training parallel corpus by the above trained word-based MT decoder.

Step 3: Obtain the cumulative  $X$ -gram score for each pair of sentences where  $X$  is 4, 3, 2, and 1.

Step 4: By the threshold described in text, we produce new reduced parallel corpus.

---

We would like to mention the logic why we choose the variant of Bleu. In Step 3 we need to set up a threshold in  $M_{WB}$  to determine *outliers*. Natural intuition is that this distribution takes some smooth distribution as Bleu takes weighted geometric mean. However, as is shown in the first

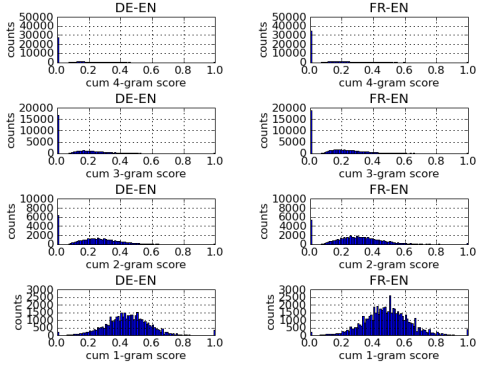


Figure 6: Each row shows the cumulative 4-, 3-, 2-, and 1-gram score, while each column shows language pairs DE-EN and FR-EN for News Commentary parallel corpus.

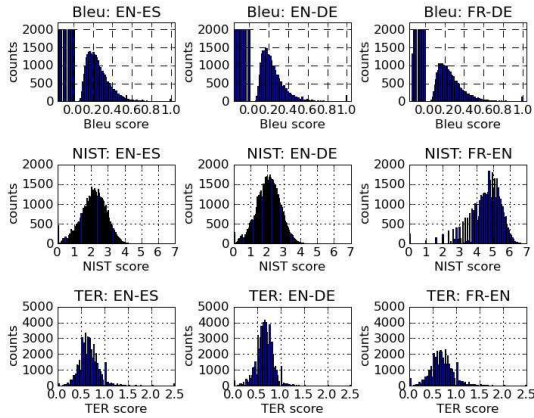


Figure 7: Each row shows Bleu, NIST, and TER, while each column shows different language pairs (EN-ES, EN-DE and FR-DE). These figures show the scores of all the training parallel corpus in configuration 5 in Algorithm 2 after training the word-based MT. In the row of Bleu, there is a trick in these figures: the area of rectangle shows the number of sentence pairs whose Bleu scores are zero. (There are a lot of sentence pairs whose Bleu score are zero: if we draw without en-folding the coordinate, these heights reach to 25000 to 30000 as in Figure 6.) There is a smooth probability distribution in the middle, while there are two non-smoothed connection at 1.0 and 0.0. Notice there are small amount of sentences whose score is 1.0. In the middle row for NIST score, similarly there is a smooth probability distribution in the middle and we have a non-smoothed connection at 0.0. In the bottom row for TER score, the 0.0 is the best score unlike Bleu and NIST, and we omit the score more than 2.5 in these figures. (The maximum was 27.0.)

row of Figure 7 typical distribution of words in this space  $M_{WB}$  is separated in two clusters: one looks like a geometric distribution and the other one is a lot of points whose value is zero. (Especially in the case of Bleu, remind that if the sentence length is less than 3 the Bleu score is zero.) In this reason, we use the variants of Bleu score: we decompose Bleu score in cumulative 4-, 3-, 2-, 1-grams, which is shown in Figure 6. In 3-gram score, the tendency to separate in two clusters is slightly decreased. Furthermore, in 1-gram score the distribution approaches to normal distribution. Although these observations are rather ad-hoc, they are the source of configurations in Table 1 which models P(outlier). It is noted that although we choose the variants of Bleu score, it is clear in our context that we can replace Bleu score as another measure, such as METEOR (Banerjee and Lavie, 05), NIST (Doddington, 02), GTM (Melamed et al., 03), TER (Snover et al., 06) and so forth (ref Figure 7).

conf	A1	A2	A3	A4
1	>0.1			
2	>0.1	>0.2		
3	>0.1	>0.2	>0.3	>0.5
4	>0.05	>0.1	>0.2	>0.4
5	>0.05	>0.05	>0.1	>0.2
6	>0.22	>0.3	>0.4	>0.6
7	>0.25	>0.4	>0.5	>0.7
8	>0.2	>0.4	>0.5	>0.8
9				>0.6
10	1	1	>0.5	>0.8

Table 1: Table shows ten configurations that we used for the experiments in Table 5. A1, A2, A3, and A4 correspond to the cumulative 4-, 3-, 2-, and 1-gram score.

Figure 2 shows outliers detected by Algorithm 2.

### 3.3 Sentence Duplication Approach

This approach is motivated purely by statistics, notably the *tails of a probability distribution*. This word means areas where the probability distribution tails off. For example in the 2-dimensional normal distribution *tails* is in the both sides of the distributions where a lot of values whose probability is miniscule. By definition, *outliers* should lie in the tails of a probability distribution. Suppose we encounter outliers with a frequency wildly un-

but this does not matter . peu importe !
we may find ourselves there once again . va-t-il en être de même cette fois-ci ?
all for the good . et c' est tant mieux !
but if the ceo is not accountable , who is ? mais s' il n' est pas responsable , qui alors ?

Table 2: Sentences judged as outliers by our Algorithm 2 (ENFR News Commentary corpus).

derpredicted by the model. If we use our generative model, i.e. IBM Model 4, the distribution that we learn by EM algorithm will result in just very close to the one that our probabilistic model predicts incorporating the generation process. Hence, our observation that we have encountered not a few outliers will not be reflected in the results by the learning process. One way to incorporate this is to make the tails heavier. Let  $\mathcal{O}$  denotes observation and  $\mathcal{M}$  denotes model. Suppose we supply some outlier model  $P(\text{outlier})$ . Then we can model this by  $(1 - \lambda)P(\mathcal{O}|\mathcal{M}) + \lambda P(\text{outlier})$  where  $\lambda \in [0, 1]$ . In our case, we do not know neither  $P(\text{outlier})$  nor  $\lambda$ . In the experiments, we use  $N$  and  $X$  by trial and errors. In sum, our algorithm becomes as follows:

---

**Algorithm 3** Sentence Duplication Algorithm

---

Step 1: Conditioned on a sentence length pair  $(l_e, l_f)$ , we count the numbers of them. We calculate the ratio  $r_{i,j}$  of this number over the number of all sentences.

Step 2: If this ratio  $r_{i,j}$  is under the threshold  $X$ , we duplicate  $N$  times.

---

### 3.4 Bad Alignment Removal Approach

The motivation of this approach is that if an aligner were not align words correctly, it could have produced wrong alignments. The heuristic consists of the following: 1) if we observe the  $1 : X$  mappings where  $X$  is big, e.g. more than 8, this mapping might be suspicious where this situation is depicted in the upper right figure of Figure 4, and 2) if we observe a lot of alignments which maps from NULL into a word or from a word into a blank word in a sentence, this mapping may be suspicious. (These two situations are often overlapped.)

Hence, the algorithm of this approach becomes

as the following:

---

**Algorithm 4** Bad Alignment Removal Algorithm

---

Step 1: Do a word alignment.

Step 2: Remove sentences whose alignment results are suspicious: 1) if it includes  $1 : X$  mappings where  $X$  is more than 8, or 2) if it includes more than  $Y$  percents of mappings in a source sentence which map into blank words (This is done by inspecting the A3 files in each direction).

---

## 4 Results

We evaluate our algorithm using News Commentary parallel corpus used in 2007 Statistical Machine Translation Workshop shared task where we use three language pairs as is shown in Table 3. We use the devset and the evaluation set provided by this Workshop. We use Moses (Koehn et al., 07) as the main MT system, with mgiza (Gao and Vogel, 08) as its word alignment tool. We do MERT in all the experiments below.

	ENFR	ENES	DEEN
baseline	0.180	0.280	0.169
size	51k	51k	60k
average	21.0	20.9	20.6
length	23.8	24.5	21.6

Table 3: The baseline is scored when  $n$  is 100 in Algorithm 1. We use three language pairs ENFR, ENES, and DEEN from News Commentary corpus. The size and average length of this corpus are shown.

**Sentence Removal Approach** Table 4 shows the results. The best scores are measured for different value of  $n$ .

**Good Points Approach** Step 1 of Algorithm 2 is, for a given parallel corpus, to make a word-based MT. We do this by Moses with option max-phrase-length set to 1, with the alignment option set to ‘union’ as it is high recall. Although we have chosen union, other selection may be possible. Step 2 is to obtain the cumulative n-gram score for all the training parallel corpus by using the word-based MT trained in Step 1. In Step 3 we score for all the sentence pairs. Statistics of cumulative 4, 3, 2, and 1-gram are shown in Figure 6. As is already mentioned, there are a lot of sentence pairs whose score are zero in cumulative

n	ENFR	ENES	DEEN
10	0.167	0.134	0.097
20	0.087	0.228	0.138
30	0.145	0.259	0.157
40	0.175	0.261	0.168
50	<u>0.229</u>	0.273	0.170
60	0.178	0.273	<u>0.171</u>
70	0.179	0.272	0.170
80	0.181	0.273	0.169
90	0.180	0.276	<u>0.171</u>
100	0.180	<u>0.280</u>	0.169

Table 4: Bleu score after cleaning of sentences whose length is greater than  $n$ . The row shows  $n$ , while the column shows the language pair.

4-gram score. In Step 4, based on ten configurations, we reduce our parallel corpus and check our performance.

$c$	ENFR	%	ENES	%	DEEN	%
1	0.187	49	0.297	56	0.201	40
2	0.188	55	0.294	60	0.205	49
3	0.187	61	0.301	66	0.208	58
4	0.190	82	0.306	85	0.215	83
5	<u>0.192</u>	96	<u>0.314</u>	97	<u>0.221</u>	96
6	0.180	32	0.299	56	0.192	29
7	0.162	30	0.271	25	0.174	18
8	0.179	31	0.283	35	0.186	25
9	0.167	17	0.264	20	0.177	18
10	0.152	11	0.260	20	0.155	10

Table 5: This algorithm attains 0.192 for ENFR, 0.314 for ENES, and 0.221 for DEEN. Notice these are the case when we use around 96 percents of parallel corpus. % denotes the effective ratio which can be considered to be the inlier ratio. This is equivalent to  $1 - (\text{outlier ratio})$ .

In the case of English-Spanish the configuration 5 deletes only 3.46 percents of sentences whose performance reaches 0.314 which is the best among these ten configurations. Similarly in the case of German-English the configuration 5 attains the best performance among ten configurations. The baseline system is shown in Table 4 where we picked up the best score among various selections of  $n$ . In this sense, our results for English-French are superior to nine configurations of  $n$  except one (when  $n$  is 50). Considering all of these results and baseline systems, it is possible that the outlier ratio of English-French may be big-

ger than English-Spanish and German-English. It is noted that the baseline system, as well as the 10 configurations below, uses the MERT as is already mentioned.

**Sentence Duplication Approach** Although the score will not get worse as this approach duplicates the sentence, the score is relatively good more than expected as is shown in Table 6. On the other hand, when we duplicate ten times we face with a problem that mgiza often get stuck in the middle or get the phrase extremely small numbers by unknown reasons.

$p$	ENFR	ENFR	DEEN
0.000001	–	–	0.221
0.000005	0.235	–	0.223
0.00005	0.228	0.238	0.223
0.00001	–	0.237	0.223
0.0001	–	0.235	0.216
0.001	–	0.236	0.213
duplication	10	2	10

Table 6: ENFR is shown when duplication is twice and 10 times.

**Bad Alignment Removal Approach** Results are shown in Table 7 where we use the parallel corpus obtained after Algorithm 2. All the results are worse than at the beginning.

$e$	ENFR	ENES	DEEN
1	0.191	0.299	0.217

Table 7: These results are obtained after Algorithm 2 where all the results are worse than Algorithm 2.

## 5 Conclusion and Further Work

This paper shows the preliminary results that the preprocessing of parallel corpus might be a useful for word alignment. We investigate the mechanism of Algorithm 1 and 2 as is shown in Figure 5. Our findings are that while Algorithm 1 is efficiently removes the sentences whose outlier ratio are possibly high without touching the *outliers* whose length is in high density area, i.e. between 5 and 40 word length, Algorithm 2 is honestly removes them even they resides in high density area.

By Algorithm 2, we observe two improvements of Bleu score from 28.0 to 31.4 in English-Spanish and 17.1 to 22.1 in German-English which are

shown in Table 5. By Algorithm 3, we observe several improvements of Bleu score as well although we should note that the results shown for sentence duplication part is better considered to be quite preliminary. This is because we face with several unexpected crash of a word aligner.

## References

- Banerjee, S., Lavie, A. *METEOR: An Automatic Metric for MT Evaluation With Improved Correlation With Human Judgments*. Workshop On Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005.
- Brown, F., Pietra, V.J.D, Pietra, A.D.P., Mercer, R.L.. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, Vol.19, Issue 2. 1993.
- Callison-Burch, C. *Paraphrasing and Translation*. PhD Thesis, University of Edinburgh. 2007.
- Callison-Burch, C., Koehn, P., Osborne, M. *Improved Statistical Machine Translation Using Paraphrases*. NAACL. 2006.
- Dempster, A.P., Laird, N.M., Rubin, D.B.. *Maximum likelihood from Incomplete Data via the EM algorithm*. Journal of the Royal Statistical Society. 1977.
- Deng, Y., Byrne, W. *HMM Word and Phrase Alignment for Statistical Machine Translation*. 2005.
- Doddington, G. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. HLT. 2002.
- Forsyth, D.A., Ponce, J. *Computer Vision*. Pearson Education. 2003.
- Gao, Q., Vogel, S. *Parallel Implementations of Word Alignment Tool*. 2008.
- Imamura, K., Sumita, E., Matsumoto, Y.. *Automatic Construction of Machine Translation Knowledge Using Translation Literalness*. EAACL. 2003.
- Koehn, P., Och, F.J., Marcu,D., *Statistical Phrase-Based Translation*. HLT/NAACL. 2003.
- Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D. *Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation*. International Workshop on Spoken Language Translation. 2005.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. *Moses: Open Source Toolkit for Statistical Machine Translation*. ACL. 2007.
- Liang, P., Taskar, B., Klein, D. *Alignment by agreement*. HLT/NAACL. 2006.
- Lin,D., Pantel, P. *Induction of Semantic Classes from Natural Language Text*. In Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01). 1999.
- Marcu, D. and Wong, D. *A Phrase-based, Joint Probability Model for Statistical Machine Translation*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002.
- Melamed, I.D., Green, R., Turian, J.P. *Precision and Recall of Machine Translation*. NAACL/HLT 2003. 2003.
- Moore, R.C. *A Discriminative Framework for Bilingual Word Alignment*. HLT/EMNLP. 2005.
- Och, F.J, Ney, H.. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, volume 20, number 1. 2003.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.J. *BLEU: A Method For Automatic Evaluation of Machine Translation*. ACL. 2002.
- Lambert, P. and Banchs, R. *Data Inferred Multi-word Expressions for Statistical Machine Translation*. Machine Translation Summit X. 2005.
- Snover. M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. *A Study of Translation Edit Rate with Targeted Human Annotation*. Association for Machine Translation in the Americas. 2006.
- Vogel, S., Ney, H., Tillmann, C. *HMM-based Word Alignment in Statistical Translation*. COLING 96. 1996.