# Data Cleaning for Word Alignment

**Tsuyoshi Okita**
CNGL / School of Computing
Dublin City University, Glasnevin, Dublin 9
tokita@computing.dcu.ie

## Abstract

Parallel corpora are made by human beings. However, as an MT system is an aggregation of state-of-the-art NLP technologies without any intervention of human beings, it is unavoidable that quite a few sentence pairs are beyond its analysis and that will therefore not contribute to the system. Furthermore, they in turn may act against our objectives to make the overall performance worse. Possible unfavorable items are $n : m$ mapping objects, such as paraphrases, non-literal translations, and multiword expressions. This paper presents a pre-processing method which detects such unfavorable items before supplying them to the word aligner under the assumption that their frequency is low, such as below 5 percent. We show an improvement of Bleu score from 28.0 to 31.4 in English-Spanish and from 16.9 to 22.1 in German-English.

## 1 Introduction

Phrase alignment (Marcu and Wong, 02) has recently attracted researchers in its theory, although it remains in infancy in its practice. However, a phrase extraction heuristic such as grow-diag-final (Koehn et al., 05; Och and Ney, 03), which is a single difference between word-based SMT (Brown et al., 93) and phrase-based SMT (Koehn et al., 03) where we construct word-based SMT by bi-directional word alignment, is nowadays considered to be a key process which leads to an overall improvement of MT systems. However, technically, this phrase extraction process after word alignment is known to have at least two limitations: 1) the objectives of uni-directional word alignment is limited only in $1 : n$ mappings and 2) an atomic unit of phrase pair used by phrase extraction is thus basically restricted in $1 : n$ or $n : 1$ with small exceptions.

Firstly, the posterior-based approach (Liang, 06) looks at the posterior probability and partially delays the alignment decision. However, this approach does not have any extension in its $1 : n$ uni-directional mappings in its word alignment. Secondly, the aforementioned phrase alignment (Marcu and Wong, 02) considers the $n : m$ mapping directly bilingually generated by some concepts without word alignment. However, this approach has severe computational complexity problems. Thirdly, linguistic motivated phrases, such as a tree aligner (Tinsley et al., 06), provides $n : m$ mappings using some information of parsing results. However, as the approach runs somewhat in a reverse direction to ours, we omit it from the discussion. Hence, this paper will seek for the methods that are different from those approaches and whose computational cost is cheap.

$n : m$ mappings in our discussion include paraphrases (Callison-Burch, 07; Lin and Pantel, 01), non-literal translations (Imamura et al., 03), multiword expressions (Lambert and Banchs, 05), and some other noise in one side of a translation pair (from now on, we call these 'outliers', meaning that these are not systematic noise). One common characteristic of these $n : m$ mappings is that they tend to be so flexible that even an exhaustive list by human beings tends to be incomplete (Lin and Pantel, 01). There are two cases which we should like to distinguish: when we use external resources and when we do not. For example, Quirk et al. employ external resources by drawing pairs of English sentences from a comparable corpus (Quirk et al., 04), while Bannard and Callison-Burch (Bannard and Callison-Burch, 05) identified English paraphrases by pivoting through phrases in another language. However, in this paper our interest is rather the case when our resources are limited within our parallel corpus.



*Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 72–80,
Suntec, Singapore, 4 August 2009. 

Imamura et al. (Imamura et al., 03), on the other hand, do not use external resources and present a method based on literalness measure called TCR (Translation Correspondence Rate). Let us define literal translation as a word-to-word translation, and non-literal translation as a non word-to-word translation. Literalness is defined as a degree of literal translation. Literalness measure of Imamura et al. is trained from a parallel corpus using word aligned results, and then sentences are selected which should either be translated by a 'literal translation' decoder or by a 'non-literal translation' decoder based on this literalness measure. Apparently, their definition of literalness measure is designed to be high recall since this measure incorporates all the possible correspondence pairs (via realizability of lexical mappings) rather than all the possible true positives (via realizability of sentences). Adding to this, the notion of literal translation may be broader than this. For example, literal translation of "C'est la vie." in French is "That's life." or "It is the life." in English. If literal translation can not convey the original meaning correctly, non-literal translation can be applied: "This is just the way life is.", "That's how things happen.", "Love story.", and so forth. Non-literal translation preserves the original meaning[1] as much as possible, ignoring the exact word-to-word correspondence. As is indicated by this example, the choice of literal translation or non-literal translation seems rather a matter of translator preference.

This paper presents a pre-processing method using the alternative literalness score aiming for high precision. We assume that the percentages of these $n : m$ mappings are relatively low. Finally, it turned out that if we focus on outlier ratio, this method becomes a well-known sentence cleaning approach. We refer to this in Section 5.

This paper is organized as follows. Section 2 outlines the $1 : n$ characteristics of word alignment by IBM Model 4. Section 3 reviews an atomic unit of phrase extraction. Section 4 explains our Good Points Algorithm. Experimental results are presented in Section 5. Section 6 discusses a sentence cleaning algorithm. Section 7 concludes and provides avenues for further research.

[1]Dictionary goes as follows: something that you say when something happens that you do not like but which you have to accept because you cannot change it [Cambridge Idioms Dictionary 2nd Edition, 06].

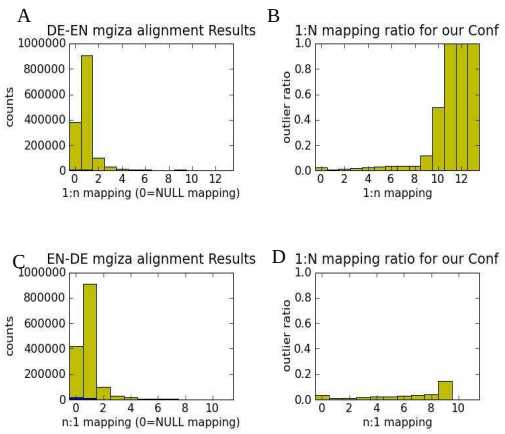


Figure 1: Figures A and C show the results of word alignment for DE-EN where outliers detected by Algorithm 1 are shown in blue at the bottom. We check all the alignment cept pairs in the training corpus inspecting so-called A3 final files by type of alignment from 1:1 to 1:13 (or NULL alignment). It is noted that outliers are miniscule in A and C because each count is only 3 percent. Most of them are NULL alignment or 1:1 alignment, while there are small numbers of alignments with 1:3 and 1:4 (up to 1:13 in the DE-EN direction in Figure A). In Figure C, 1:11 is the greatest. Figure B and D show the ratio of outliers over all the counts. Figure B shows that in the case of 1:10 alignments, 1/2 of the alignments are considered to be outliers by Algorithm 1, while 100 percent of alignment from 1:11 to 1:13 are considered to be outliers (false negative). Figure D shows that in the case of EN-DE, most of the outlier ratios are less than 20 percent.

## 2 $1 : n$ Word Alignment

Our discussion of uni-directional alignments of word alignment is limited 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 sentence length of $e_i$, and similarly for $|f_i|$. We are given a pair of sentence aligned bilingual texts $(f_1, e_1), \ldots, (f_n, e_n) \in \mathcal{X} \times \mathcal{Y}$, where $f_i = (\bar{f}_{i,1}, \ldots, \bar{f}_{i,|f_i|})$ and $e_i = (\bar{e}_{i,1}, \ldots, \bar{e}_{i,|e_i|})$. 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 $p_{\bar{f}_i} : \bar{e}_i \rightarrow p_{\bar{f}_j}(\bar{e}_i)$ such that $\Sigma p_{\bar{f}_j}(\bar{e}_i) = 1$ and $\forall \bar{e}_i : 0 \leq p_{\bar{f}_j}(\bar{e}_i) \leq 1$ (It is noted that some models such*

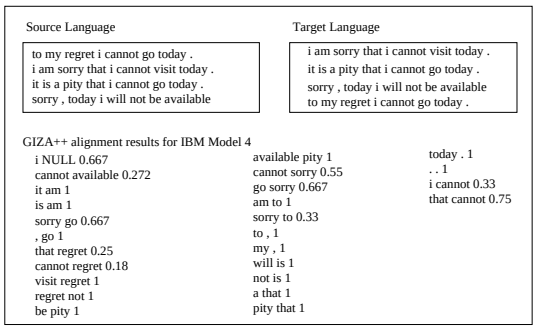


Figure 2: Example shows an example alignment of paraphrases in a monolingual case. Source and target use the same set of sentences. Results show that only the matching between the colon is correct[3].

*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.*

As the above definition shows, the purpose of the word alignment task is to obtain a lexical translation probability $p(\bar{f}_i|\bar{e}_i)$, which is a $1:n$ uni-directional word alignment. The initial idea underlying the IBM Models, consisting of five distinctive models, is that it introduces an alignment function $a(j|i)$, or alternatively the distortion function $d(j|i)$ or $d(j-\odot_i)$, when the task is viewed as a missing value problem, where $i$ and $j$ denote the position of a cept in a sentence and $\odot_i$ denotes the center of a cept. $d(j|i)$ denotes a distortion of the absolute position, while $d(j-\odot_i)$ denotes the distortion of relative position. Then this missing value problem can be solved by EM algorithms : E-step is to take expectation of all the possible alignments and M-step is to estimate maximum likelihood of parameters by maximizing the expected likelihood obtained in the E-step. The second idea of IBM Models is in the mechanism of fertility and a NULL insertion, which makes the performance of IBM Models competitive. Fertility and a NULL insertion is used to adjust the length $n$ when the length of the source sentence is different from this $n$. Fertility is a mechanism to augment one source word into several source words or delete a source word, while a NULL insertion is a mechanism of generating several words from blank words. Fertility uses a conditional probability depending only on the lexicon. For example, the length of 'today' can be conditioned only on the lexicon 'today'.

As is already mentioned, the resulting alignments are $1:n$ (shown in the upper figure in Figure 1). For DE-EN News Commentary corpus, most of the alignments fall in either 1:1 mapping or NULL mappings whereas small numbers are 1:2 mappings and miniscule numbers are from 1:3 to 1:13. However, this $1:n$ nature of word alignment will cause problems if we encounter $n:m$ mapping objects, such as a paraphrase, non-literal translation, or multiword expression. Figure 2 shows such difficulties where we show a monolingual paraphrase. Without loss of generality this can be easily extended to bilingual paraphrases. In this case, results of word alignment are completely wrong, with the exception of the example consisting of a colon. Although these paraphrases, non-literal translations, and multiword expressions do not always become outliers, they may face the potential danger of producing the incorrect word alignments with incorrect probabilities.

## 3 Phrase Extraction and Atomic Unit of Phrases

The phrase extraction is a process to exploit phrases for a given bi-directional word alignment (Koehn et al., 05; Och and Ney, 03). If we focus on its generative process, this would become as follows: 1) add intersection of two word alignments as an alignment point, 2) add new alignment points that exist in the union with the constraint that a new alignment point connects at least one previously unaligned word, 3) check the unaligned row (or column) as unaligned row (or column, respectively), 4) if $n$ alignment points are contiguous in horizontal (or vertical) direction we consider that this is a contiguous $1:n$ (or $n:1$) phrase pair (let us call these type I phrase pairs), 5) if a neighborhood of a contiguous $1:n$ phrase pair is (an) unaligned row(s) or (an) unaligned column(s) we grow this region (with consistency constraint) (let us call these type II phrase pair), and 6) we consider all the diagonal combinations of type I and

[3]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 the $1:n$ mapping, it would be quite lucky if we obtain $n:m$ mapping after phrase extraction (Our focus is not on the incorrect probability, but rather on the incorrect matching.)

type II phrase pairs generatively.

The atomic unit of type I phrase pairs is $1:n$ or $n:1$, while that of type II phrase pairs is $n:m$ if unaligned row(s) and column(s) exist in neighborhood. So, whether they form a $n:m$ mapping or not depends on the existence of unaligned row(s) and column(s). And at the same time, $n$ or $m$ should be restricted to a small value. There is a chance that a $n:m$ phrase pair can be created in this way. This is because around one third of word alignments, which is quite a large figure, are $1:0$ as is shown in Figure 1. Nevertheless, our concern is if the results of word alignment is very low quality, e.g. similar to the situation depicted in Figure 2, this mechanism will not work. Furthermore, this mechanism is only restricted in the unaligned row(s) and column(s).

## 4 Our Approach: Good Points Approach

Our approach aims at removing *outliers* by the literalness score, which we defined in Section 1, between a pair of sentences. Sentence pairs with low literalness score 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 an ideal MT system $M_{PB}$, a paraphrase is an inlier (or realizable), and*

**Proposition 2** *Under an ideal 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, what we should do is, initially, to score it under a word-based MT $M_{WB}$ using Bleu, for example. (Later we replace it with a variant of Bleu, i.e. cumulative n-gram score). However, despite Proposition 1, our MT system at hand is unfortunately not ideal. What we can currently do is the following: if we witness bad sentence-based scores in word-based MT, we can consider our MT system failing to incorporating a $n:m$ mapping object for those sentences. Later in our revised version, we use both of word-based MT and phrase-based MT. The summary of our first 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

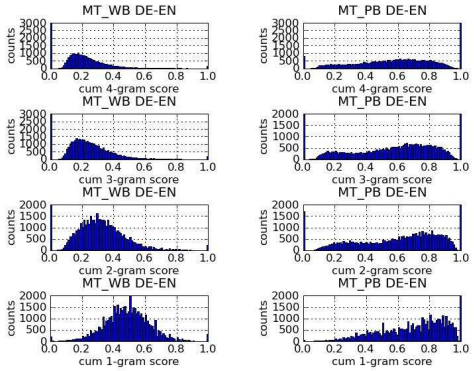


Figure 3: Left figure shows sentence-based Bleu score of word-based SMT and right figure shows that of phrase-based SMT. Each row shows the cumulative n-gram score (n = 1,2,3,4) and we use News Commentary parallel corpus (DE-EN).

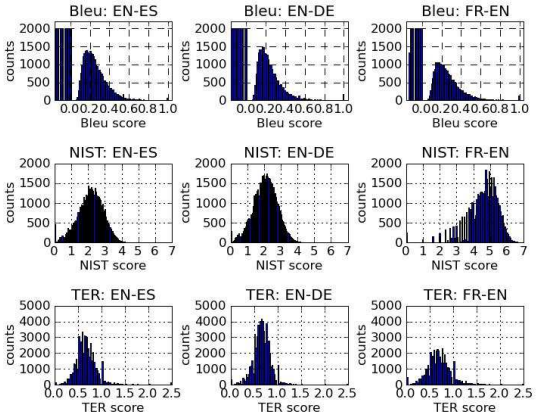


Figure 4: Each row shows Bleu, NIST, and TER, while each column shows different language pairs (EN-ES, EN-DE and FR-DE). These figures show the scores of all the training sentences by the word-based SMT system. In the row for Bleu, note that the area of rectangle shows the number of sentence pairs whose 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 25,000 to 30,000.) There is a smooth probability distribution in the middle, while there are two non-smoothed connections at 1.0 and 0.0. Notice there is a small number 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 scores more than 2.5 in these figures. (The maximum was 27.0.)

of Bleu score as the measure of literalness, and 3) based on this score, we reduce the sentences in parallel corpus. Our algorithm is as follows:

**Algorithm 1** Good Points Algorithm

Step 1: Train word-based MT.
Step 2: Translate all training sentences 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 Table 1, we produce new reduced parallel corpus.
(Step 5: Do the whole procedure of phrase-based SMT using the reduced parallel corpus which we obtain from Step 1 to 4.)

| conf | A1 | A2 | A3 | A4 |
|---|---|---|---|---|
| Ours | 0.05 | 0.05 | 0.1 | 0.2 |
| 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 | 0.22 | 0.3 | 0.4 | 0.6 |
| 6 | 0.25 | 0.4 | 0.5 | 0.7 |
| 7 | 0.2 | 0.4 | 0.5 | 0.8 |
| 8 | | | | 0.6 |

Table 1: Table shows our threshold where A1, A2, A3, and A4 correspond to the absolute cumulative n-gram precision value (n=1,2,3,4 respectively). In experiments, we compare ours with eight configurations above in Table 6.

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

Table 2: Sentences judged as outliers by Algorithm 1 (ENFR News Commentary corpus).

We would like to mention our motivation for choosing 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

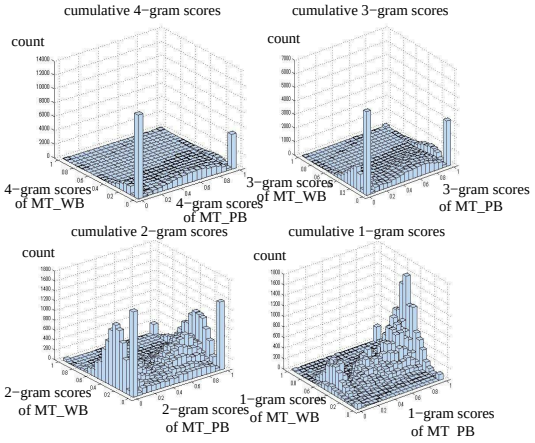


Figure 5: Four figures show the sentence-based cumulative n-gram scores: x-axis is phrase-based SMT and y-axis is word-based SMT. Focus is on the worst point (0,0) where both scores are zero. Many points reside in (0,0) in cumulative 4-gram scores, while only small numbers of point reside in (0,0) in cumulative 1-gram scores.

in the first row of Figure 4, typical distribution of words in this space $M_{WB}$ is separated in two clusters: one looks like a geometric distribution and the other one contains a lot of points whose value is zero. (Especially in the case of Bleu, if the sentence length is less than 3 the Bleu score is zero.) For this reason, we use the variants of Bleu score: we decompose Bleu score in cumulative n-gram score (n=1,2,3,4), which is shown in Figure 3. It is noted that the following relation holds: $S_4(e, f) \leq S_3(e, f) \leq S_2(e, f) \leq S_1(e, f)$ where $e$ denotes an English sentence, $f$ denotes a foreign sentence, and $S_X$ denotes cumulative $X$-gram scores. For 3-gram scores, the tendency to separate in two clusters is slightly decreased. Furthermore, for 1-gram scores, the distribution approaches to normal distribution. We model P(outlier) taking care of the quantity of $S_2(e, f)$, where we choose 0.1: other configurations in Table 1 are used in experiments. It is noted that although we choose the variants of Bleu score, it is clear, in this context, that we can replace Bleu with any other measure, such as METEOR (Banerjee and Lavie, 05), NIST (Doddington, 02), GTM (Melamed et al., 03), TER (Snover et al., 06), labeled dependency approach (Owczarzak et al., 07) and so forth (see Figure 4). Table 2 shows outliers detected by Algorithm 1.

Finally, a revised algorithm which incorporates sentence-based $X$-gram scores of phrase-based MT is shown in Algorithm 2. Figure 5 tells us

that there are many sentence pair scores actually improved in phrase-based MT even if word-based score is zero.

**Algorithm 2** Revised Good Points Algorithm

Step 1: Train word-based MT for full parallel corpus. Translate all training sentences by the above trained word-based MT decoder.

Step 2: Obtain the cumulative $X$-gram score $S_{WB,X}$ for each pair of sentences where $X$ is 4, 3, 2, and 1 for word-based MT decoder.

Step 3: Train phrase-based MT for full parallel corpus. Note that we do not need to run a word aligner again in here, but use the results of Step 1. Translate all training sentences by the above trained phrase-based MT decoder.

Step 4: Obtain the cumulative $X$-gram score $S_{PB,X}$ for each pair of sentences where $X$ is 4, 3, 2, and 1 for phrase-based MT decoder.

Step 5: Remove sentences whose $(S_{WB,2}, S_{PB,2}) = (0, 0)$. We produce new reduced parallel corpus.

(Step 6: Do the whole procedure of phrase-based SMT using the reduced parallel corpus which we obtain from Step 1 to 5.)

## 5 Results

We evaluate our algorithm using the News Commentary parallel corpus used in 2007 Statistical Machine Translation Workshop shared task (corpus size and average sentence length are shown in Table 8). We use the devset and the evaluation set

| alignment | ENFR | ESEN |
|---|---|---|
| grow-diag-final | 0.058 | 0.115 |
| union | 0.205 | 0.116 |
| intersection | 0.164 | 0.116 |

Table 3: Performance of word-based MT system in different alignment methods. The above is between ENFR and ESEN.

| pair | ENFR | FREN |
|---|---|---|
| score | 0.205 | 0.176 |
| ENES | ENDE | DEEN |
| 0.276 | 0.134 | 0.208 |

Table 4: Performance of word-based MT system for different language pairs with union alignment method.

provided by this workshop. We use Moses (Koehn et al., 07) as the baseline system, with mgiza (Gao and Vogel, 08) as its word alignment tool. We do MERT in all the experiments below.

Step 1 of Algorithm 1 produces, for a given parallel corpus, a word-based MT. We do this using Moses with option max-phrase-length set to 1, alignment as union as we would like to extract the bi-directional results of word alignment with high recall. Although we have chosen union, other selection options may be possible as Table 3 suggests. Performance of this word-based MT system is as shown in Table 4.

Step 2 is to obtain the cumulative n-gram score for the entire training parallel corpus by using the word-based MT system trained in Step 1. Table 5 shows the first two sentences of News Commentary corpus. We score for all the sentence pairs.

| |
|---|
| c_score = [0.4213,0.4629,0.5282,0.6275]<br>consider the number of clubs that have qualified for the european champions ' league top eight slots .<br>considérons le nombre de clubs qui se sont qualifiés parmi les huit meilleurs de la ligue des champions europenne . |
| c_score = [0.0000,0.0000,0.0000,0.3298]<br>estonia did not need to ponder long about the options it faced .<br>l' estonie n' a pas eu besoin de longuement rflchir sur les choix qui s' offraient à elle . |

Table 5: Four figures marked as score shows the cumulative n-gram score from left to right. The following EN and FR are the calculated sentences used by word-based MT system trained on Step 1.

In Step 3, we obtain the cumulative $n$-gram score (shown in Figure 3). As is already mentioned, there are a lot of sentence pairs whose cumulative 4-gram score is zero. In the cumulative 3-gram score, this tendency is slightly decreased. For 1-gram scores, the distribution approaches to normal distribution. In Step 4, other than our configuration we used 8 different configurations in Table 6 to reduce our parallel corpus.

Now we obtain the reduced parallel corpus. In Step 5, using this reduced parallel corpus we carried out training of MT system from the beginning: we again started from the word alignment, followed by phrase extraction, and so forth. The results corresponding to these configurations are shown in Table 6. In Table 6, in the case of

| ENES | Bleu | effective sent | UNK |
|---|---|---|---|
| Base | 0.280 | 99.30 % | 1.60% |
| Ours | 0.314 | 96.54% | 1.61% |
| 1 | 0.297 | 56.21% | 2.21% |
| 2 | 0.294 | 60.37% | 2.09% |
| 3 | 0.301 | 66.20% | 1.97% |
| 4 | 0.306 | 84.60% | 1.71% |
| 5 | 0.299 | 56.12% | 2.20% |
| 6 | 0.271 | 25.05% | 2.40% |
| 7 | 0.283 | 35.28% | 2.26% |
| 8 | 0.264 | 19.78% | 4.22% |

| | DEEN | % | ENFR | % |
|---|---|---|---|---|
| Base | 0.169 | 99.10% | 0.180 | 91.81% |
| Ours | 0.221 | 96.42% | 0.192 | 96.38% |
| 1 | 0.201 | 40.49% | 0.187 | 49.37% |
| 2 | 0.205 | 48.53% | 0.188 | 55.03% |
| 3 | 0.208 | 58.07% | 0.187 | 61.22% |
| 4 | 0.215 | 83.10% | 0.190 | 81.57% |
| 5 | 0.192 | 29.03% | 0.180 | 31.52% |
| 6 | 0.174 | 17.69% | 0.162 | 29.97% |
| 7 | 0.186 | 24.60% | 0.179 | 30.52% |
| 8 | 0.177 | 18.29% | 0.167 | 17.11% |

Table 6: Table shows Bleu score for ENES, DEEN, and ENFR: 0.314, 0.221, and 0.192, respectively. All of these are better than baseline. Effective ratio can be considered to be the inlier ratio, which is equivalent to 1 - (outlier ratio). The details for the baseline system are shown in Table 8.

| ENES | Bleu | effective sent |
|---|---|---|
| Base | 0.280 | 99.30 % |
| Ours | 0.317 | 97.80 % |
| DEEN | Bleu | effective sent |
| Base | 0.169 | 99.10 % |
| Ours | 0.218 | 97.14 % |

Table 7: This table shows results for the revised Good Points Algorithm.

English-Spanish our configuration discards 3.46 percent of sentences, and the performance reaches 0.314 which is the best among other configurations. Similarly in the case of German-English our configuration attains the best performance among configurations. It is noted that results for the baseline system are shown in Table 8 where we picked up the score where $n$ is 100. It is noted that the baseline system as well as other configurations use MERT. Similarly, results for a revised Good Points Algorithm is shown in Table 7.

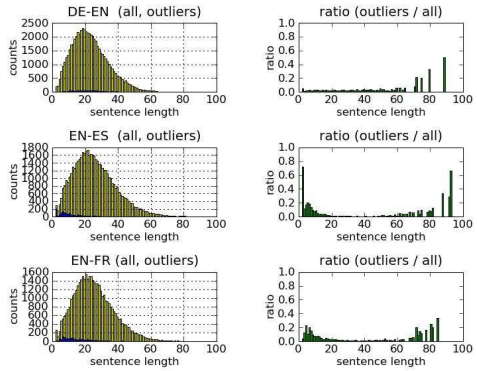


Figure 6: Three figures in the left show the histogram of sentence length (main figures) and histogram of sentence length of outliers (at the bottom). (As the numbers of outliers are less than 5 percent in each case, outliers are miniscule. In the case of EN-ES, we can observe the blue small distributions at the bottom from 2 to 16 sentence 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 percent from 80 to 100 sentence length. The lower two figures show that sentence length 1 to 4 tend to be more than 10 percent.

## 6 Discussion

In Section 1, we mentioned that if we aim at outlier ratio using the indirect feature *sentence length*, this method reduces to a well-known sentence cleaning approach shown below in Algorithm 3.

**Algorithm 3** Sentence Cleaning Algorithm

Remove sentences with lengths greater than $X$ (or remove sentences with lengths smaller than $X$ in the case of short sentences).

This approach is popular although the reason behind why this approach works is not well understood. Our explanation is shown in the right-hand side of Figure 6 where outliers are shown at the bottom (almost invisible) which are extracted by Algorithm 1. The region that Algorithm 3 removes via sentence length $X$ is possibly the region where the ratio of outliers is high.

This method is a high recall method. This method does not check whether the removed sentences are really sentences whose behavior is bad or not. For example, look at Figure 6 for sen-

| X | ENFR | FREN | ESEN | DEEN | ENDE |
|---|---|---|---|---|---|
| 10 | 0.167 | 0.088 | 0.143 | 0.097 | 0.079 |
| 20 | 0.087 | 0.195 | 0.246 | 0.138 | 0.127 |
| 30 | 0.145 | 0.229 | 0.279 | 0.157 | 0.137 |
| 40 | 0.175 | 0.242 | 0.295 | 0.168 | 0.142 |
| 50 | 0.229 | 0.250 | 0.297 | 0.170 | 0.145 |
| 60 | 0.178 | 0.253 | 0.297 | 0.171 | 0.146 |
| 70 | 0.179 | 0.251 | 0.298 | 0.170 | 0.146 |
| 80 | 0.181 | 0.252 | 0.301 | 0.169 | 0.147 |
| 90 | 0.180 | 0.252 | 0.297 | 0.171 | 0.147 |
| 100 | 0.180 | 0.251 | 0.302 | 0.169 | 0.146 |
| # | 51k | 51k | 51k | 60k | 60k |
| ave len | 21.0/23.8(EN/FR) 20.9/24.5(EN/ES) 20.6/21.6(EN/DE) | | | | |

Table 8: Bleu score after cleaning of sentences with length greater than $X$. The row shows $X$, while the column shows the language pair. Parallel corpus is News Commentary parallel corpus. It is noted that the default setting of MAX_SENTENCE_LENTH_ALLOWED in GIZA++ is 101.

tence length 10 to 30 where there are considerably many outliers in the region that a lot of inliers reside. However, this method cannot cope with such outliers. Instead, the method cope with the region that the outlier ratio is possibly high at both ends, e.g. sentence length $> 60$ or sentence length $< 5$. The advantage is that sentence length information is immediately available from the sentence which is easy to implement. The results of this algorithm is shown in Table 8 where we varies $X$ and language pair. This table also suggests that we should refrain from saying that $X = 60$ is best or $X = 80$ is best.

## 7 Conclusions and Further Work

This paper shows some preliminary results that data cleaning may be a useful pre-processing technique for word alignment. At this moment, we observe two positive results, improvement of Bleu score from 28.0 to 31.4 in English-Spanish and 16.9 to 22.1 in German-English which are shown in Table 6. Our method checks the realizability of target sentences in training sentences. If we witness bad cumulative $X$-gram scores we suspect that this is due to some problems caused by the $n : m$ mapping objects during word alignment followed by phrase extraction process.

Firstly, although we removed training sentences whose $n$-gram scores are low, we can duplicate such training sentences in word alignment. This method is appealing, but unfortunately if we use mgiza or GIZA++, our training process often ceased in the middle by unrecognized errors. However, if we succeed in training, the results often seem comparable to our results. Although we did not supply back removed sentences, it is possible to examine such sentences using the T-tables to extract phrase pairs.

Secondly, it seems that one of the key matters lies in the quantities of $n : m$ mapping objects which are difficult to learn by word-based MT (or by phrase-based MT). It is possible that such quantities are different depending on their language pairs and on their corpora size. A rough estimation is that this quantity may be somewhere less than 10 percent (in FR-EN Hansard corpus, recall and precision reach around 90 percent (Moore, 05)), or less than 5 percent (in News Commentary corpus, the best Bleu scores by Algorithm 1 are when this percentage is less than 5 percent ). As further study, we intend to examine this issue further.

Thirdly, this method has other aspects that it removes discontinuous points: such discontinuous points may relate to the smoothness of optimization surface. One of the assumptions of the method such as Wang et al. (Wang et al., 07) relates to smoothness. Then, our method may improve their results, which is our further study.

In addition, although our algorithm runs a word aligner more than once, this process can be reduced since removed sentences are less than 5 percent or so.

Finally, we did not compare our method with TCR of Imamura. In our case, the focus was 2-gram scores rather than other $n$-gram scores. We intend to investigate this further.

## 8 Acknowledgements

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142). Thanks to Yvette Graham and Sudip Naskar for proof reading, Andy Way, Khalil Sima'an, Yanjun Ma, and annonymous reviewers for comments, and Machine Translation Marathon.

## References

Colin Bannard and Chris Callison-Burch. 2005. *Paraphrasing with bilingual parallel corpora*. ACL.

Satanjeev Banerjee and Alon Lavie. 2005. *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.

Peter F. Brown, Vincent J.D. Pietra, Stephen A.D. Pietra, and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics, Vol.19, Issue 2.*

Chris Callison-Burch. 2007. *Paraphrasing and Translation.* PhD Thesis, University of Edinburgh.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. *Improved Statistical Machine Translation Using Paraphrases.* NAACL.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapala. 2008. *ParaMetric: An Automatic Evaluation Metric for Paraphrasing.* COLING.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. *Maximum likelihood from Incomplete Data via the EM algorithm.* Journal of the Royal Statistical Society.

Yonggang Deng and William Byrne. 2005. *HMM Word and Phrase Alignment for Statistical Machine Translation.* Proc. Human Language Technology Conference and Empirical Methods in Natural Language Processing.

George Doddington. 2002. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.* HLT.

David A. Forsyth and Jean Ponce. 2003. *Computer Vision.* Pearson Education.

Qin Gao and Stephan Vogel. 2008. *Parallel Implementations of Word Alignment Tool.* Software Engineering, Testing, and Quality Assurance for Natural Language Processing.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003. *Automatic Construction of Machine Translation Knowledge Using Translation Literalness.* EACL.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. *Statistical Phrase-Based Translation.* HLT/NAACL.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. *Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation.* International Workshop on Spoken Language Translation.

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.* ACL.

Patrik Lambert and Rafael E. Banchs. 2005. *Data Inferred Multiword Expressions for Statistical Machine Translation.* Machine Translation Summit X.

Percy Liang, Ben Taskar, and Dan Klein. 2006. *Alignment by agreement.* HLT/NAACL.

Dekang Lin and Patrick Pantel. 1999. *Induction of Semantic Classes from Natural Language Text.* In Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01).

Daniel Marcu and William Wong. 2002. *A Phrase-based, Joint Probability Model for Statistical Machine Translation.* In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

I. Dan Melamed, Ryan Green, and Joseph Turian. 2003. *Precision and Recall of Machine Translation.* NAACL/HLT 2003.

Robert C. Moore. 2005. *A Discriminative Framework for Bilingual Word Alignment.* HLT/EMNLP.

Franz Josef Och and Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, volume 20,number 1.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. *Evaluating Machine Translation with LFG Dependencies.* Machine Translation, Springer, Volume 21, Number 2.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A Method For Automatic Evaluation of Machine Translation* ACL.

Chris Quirk, Chris Brockett, and William Dolan. 2004. *Monolingual machine translation for paraphrase generation.* EMNLP-2004.

Matthew Snover. Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. *A Study of Translation Edit Rate with Targeted Human Annotation.* Association for Machine Translation in the Americas.

John Tinsley, Ventsisiav Zhechev, Mary Hearne, and Andy Way. 2006. *Robust Language Pair-Independent Sub-Tree Alignment.* Translation Summit XI.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. *HMM-based Word Alignment in Statistical Translation.* COLING 96.

Zhuoran Wang, John Shawe-Taylor, and Sandor Szedmak. 2007. *Kernel Regression Based Machine Translation.* Proceedings of NAACL-HLT 2007.