Training Data Modification for SMT

Considering Groups of Synonymous Sentences

Hideki KASHIOKA
Spoken Language Communication Research Laboratories, ATR
2-2-2 Hikaridai, Keihanna Science City
Kyoto, 619-0288, Japan
hideki.kashioka@atr.jp

Abstract

Generally speaking, statistical machine translation systems would be able to attain better performance with more training sets. Unfortunately, well-organized training sets are rarely available in the real world. Consequently, it is necessary to focus on modifying the training set to obtain high accuracy for an SMT system. If the SMT system trained the translation model, the translation pair would have a low probability when there are many variations for target sentences from a single source sentence. If we decreased the number of variations for the translation pair, we could construct a superior translation model. This paper describes the effects of modification on the training corpus when consideration is given to synonymous sentence groups. We attempt three types of modification: compression of the training set, replacement of source and target sentences with a selected sentence from the synonymous sentence group, and replacement of the sentence on only one side with the selected sentence from the synonymous sentence group. As a result, we achieve improved performance with the replacement of source-side sentences.

1 Introduction

Recently, many researchers have focused their interest on statistical machine translation (SMT) systems, with particular attention given to models and decoding algorithms. The quantity of the training corpus has received less attention, although of course the earlier reports do address the quantity issue. In most cases, the larger the training corpus becomes, the higher accuracy is achieved. Usually, the quantity problem of the training corpus is discussed in relation to the size of the training corpus and system performance; therefore, researchers study line graphs that indicate the relationship between accuracy and training corpus size.

On the other hand, needless to say, a single sentence in the source language can be used to translate several sentences in the target language. Such various possibilities for translation make MT system development and evaluation very difficult. Consequently, here we employ multiple references to evaluate MT systems like BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). Moreover, such variations in translation have a negative effect on training in SMT because when several sentences of input-side language are translated into the exactly equivalent output-side sentences, the probability of correct translation decreases due to the large number of possible pairs of expressions. Therefore, if we can restrain or modify the training corpus, the SMT system might achieve high accuracy.

As an example of modification, different output-side sentences paired with the exactly equivalent input-side sentences are replaced with one target sentence. These sentence replacements are required for synonymous sentence sets. Kashioka (2004) discussed synonymous sets of sentences. Here, we employ a method to group them as a way of modifying the training corpus for use with SMT. This paper focuses on how to control the corpus while giving consideration to synonymous sentence groups.

Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 19–24,
Ann Arbor, June 2005.

2 Target Corpus

In this paper, we use a multilingual parallel corpus called BTEC (Takezawa et al., 2002) for our experiments. BTEC was used in IWSLT (Akiba et al., 2004). This parallel corpus is a collection of Japanese sentences and their translations into English, Korean and Chinese that are often found in phrase books for foreign tourists. These parallel sentences cover a number of situations (e.g., hotel reservations, troubleshooting) for Japanese going abroad, and most of the sentences are rather short. Since the scope of its topics is quite limited, some very similar sentences can be found in the corpus, making BTEC appropriate for modification with compression or replacement of sentences. We use only a part of BTEC for training data in our experiments. The training data we employ contain 152,170 Japanese sentences, with each sentence combined with English and Chinese translations. In Japanese, each sentence has 8.1 words on average, and the maximum sentence length is 150 words. In English, each sentence contains an average of 7.4 words, with a maximum sentence length of 117 words. In Chinese, each sentence has an average of 6.7 words and maximum length of 122 words. Some sentences appear twice or more in the training corpus. In total, our data include 94,268 different Japanese sentences, 87,061 different Chinese sentences, and 91,750 different English sentences. Therefore, there are some sentence pairs that consist of exactly the same sentence in one language but a different sentence in another language, as Fig. 1 shows. This relationship can help in finding the synonymous sentence group.

The test data contain 510 sentences from different training sets in the BTEC. Each source sentence in the test data has 15 target sentences for evaluations. For the evaluation, we do not use any special process for the grouping process. Consequently, our results can be compared with those of other MT systems.

S1 ⇔ T1
S2 ⇔ T1
S1 ⇔ T2
S3 ⇔ T1

Figure 1. Sample sentence pairs

3 Modification Method

When an SMT system learns the translation model, variations in the translated sentences of the pair are critical for determining whether the system obtains a good model. If the same sentence appears twice in the input-side language and these sentences form pairs with two different target sentences in the output-side language, then broadly speaking the translation model defines almost the same probability for these two target sentences.

In our model, the translation system features the ability to generate an output sentence with some variations; however, for the system to generate the most appropriate output sentence, sufficient information is required. Thus, it is difficult to prepare a sufficiently large training corpus.

3.1 Synonymous Sentence Group

Kashioka (2004) reported two steps for making a synonymous sentence group. The first is a concatenation step, and the second is a decomposition step. In this paper, to form a synonymous sentence group, we performed only the concatenation step, which has a very simple idea. When the expression "Exp_A_1" in language A is translated into the expressions "Exp_B_1, Exp_B_2, ..., Exp_B_n" in language B, that set of expressions form one synonymous group. Furthermore, when the sentence "Exp_A_2" in language A is translated into the sentences "Exp_B_1, Exp_$B_{n+1,}$ ..., Exp_B_m" in language B, "Exp_B_1, Exp_$B_{n+1,}$ ..., Exp_B_m ($n < m$)" form one synonymous group. In this situation, "Exp_A_1" and "Exp_A_2" form a synonymous group because both "Exp_A_1" and "Exp_A_2" have a relationship with the translation pairs of "Exp_B_1." Thus, "Exp_$A_{1,}$ Exp_A_2" in language A and "Exp_$B_{1,}$ ..., Exp_B_m" in language B form a synonymous group. If other language information is available, we can extend this synonymous group using information on translation pairs for other languages.

In this paper, we evaluate an EJ/JE system and a CJ/JC system, and our target data include three languages, i.e., Japanese, English, and Chinese. We make synonymous sentence groups in two different environments. One is a group using Japanese and English data, and other is a group that uses Japanese and Chinese data.

The JE group contained 72,808 synonymous sentence groups, and the JC group contained 83,910 synonymous sentence groups as shown in Table 1.

	# of Groups	# of Sent per Group
JE	72,808	2.1
JC	83,910	1.8

Table 1 Statistics used in BTEC data

3.2 Modification

We prepared the three types of modifications for training data.

1. Compress the training corpus based on the synonymous sentence group (Fig. 2).
2. Replace the input and output sides' sentences with the selected sentence, considering the synonymous sentence group (Fig. 3).
3. Replace one side's sentences with a selected sentence, considering the synonymous sentence group (Figs. 4, 5).

We describe these modifications in more detail in the following subsections.

3.2.1 Modification with Compression

Here, a training corpus is constructed with several groups of synonymous sentences. Then, each group keeps only one pair of sentences and the other pairs are removed from each group, thereby decreasing the total number of sentences and narrowing the variation of expressions. Figure 2 shows an example of modification in this way. In the figure, S1, S2, and S3 indicate the input-side sentences while T1 and T2 indicate the output-side sentences. The left-hand side box shows a synonymous sentence group in the original training corpus, where four sentence pairs construct one synonymous sentence group. The right-hand side box shows a part of the modified training corpus. In this case, we keep the S1 and T1 sentences, and this resulting pair comprises a modified training corpus.

The selection of what sentences to keep is an important issue. In our current experiment, we select the most frequent sentence in each side's language from within each group. In Fig. 2, S1 appeared twice, while S2 and S3 appeared only once in the input-side language. As for the output-side language, T1 appeared three times and T2 appeared once. Thus, we keep the pair consisting of S1 and T1. When attempting to separately select the most frequent sentence in each language, we may not find suitable pairs in the original training corpus; however, we can make a new pair with the extracted sentences for the modified training corpus.

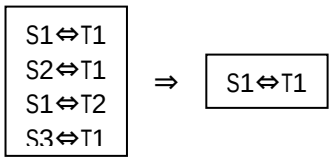

Figure 2. Modification sample for compression

3.2.2 Modification of replacing the sentences of both sides

In the compression stage, the total number of sentences in the modified training corpus is decreased, and it is clear that fewer sentences in the training corpus leads to diminished accuracy. In order to make a comparison between the original training corpus and a modified training corpus with the same number of sentences, we extract one pair of sentences from each group, and each pair appears in the modified training corpus in the same number of sentences. Figure 3 shows an example of this modification. The original training data are the same as in Fig. 2. Then we extract S1 and T1 by the same process from each side with this group, and replacing all of the input-side sentences with S1 in this group. The output side follows the same process. In this case, the modified training corpus consists of four pairs of S1 and T1.

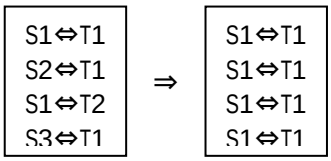

Figure 3. Sample modifications for replacement of both sentences

3.2.3 Modification to replace only one side's sentence

With the previous two modifications, the language variations in both sides decrease. Next, we propose the third modification, which narrows the range of one side's variations.

The sentences of one side are replaced with the selected sentence from that group. The sentence for replacement is selected by following the same process used in the previous modifications. As a result, two modified training corpora are available

as shown in Figs. 4 and 5. Figure 4 illustrates the output side's decreasing variation, while Fig. 5 shows the input side's decreasing variation.

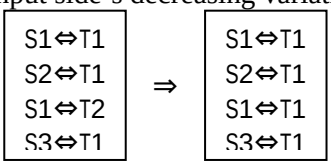

Figure 4. Modification example of replacing the output side's sentence

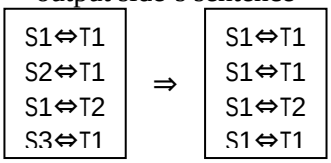

Figure 5. Modification example of replacing the input side's sentence

4 SMT System and Evaluation method

In this section, we describe the SMT systems used in these experiments. The SMT systems' decoder is a graph-based decoder (Ueffing et al., 2002; Zhang et al., 2004). The first pass of the decoder generates a word-graph, a compact representation of alternative translation candidates, using a beam search based on the scores of the lexicon and language models. In the second pass, an *A** search traverses the graph. The edges of the word-graph, or the phrase translation candidates, are generated by the list of word translations obtained from the inverted lexicon model. The phrase translations extracted from the Viterbi alignments of the training corpus also constitute the edges. Similarly, the edges are also created from dynamically extracted phrase translations from the bilingual sentences (Watanabe and Sumita, 2003). The decoder used the IBM Model 4 with a trigram language model and a five-gram part-of-speech language model. Training of the IBM model 4 was implemented by the GIZA++ package (Och and Ney, 2003). All parameters in training and decoding were the same for all experiments. Most systems with this training can be expected to achieve better accuracy when we run the parameter tuning processes. However, our purpose is to compare the difference in results caused by modifying the training corpus.

We performed experiments for JE/EJ and JC/CJ systems and four types of training corpora:

1) Original BTEC corpus;
2) Compressed BTEC corpus (see 3.2.1);
3) Replace both languages (see 3.2.2);
4) Replace one side language (see 3.2.3)
   4-1) replacement on the input side
   4-2) replacement on the output side.

For the evaluation, we use BLEU, NIST, WER, and PER as follows:

BLEU: A weighted geometric mean of the n-gram matches between test and reference sentences multiplied by a brevity penalty that penalizes short translation sentences.

NIST: An arithmetic mean of the n-gram matches between test and reference sentences multiplied by a length factor, which again penalizes short translation sentences.

mWER (Niessen et al., 2000): Multiple reference word-error rate, which computes the edit distance (minimum number of insertions, deletions, and substitutions) between test and reference sentences.

mPER: Multiple reference position-independent word-error rate, which computes the edit distance without considering the word order.

5 Experimental Results

In this section, we show the experimental results for the JE/EJ and JC/CJ systems.

5.1 EJ/JE-system-based JE group

Tables 2 and 3 show the evaluation results for the EJ/JE system.

EJ	BLEU	NIST	mWER	mPER
Original	0.36	3.73	0.55	0.51
Compress	0.47	5.83	0.47	0.44
Replace Both	0.42	5.71	0.50	0.47
Replace J.	0.44	2.98	0.60	0.58
Replace E.	**0.48**	**6.05**	**0.44**	**0.41**

Table 2. Evaluation results for EJ System

JE	BLEU	NIST	mWER	mPER
Original	0.46	3.96	0.52	0.49
Compress	0.53	8.53	**0.42**	**0.38**
Replace Both	0.49	8.10	0.46	0.41
Replace J.	**0.54**	**8.64**	**0.42**	**0.38**
Replace E.	0.51	6.10	0.52	0.49

Table 3. Evaluation results for JE system

Modification of the training data is based on the synonymous sentence group with the JE pair.

The EJ system performed at 0.55 in mWER with the original data set, and the system replacing the Japanese side achieved the best performance of 0.44 in mWER. The system then gained 0.11 in mWER. On the other hand, the system replacing the English side lost 0.05 in mWER. The mPER score also indicates a similar result. For the BLEU and NIST scores, the system replacing the Japanese side also attained the best performance.

The JE system attained a score of 0.52 in mWER with the original data set, while the system with English on the replacement side gave the best performance of 0.42 in mWER, a gain of 0.10. On the other hand, the system with Japanese on the replacement side showed no change in mWER, and the case of compression achieved good performance. The ratios of mWER and mPER are nearly the same for replacing Japanese. Thus, in both directions replacement of the input-side language derives a positive effect for translation modeling.

5.2 CJ/JC system-based JC group

Tables 4 and 5 show the evaluation results for the EJ/JE system based on the group with a JC language pair.

CJ	BLEU	NIST	mWER	mPER
Original	0.51	6.22	0.41	0.38
Compress	0.52	**6.43**	0.43	0.40
Replace both	**0.53**	5.99	***0.40***	***0.37***
Replace J.	0.50	5.98	0.41	0.39
Replace C.	0.51	6.22	0.41	0.38

Table 4. Evaluation results for CJ based on the JC language pair

JC	BLEU	NIST	mWER	mPER
Original	0.56	**8.45**	**0.38**	**0.34**
Compress	0.55	8.22	0.41	0.36
Replace both	0.56	8.32	0.39	0.35
Replace J.	0.56	8.25	0.40	0.36
Replace C.	**0.57**	8.33	**0.38**	0.35

Table 5. Evaluation results for JC based on the JC language pair

The CJ system achieved a score of 0.41 in mWER with the original data set, with the other cases similar to the original; we could not find a large difference among the training corpus modifications. Furthermore, the JC system performed at 0.38 in mWER with the original data, although the other cases' results were not as good. These results seem unusual considering the EJ/JE system, indicating that they derive from the features of the Chinese part of the BTEC corpus.

6 Discussion

Our EJ/JE experiment indicated that the system with input-side language replacement achieved better performance than that with output-side language replacement. This is a reasonable result because the system learns the translation model with fewer variations for input-side language.

In the experiment on the CJ/JC system based on the JC group, we did not provide an outline of the EJ/JE system due to the features of BTEC. Initially, BTEC data were created from pairs of Japanese and English sentences in the travel domain. Japanese-English translation pairs have variation as shown in Fig. 1. However, when Chinese data was translated, BTEC was controlled so that the same Japanese sentence has only one Chinese sentence. Accordingly, there is no variation in Chinese sentences for the pair with the same Japanese sentence. Therefore, the original training data would be similar to the situation of replacing Chinese. Moreover, replacing the Japanese data was almost to the same as replacing both sets of data. Considering this feature of the training corpus, i.e. the results for the CJ/JC system based on the group with JC language pairs, there are few differences between keeping the original data and replacing the Chinese data, or between replacing both side's data and replacing only the Japanese data. These results demonstrate the correctness of the hypothesis that reducing the input side's language variation makes learning models more effective.

Currently, our modifications only roughly process sentence pairs, though the process of making groups is very simple. Sometimes a group may include sentences or words that have slightly different meanings, such as. *fukuro* (bag), *kamibukuro* (paper bag), *shoppingu baggu* (shopping bag), *tesagebukuro* (tote bag), and *biniiru bukuro* (plastic bag). In this case if we select *tesagebukuro* from the Japanese side and "paper bag" from the English side, we have an incorrect word pair in the translation model. To handle such a problem, we would have to arrange a method to select the sen-

tences from a group. This problem is discussed in Imamura et al. (2003). As one solution to this problem, we borrowed the measures of literalness, context freedom, and word translation stability in the sentence-selection process.

In some cases, the group includes sentences with different meanings, and this problem was mentioned in Kashioka (2004). In an attempt to solve the problem, he performed a secondary decomposition step to produce a synonymous group. However, in the current training corpus, each synonymous group before the decomposition step is small, so there would not be enough difference for modifications after the decomposition step.

The replacement of a sentence could be called paraphrasing. Shimohata et al. (2004) reported a paraphrasing effect in MT systems, where if each group would have the same meaning, the variation in the phrases that appeared in the other groups would reduce the probability. Therefore, considering our results in light of their discussion, if the training corpus could be modified with the module for paraphrasing in order to control phrases, we could achieve better performance.

7 Conclusion

This paper described the modification of a training set based on a synonymous sentence group for a statistical machine translation system in order to attain better performance. In an EJ/JE system, we confirmed a positive effect by replacing the input-side language. Because the Chinese data was specific in our modification, we observed an inconclusive result for the modification in the CJ/JC system based on the synonymous sentence group with a JC language pair. However, there was still some effect on the characteristics of the training corpus. In this paper, the modifications of the training set are based on the synonymous sentence group, and we replace the sentence with rough processing. If we paraphrased the training set and controlled the phrase pair, we could achieve better performance with the same training set.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

References

Yasuhiro AKIBA, Marcello FEDERICO, Noriko KANDO, Hiromi NAKAIWA, Michael PAUL, and Jun'ichi TSUJII, 2004. *Overview of the IWSLT04 Evaluation Campaign,* In *Proc. of IWSLT04*, 1 – 12.

George Doddington. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. In *Proceedings of the HLT Conference*, San Diego, California.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto, 2003. *Automatic Construction of Machine Translation Knowledge Using Translation Literalness*, in *Proc. of EACL* 2003, 155 – 162.

Hideki Kashioka, 2004. *Grouping Synonymous Sentences from a Parallel Corpus.* In *Proc. of LREC 2004*, 391 - 394.

Sonja Niessen, Franz J. Och, Gregor Leusch, and Hermann Ney. 2000. *An evaluation tool for machine translation: Fast evaluation for machine translation research.* In *Proc.of LREC 2000*, 39 – 45.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation.* In *Proc. of ACL 2002*, 311–318.

Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto, 2004. *Building a Paraphrase Corpus for Speech Translation.* In *Proc. of LREC 2004*, 1407 - 1410.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. *Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world*, In *Proc. of LREC 2002*, 147–152.

Nicola Ueffing, Franz Josef Och, and Hermann Ney. 2002. *Generation of word graphs in statistical machine translation.* In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP02)*, 156 – 163.

Taro Watanabe and Eiichiro Sumita. 2003. *Example-based decoding for statistical machine translation.* In *Machine Translation Summit IX*, 410 – 417.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Frank Soong, Taro Watanabe and Wai Kit Lo, 2004. A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech recognition and Machine Translation, In Proc. of COLING 2004, 1168 - 1174.