Improving Phrase-based Korean-English Statistical Machine Translation

Jonghoon Lee, Donghyeon Lee, Gary Geunbae Lee.

Department of Computer Science & Engineering
Pohang University of Science and Technology, Pohang, South Korea
{jh21983, semko, gblee}@postech.ac.kr

10.21437/Interspeech.2006-263

ABSTRACT

In this paper, we describe several techniques to improve Korean-English statistical machine translation. We have built a phrase-based statistical machine translation system in a travel domain. On the baseline phrase-based system, several techniques are applied to improve the translation quality. Each technique can be applied or removed easily since the techniques are part of the preprocessing method or corpus processing method. Our experiments show that most of the techniques were successful except reordering the word sequence. The combination of the successful techniques has significantly improved the translation quality.
Index Terms: statistical machine translation

1. INTRODUCTION

Recently, most of the researchers have been using the cascading approach to achieve a speech-to-speech translation (SST) task. In the cascading approach, SST system is usually composed of three major components: a speech recognition part, a machine translation part and a speech synthesis part. Although all the components are important, we are especially interested in the machine translation component among the others, because recent statistical decoding techniques make the machine translation well integrated into speech recognition systems.

Machine translation itself has been studied by various researchers for a long time, and various approaches have been developed. Currently, a significant portion of the outperforming systems are based on some form of statistical method. Among the others, Pharaoh [1] is a state-of-the-art system based on the phrase-based approach. We chose the Pharaoh decoder as our guide since Koehn et al. [2] have reported that the Pharaoh decoder records Korean-English translation task with the best result. Consequently, we used the phrase-based statistical machine translation (SMT) method to achieve high quality translation.

Despite of many researchers working on SST for various language pairs, there is almost no research focused on Korean-English SST task. To start the research on Korean-English SST, we performed some experiments on Korean-English machine translation with conversational style parallel text. And also, based on our experiments, we proposed several methods to improve the translation quality. Although our experiments were not directly performed on SST task due to the lack of high performance Korean continuous speech recognizer (CSR), our work targets to the Korean-English SST task.

The remaining part of this paper is organized as follows. In the next section our baseline phrase-based decoding model is described. In section 3, we introduce several techniques to improve the translation quality. Finally, the experimental results are described in section 4, and the conclusions will be drawn in section 5.

2. BASELINE SYSTEM

In order to build the baseline system, we need (1) parallel texts, (2) programs to generate translation table and language model and (3) a decoder for machine translation. We did not use any additional language processing tools that are not mentioned here for the baseline environment.

2.1 Parallel Text

The corpus[1] used for our experiment is a Korean-English parallel text which is sentence-by-sentence aligned. The corpus is manually collected from several kinds of travel guide books. Consequently, the corpus is a conversational style text, and is spaced with a natural spacing unit because no additional processing is applied. That is, for Korean side texts, the texts are spaced eojeol by eojeol[2], while English side texts are spaced word by word. We selected the last sentence from every ten sentences to form a test corpus, and formed the training corpus with the reaming sentences. The statistics of our training and test corpus are shown in table 1. The average number of eojeols/words in a sentence is about 5 for Korean side texts and about 7 for English side texts, so the corpus contains relatively short sentences compared with newswire text.

Table 1. Statistics of the baseline corpus

		Korean	English
training	sentence	41,566	
	eojeol/word	190,418	279,918
	vocabulary	28,391	8,914
test	sentence	4,619	
	eojeol/word	21,111	31,042
	vocabulary	6,924	3,145

[1] We thank to Infinity Telecom, Inc. for this valuable corpus.
[2] An eojeol is a Korean spacing unit which may consist of more than one morphemes

2.2 Translation Table and Language Model

We used the Pharaoh training module and GIZA++ [3] to make a phrase translation table. In the experiments described here, we did not modify or constrain the table with additional condition. And all the options of the Pharaoh training module were set to default value. For language modeling, SRILM toolkit [4] is used to build a trigram language model.

2.3 Phrase-based Decoder

The decoder used for our experiment is a phrase-based SMT decoder. We implemented the decoder based on the Pharaoh decoder. Although the two decoders have some slight differences on details of decoding process and the performance, we did not describe the details here because the implementation of the decoder is not the focus of this paper.

3. TECHNIQUES TO IMPROVE SMT

In order to improve SMT system, several techniques were applied to our baseline system. Although the purpose of applying these techniques is focused on the improvement of the Korean-English text translation task, we expect that similar techniques can be applied to SST tasks or tasks on other similar language pairs such as Japanese-English translation.

3.1 Adding Part-Of-Speech Information

Usually, spacing unit of written text is different from the unit of meaning, that is, morpheme in Korean. Because translation tasks are related to meanings, it is obvious that morpheme unit is better than the eojeol unit in the translation task. Besides, our speech recognizer output is segmented morpheme by morpheme. It is clear that we should change the spacing unit into morphemes to get a better translation result and easy connection with speech recognizer.

Basic spacing unit of Korean text, eojeol, is composed of several morphemes with some complex inflections. Because of the ambiguity introduced by the complex eojeol, Korean morpheme analysis is usually accomplished by full Part-Of-Speech (POS) tagging. Thus changing spacing unit of Korean side text into morpheme gives us POS tag information as an additional gain.

Baseline text	이 시계 가격은 얼마입니까
POS tagged text	이/MM 시계/NNG 가격/NNG 은/JX 얼마/NNG 이/VCP ㅂ니까/EF

Figure 1. An example of the POS tagged corpus

Our intuition tells us that adding POS tag information would help word alignment, because some of Korean homographs can be distinguished by POS tag information. Some previous researchers tried to use POS tag information in various ways. For example, [5] directly updated translation table using POS tag sequences. We simply used the POS tag information by leaving the information to the Korean side texts of corpus. As a result, the original morpheme and its POS tag are regarded as a single word. An example of POS tagged corpus is shown in figure 1. For the English side of the texts, words including apostrophes are separated. We performed word alignment and phrase extraction using this POS tagged corpus, and both training and test corpus were tagged.

POS tagging changed our corpus statistics. The average number of words in a sentence and the total number of words have increased but the vocabulary size has decreased because of the changed spacing unit. New corpus statistics are shown in table 2. The number of sentences is the same as table 1. The average number of words in a Korean sentence is increased to about 9, which has almost doubled compared to the baseline. However, the English side has not changed significantly.

Table 2. Statistics of the POS tagged corpus

		Korean	English
training	morpheme/word	360,102	296,908
	vocabulary	8,473	7,609
test	morpheme/word	39,955	32,936
	vocabulary	3,383	2,854

3.2 Re-ordering Word Sequence

M. Collins et al. [6] reported that clause restructuring method can help the translation task between two languages that have different word order such as German-English. We expected that the method could help our task because Korean is also a relatively free order language, like German.

First, we parsed our training corpus with our Korean parser. By analyzing the parse trees, we manually generated a number of rules. Using these rules, we restructured the parse trees. Consequently, we got English-ordered Korean texts for the Korean side of corpus. Then, we performed word alignment and phrase extraction with the reordered corpus. When we evaluated the performance we also restructured the test corpus using the rules as well.

3.3 Deleting Useless Words

Analyzing the word alignment results that GIZA++ had generated, we noticed an interesting fact. Some Korean words have a tendency to not align definitely with any English words. We found that even for a human annotator, aligning these kinds of words to a specific English word is a very difficult and confusing task. Actually, no English words have the same meaning or function to such a word. It means that there is no method to make a correct alignment with sentences including these particular words. To resolve this problem, we automatically deleted these untranslatable words from the Korean side texts of training and test corpus.

Furthermore, most of the useless words have been found in some specific POS tags: case particles, final endings and auxiliary particles. We can easily delete such useless words by using POS tag information and some additional rules. But it is a difficult task if we do not have POS tag information. Thus, this technique is applied on the system described in the section 3.1 rather than the baseline. Figure 2 shows an example of useless words. In the example, the arrow and the small

rectangles represent the ideal alignment. In the alignment, 2 Korean words are not matched to any English word. First of the unmatched word is an objective case mark, and the second one is a final ending.

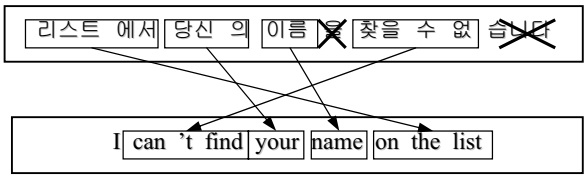

Figure 2. An example of useless words

Deleting useless words decreased the total number of words and the vocabulary size on Korean side texts of the corpus compared to the corpus shown in Table 2. The average number of words in a sentence is decreased to about 7 and the new corpus statistics is shown in Table 3.

Table 3. Statistics of the Useless words deleted corpus

		Korean	English
training	morpheme/word	290,991	296,908
	vocabulary	8,302	7,609
test	morpheme/word	32,346	32,936
	vocabulary	3,287	2,854

3.4 Language Modeling by Parts

Generally, a large vocabulary size complicates the decoding problem. This means that if we have a method to effectively reduce the vocabulary size, we can lessen the problem. To address this point, we set up an assumption:

- Category preserving assumption: translation does not change the category of a given sentence. Thus, if a source sentence is an interrogative sentence then its translation is also interrogative.

With this assumption, we can expect that dividing the language model can be helpful. Hence, we divided the language model into two models: one for interrogative sentences and another for the other kind of sentences.[1] For the Korean side text, interrogative sentences can be distinguished easily among the others because they have special endings. In this way, about 40% of sentences are marked automatically as interrogative for both training and test corpus. From the result, we obtained two different language models. And the experiment was performed using these two models along with a phrase translation table. In the experiment, the system could decide directly whether an input sentence is interrogative or not just by observing the end of each sentence. If the sentence is interrogative, decoding is performed using the interrogative language model. If not, it works in the same manner with the other language model.

[1] This is possible because we have a large portion of interrogative sentences in our travel domain conversational corpus

3.5 Appending Dictionary

A dictionary is a fundamental tool for learning a foreign language. From the dictionary, we are able to learn which word in a foreign language has the same meaning to the source language. We expect this notion can be applied to the machines as well.

Our dictionary for the machine translation task has the form of a parallel text. The difference between the parallel text and the dictionary is the contents of the entries. Parallel corpus has a pair of full sentences as its entry whereas the dictionary has a pair of words or phrases. The dictionary is composed of about 160k entries from the general domain. Although our task is a domain-limited task, the general dictionary can be used, and even the domain can be changed, because the dictionary is not bounded in a specific domain.

Our approach to use the dictionary is also very simple. We just appended the dictionary into the training corpus before word alignment and phrase extraction processes. In other words, we regarded the dictionary as an additional corpus.

Using the dictionary we can expect the two effects; the first is a boosting for correct word alignment. Since the dictionary offers exactly aligned pairs, it contributes one more count to the correct alignment. The second effect is a larger coverage of vocabulary. In the text translation task, the dictionary lowers the chance of the unknown word occurrences.

We did not apply the dictionary when we built the language model, because it can cause biases and increase the vocabulary size greatly.

4. EXPERIMENTAL RESULTS

We introduced five improvement techniques in the previous section. Now, we will describe our experiment for verifying those ideas. At first, we set up an experiment with the baseline environment. On the baseline environment, the five techniques are added one by one. However, because reordering word sequence and deleting useless words need POS tag information they were applied to the POS tagged corpus. And some possible combinations of those techniques are also tested.

The whole results are shown in table 4. In this paper, all the results are measured in terms of BLEU score [7]. The number in the parenthesis represents an improvement of the performance compared with the shaded experiment. The improvements shown in the parenthesis that are underlined are statistically significant. To make out the statistical significance, we empirically performed the statistical significance test, using the method described in [8] with p-value 0.05. Roughly, the critical point that makes the difference was around 1.5% for each pair of the experiments.

We tested the five techniques to improve the SMT system. The first one produced the best increase of the performance: adding POS tag information which was described in section 3.1. It improved the baseline to 5th result by 4.57% in terms of BLEU score. However, it includes the effect of the changes of the spacing unit into morpheme. To observe the pure effect of the POS tagging (disambiguation), we should compare 4th and 5th results. The difference of the two results is only 0.05%. It

is not a significant improvement but we still need this technique to apply other improvement techniques.

Except for the POS tag information, the best performed technique is the third one: "deleting useless words" which was described in the section 3.3. It improved 5th result to 9th one, that is, 31.54% to 33.94% in terms of BLEU score.

Language modeling by parts and appending dictionary are also well performed methods. Three pairs of experiments demonstrate the effect of Language modeling by parts: baseline and 2nd result, 5th and 7th results, and 9th and 10th results. Although the three experiments do not show statistically significant improvements, we conclude that the method is still meaningful because the results show consistent positive values.

In order to see the improvement resulted by adding the dictionary, three pairs of experiments are compared: baseline and 3rd results, 5th and 6th results, and 9th 11th results. First two of the three results are statistically significant, but third one is not.

The reordering word sequence is not performed well. It harms the translation quality instead of helping the task. The difference between 5th and 8th results shows that this technique did not work at all. We guessed that this discouraging result might be caused by our hasty application of the parser. Actually, we do not have a high performance Korean parser which is well adapted to the conversational style text. We think that the parsing errors have been propagated to the SMT system. This result tells us that the syntactic features can be harmful if the syntactic analysis tool does not guarantee enough level of performance.

In addition to the tests, we set up an experiment with the combination of all the techniques except reordering word sequence. Translation quality shows an immense improvement when the four well-performed techniques are combined. The difference between 1st and 12th results shows that the combination improved the system performance from 26.97% to 35.57% in terms of BLEU score.

5. CONCLUSIONS

In order to approach SST task, we worked on the conversational style text translation. We made up a baseline environment with a phrase-based SMT system. To improve the system, five techniques and some of their combinations are compared through twelve different experiments. Three of five techniques improved SMT system by statistically significant difference: changing spacing unit into morphemes, adding the dictionary and deleting useless words. The techniques of purely adding POS tag information and language modeling by parts have also improved the system, although the improvement is not significant but, nevertheless, still helpful. However, the technique of reordering word sequence did not improve the system due to the parsing errors. The combinations of well-performing methods give much better results.

We demonstrated that the phrase-based SMT can improve the performance significantly by applying our techniques described in section 3 and their combinations. We believe that our methods described in this paper can be applied to other translation tasks, such as SST tasks or tasks on the similar language pairs such as Japanese-English. Through this research, we confirmed that the phrase-based SMT is one of the promising approaches on Korean-English translation task. We are working on applying these techniques to Korean-English SST tasks.

Table 4. Experimental results

No.	Experimental setup	%BLEU
1	Baseline	26.97(+0.00)
2	+Language Modeling by Parts	27.33(+0.36)
3	+Adding Dictionary	28.60(+1.63)
4	+Changing spacing Unit (No tag)	31.49(+4.52)
5	Baseline + POS tag	31.54(+0.00)
6	+Adding Dictionary	33.44(+1.90)
7	+Language Modeling by Part	32.24(+0.70)
8	+Reordering word sequence	29.13(-2.41)
9	Baseline + POS tag + Deleting	33.94(+0.00)
10	+Language Modeling by Parts	34.44(+0.50)
11	+Adding Dictionary	35.19(+1.25)
12	All, except Reordering	35.57

6. ACKNOWLEDGEMENTS

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment)" (IITA-2005-C1090-0501-0018)

7. REFERENCES

[1] P. Koehn, "Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models," *in Proc. of AMTA, Washington DC, 2004*

[2] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, D. Talbot., "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation," *in Proc. of IWSLT, Pittsburgh, 2005*

[3] F. J. Och and H. Ney, "Improved statistical alignment models," *in Proc. of 38th Annual Meeting of the ACL, page 440-447, Hongkong, China, October 2000.*

[4] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," *in Proc. of ICSLP, 2002*

[5] C. Lioma, I. Ounis, "Deploying Part-of-Speech Patterns to Enhance Statistical Phrase-Based Machine Translation Resources," *in Proc. of the ACL workshop on Building and Using Parallel Texts, page 163-166, June, 2005.*

[6] M. Collins, P. Koehn, I Kucerova, "Clause Restructuring for Statistical Machine Translation," *in Proc. of 43rd Annual Meeting of the ACL, June, 2005.*

[7] K. Papineni, S. Roukos, T. Ward, W. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, York town Heights, NY, September, 2001.*

[8] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation", *in Proc. of EMNLP, June, 2004.*