

Experiments with POS-based restructuring and alignment-based reordering for statistical machine translation

Shuo Li, Derek F. Wong and Lidia S. Chao

Department of Computer and Information Science

University of Macau, Macau S.A.R., China.

leevis1987@gmail.com, {derekfw, lidiasc}@umac.mo

Abstract

This paper presents the methods which are based on the part-of-speech (POS) and auto alignment information to improve the quality of machine translation result and the word alignment. We utilize different types of POS tag to restructure source sentences and use an alignment-based reordering method to improve the alignment. After applying the reordering method, we use two phrase tables in the decoding part to keep the translation performance. Our experiments on Korean-Chinese show that our methods can improve the alignment and translation results. Since the proposed approach reduces the size of the phrase table, multi-tables are considered. The combination of all these methods together would get the best translation result.

1 Introduction

Translating between two morphological different languages is more difficult in the descriptions by Koehn (2005). In Statistical Machine Translation (SMT) system, the surface word in a morphologically poor language is difficult to be generated from a morphologically richer language. Take the example of Korean and Chinese, their morphologies are different from European languages. Korean is a kind of subject-object-verb (SOV) language while Chinese is subject-verb-object (SVO) language which is a little similar to English. This leads to a problem of word order: despite the automatic word alignment tool GIZA++ (Och and Ney, 2003) is widely applied, there are still many generated misaligned language pairs among these two languages.

In Korean, a functional word may have different morphologies under different conditions. The verb and adjective usually end with suffixes in a

sentence to represent different meanings (Li et al., 2012). On the other hand, alignment mistakes are often generated when many Korean words with different morphologies are aligned with the same Chinese tokens in Korean-Chinese translation. We applied a simple but efficient approach by utilizing different part-of-speech (POS) information to restructure Korean, after restructuring, many Korean words share the same Chinese meaning with different morphologies can be restored to their original forms. In particular, we expect to reduce the problem of misalignment due to the verb and adjective variations. Besides word restructuring, an alignment-based word reordering method which would improve the alignment result indirectly was applied in our experiment. This method is simple but effective and language-independent by modifying some alignment files. The lack of the off-the-shelf Korean-Chinese corpus is also an important problem. Most of these corpora are not open source for users, so it is hard for people applying Korean-Chinese corpus in the experiments like Euro-parl (Koehn, 2005), we built a small size corpus by ourselves in a short time to do the experiments based on the proposed methods. A script is developed for crawling parallel corpus of some specific websites.

In this paper, section 2 will review previous related works. In section 3, the POS-based restructuring method and alignment-based reordering approaches to improve the quality of alignment will be introduced. Experimental results and the analysis will be given in the following section 4. Finally, section 5 is the conclusion.

2 Related work

Several studies have been proposed to use POS tags and morphological information to enrich

languages to tackle some problems in SMT: Li et al. (2009) proposed an approach focused on using pre-processing and post-processing methods, such as reordering the source sentences in a Chinese-Korean phrase-based SMT using syntactic information. Lee et al. (2010) transformed the syntactic relations of Chinese SVO patterns and inserted the corresponding transferred relations as pseudo words to solve the problem of word order. In order to reduce the morpheme-level translation ambiguity in an English-Chinese SMT system, Wu et al. (2008) grouped the morphemes into morpheme phrase and used the domain information for translation candidate selection. A contraction separation for Spanish in a Spanish-English SMT system was proposed in (Gispert and Mariño, 2008). Habash et al. (2009) proposed methods to tackle the Arabic enclitics. The experiment in Stymne et al. (2008) described that using POS information to split the compounds in a morphologically rich language (German nouns and adjectives) gave an effect for translation output. Holmqvist et al. (2009) also reported that using POS-based and morphology-based sequence model would give an improvement to the translation quality between English and German in WMT09 shared task.

In accordance with adding richer information to the training model, reordering the source language text to make it more similar to the target side is confirmed to be another kind of method to improve the word alignment. Collins et al. (2005) employed the forms of syntactic analysis and hand-written rules on the corpus, Xia and McCord (2004) extracted the rules from a parallel text automatically. A statistical machine pre-ordering method which addressed the reordering problems as a translation from the source sentence to a monotonized source sentence was proposed by Costa-jussà and Fonollosa (2006). Visweswariah et al. (2011) proposed a method which learns a model that can directly reorder source side text from a small parallel corpus with high quality word alignment, but this is hard for people to get such a high-quality aligned parallel corpus. Ma et al. (2007) packed some words together with the help of the existing statistical word aligner, which simplify the task of automatic word alignment by packing consecutive words together.

These approaches are integrated with morphological information in the translation and decoding model. Our approach is inspired by the approach proposed by Lee et al (2006) which added

POS information; reordered the word sequence in the source corpus; deleted case particle and final ending words in Korean; appended the external dictionary in the training step between Korean and English. In the experiment reported by Li et al. (2012), these pre-processing methods on the Korean to Chinese translation system took advantage of POS in their additional factored translation model. In these studies, POS information was reported that it would improve the translation quality, but their taxonomy of POS tag is sole and less. On the word alignment side, we try to implement the idea proposed by Holmqvist et al. (2012), which was reported as a simple, language-independent reordering method to improve the quality of word alignment. But their method did not consider the problem that the probability and the amount would be changed when updated with an improved word alignment. The accuracy of alignment would be improved but the size of phrase-table would be less than the original one because there are more sure alignments generated. The probabilities of word and phrase also have the same problem.

Our works are based on the integration of these two methods. We utilized POS information and applied a richer taxonomy of POS tags in the restructuring of Korean, applied reordering method on Korean-Chinese, and combined the POS-based restructuring and alignment-based reordering together in the experiment.

3 POS-based restructuring and alignment-based reordering

The POS information is helpful when dealing with morphologically rich languages. In the morphological analysis, the Korean POS tagger involves the analytical task to identify the stem and suffixes of Korean, followed by assigning corresponding POS tags to both the morphemes and extracted stems. As described in Li et al. (2012), Korean is considered as a highly agglutinative language: the verbs, adjectives and adverbs are able to attach with affixes and particles. We considered that different category of POS tag would lead to different results of the translation. The more complex of tag would get a better result of alignment and the quality of translation. The method of processing Chinese POS is similar to Korean, which applies a more complex POS tag category from a Chinese POS tagger. Another simple but effective and language-independent reordering method which

형식상의 번거롭고 불필요한 예절을 피하다.

9 tags:

형식상의/N 번거/N+롭/X+고/E 불필요한/N 예절/N+을/J 피하/P+다/E ./S

22 tags:

형식상의/NC 번거/NC+롭/XS+고/EC 불필요한/NC 예절/NC+을/JC 피하/PV+어다/EC ./SF

9 tags deleted:

형식상의 번거 롭 고 불필요한 예절 을 피하 다 .

22 tags deleted:

형식상의 번거 롭 고 불필요한 예절 을 피하 어다 .

Figure 1. Different types of POS tag of Korean

improves the quality of automatic word alignment is applied on Korean-Chinese. The method is implemented by modifying the alignment file in Moses (Koehn et al., 2007), which needs two runs of GIZA++. After this step, an improved word alignment is generated potentially. Then, we combine the restructuring and reordering together to compare the superposed quality of these two methods.

3.1 POS-based restructuring

Because Korean is a kind of morphologically rich language, most of Korean verbs, adjectives and adverbs can be taken as the compound words like Germany. For example, the negative verb “가지 않다 (do not go)” should be restored to its original form “가다 (go)” and the negative verb suffix “지 않다 (do not)”. Another example is the future tense verb “가겠다 (will go)” is the combination of original stem “가 (go)” and suffix with future tense “겠다 (will)”. With the help of POS tagger, we can restructure the Korean with the 22 tags category instead of 9 tags in (Li et al., 2012). Here is an example of Korean restructuring in Figure 1, POS tagger can be detected the compound word and analyze its combination (tagged with “+”). The taxonomy with 22 tags is more specific than 9 tags, when tagging a noun, 22 tags will use NC (normal noun) instead of N (noun, pronoun, numeral) in 9 tags. When dealing with the compound verb “피하다 (in order to avoid)”, “피하 (avoid) +어다” is more reasonable than “피하+다”, because “어다” represents “in order to” in corresponding Chinese grammar. Then the tags were removed and restructured to a new sentence. After restructuring, the length of original sentence increased from 5 to 9 (9 tags) and 10 (22 tags). Based on previous relative simple tags, more complex taxonomy gives a deeper analysis of the sentence which would influence the alignment and the lexical possibility between the source and target language.

3.2 Alignment-based reordering

The aim of utilizing the alignment information is to make the order in source text same as the target text. It is believed that statistical word alignment methods perform better on translation with similar word orders.

The method needs two runs of word alignment, in the first run of GIZA++: the alignment information is acquired based on the original order. Then the source text is reordered by the order of the target text based on the information in the first alignment. Next, the reordered source text and the original target text are applied on the second run of GIZA++, which means this new parallel corpus includes the word with more similar order than before. After this step, a new alignment file would be generated, which covers potential improved word alignment with the reordered source language. Finally, the order of source text in the new alignment file is restored in its original order but kept with its new alignment information.

First alignment:

Bill 은 아주 조용한데 그가 말을 하도록 하려고 한다
 比尔 1 很 2 文静 3 6 7 8 设法 null 鼓励 null 他 4 说话 5
 (Bill is very quiet, try to encourage him to speak)

Second alignment (reordered):

Bill 은 아주 조용한데 하도록 하려고 한다 그가 말을
 比尔 1 很 2 文静 3 4 设法 5 6 鼓励 null 他 7 说话 8

Figure 2. The alignment

The algorithm processes the corpus and the alignment results in a single direction: source side (Holmqvist et al., 2012). As an example, in the Korean-Chinese single direction in Figure 2, each Korean word aligns with the corresponding Chinese word, but Chinese is different. There are cross alignments between “그가”, “말을”, “하도록”, “하려고”, “한다” and “文静”, “他”,

“说话”. Moreover, the alignment of these words is not totally correct.

Before the second run of GIZA++, the original Korean is reordered to the alignment of the Chinese side, so a new Korean sentence is generated by the Chinese order. After the second alignment, in Figure 2, there are no cross points and the additional correct alignment between “하려고”, “한다” and “文静” is generated, the misalignment is decreased.

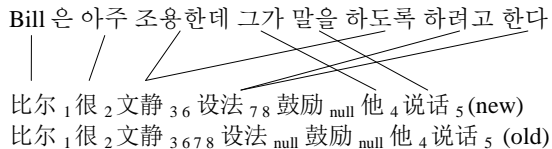


Figure 3. The improved alignment

In Figure 3, the new alignment information (new) is kept in the restored file. The crossing alignment still exists but it is more correct than the previous one (old). Based on this alignment, the establishment of word alignment, the estimation of lexical translation table and the extraction of phrase table are changed.

We assume that after applying our method, the size of the extracted words would increase because more alignments are generated at the end of the second run of GIZA++. Another assumption is that the size of the phrase table would decrease if two languages share such a different word order, because additional alignments would result in some cross alignments but the phrase extraction algorithm could not extract them. Based on two assumptions, we utilize multiple models in the decoding stage. This approach was proposed in (Koehn and Schroeder, 2007; Axelrod et al., 2011) which passes phrase and reordering tables in parallel. We used our modified tables (small size) as the main tables, and the baseline tables (big size) as the additional table when decoding. This can guarantee that if a phrase in testing sentence does not occur in the modified tables, the decoder would find the phrase in the original table. This method is effective in avoiding translation mistakes if our method harms the result.

4 Experimental results

We apply our methods on Korean-Chinese phrase-based statistical machine translation systems. The system is built based on Moses, and our reordering method is applied at the second step among the nine steps during the training in

Moses. An additional combination of the POS-based restructuring and alignment-based reordering is considered in our experiment.

4.1 Corpus and system information

The Korean-Chinese (KOR-CHN) corpus is crawled from the Internet by our script¹ and we limited the length of sentence to be under 25 words. We use 990 sentences as the testing corpus. On the other hand, we use a monolingual corpus of 600k Chinese sentences to build a Chinese 3-gram language model. ICTCLAS² is an open source Chinese segmenter applied to delimiter the word boundaries and label with proper POS tags, while the Korean text is processed by the Korean POS tagger, HanNanum³. Table 1 shows the average information of each corpus. All the experiment was trained without tuning.

| | Token | Avg. Length | Sentence |
|----------|-----------|-------------|----------|
| CHN | 664,290 | 7.36 | 90,237 |
| KOR | 539,903 | 5.98 | |
| KOR (9) | 969,445 | 10.74 | |
| KOR (22) | 1,010,117 | 11.19 | |

Table 1. Summary of training corpora

4.2 Korean-Chinese machine translation

The Korean-Chinese translation system contains a reordering model in the translation model. The reordering model is trained as the default setting from the training corpus itself. The “grow-diag-final-and” symmetrization heuristic is applied in two directions word alignment. As described in the previous section, we restructured the Korean by POS tagger and applied our reordering approach to the translation system. Since the restructured Korean can be considered as a new corpus, it could be applied to our reordering method.

According to the study of Holmqvist et al. (2012), when dealing with the morphologically different language pair, reordering the morphologically richer side performs better. In the experiments, Korean was reordered and the experimental result is shown in Table 2.

From the results, applying more POS tags on the morphological analysis of Korean got a better performance and our reordering method im-

¹ <http://nlp2ct.sftw.umac.mo/views/tools/WebpageCrawler>

² <http://ictclas.nlpir.org/>

³ <http://semanticweb.kaist.ac.kr/home/index.php/HanNanum>

proved the translation result from 14.98 to 15.50 in BLEU (Papineni et al., 2002). The combination of POS and reordering methods based on the multiple phrase tables and reordering tables got the best performance with BLEU score 17.35.

| Corpus | KOR-CHN BLEU |
|--|--------------|
| Baseline | 14.98 |
| POS-based (9 tags) | 16.61 |
| POS-based (22 tags) | 16.92 |
| Alignment-based | 15.50 |
| POS (9 tags) + Alignment | 16.71 |
| POS (22 tags) + Alignment | 17.03 |
| POS (22 tags) + Alignment + two tables | 17.35 |

Table 2. The translation results

4.3 Analysis and discussion

After the modification of the alignment file, the changes of size of the lexical file and the tables (phrase and reordering) file are shown in Table 3.

| Tables | KOR-CHN baseline | KOR-CHN modified |
|------------|------------------|------------------|
| Word | 12.39 MB | 12.52 MB |
| Phrase | 19.41 MB | 19.04 MB |
| Reordering | 10.01 MB | 9.83 MB |

Table 3. The size changes of the word and phrase tables

From the table we found that the lexical extraction is bigger than the original system, but the size of phrase tables and the reordering tables decreased slightly. The result of these changes shows that our assumption is reasonable: our method can improve the quality of automatic alignment, but the phrase extracted from the corpus would decrease. The more word alignment points were generated by using our method, the more words would be extracted. But this will bring some cross alignments when dealing with two morphological different languages.

5 Conclusion

In this paper, we presented some pre-processing methods to deal with Korean, which is a morphological rich language. POS-based restructuring restores most of the Korean verbs, adjectives and adverbs to their original format. It

is shown that the POS tag set with a richer taxonomy gives a higher translation result. Moreover, two runs of automatic alignment information got better results on the morphologically richer side. All of these methods can be combined together and improve the final translation. Finally, using two tables instead of one modified table in the decoding part will guarantee the translation quality if the reordering model harms the translation result.

Acknowledgments

This work is partially supported by the Research Committee of University of Macau, and Science and Technology Development Fund of Macau under the grants RG060/09-10S/CS/FST, and 057/2009/A2.

References

- Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK.
- Michael Collins, Philipp Koehn, and Ivoa Kucerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan.
- Marta Ruiz Costa-jussà and José A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 70–76, Sydney, Australia.
- Adrià de Gispert and José B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*. Pages 50:1034–1046.
- Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece.
- Maria Holmqvist, Sara Stymne, Lars Ahrenberg, and Magnus Merkel. 2012. Alignment-based reordering for SMT. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pages 3436–3440, Istanbul, Turkey.

- Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, pages 102–109, Cairo, Egypt.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demonstration session*, Prague, Czech Republic.
- Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.
- Shuo Li, Derek F. Wong, and Lidia S. Chao. 2012. Korean-Chinese statistical translation model. In *Machine Learning and Cybernetics (ICMLC), 2012 International Conference*, 2:767–772, Xian, Shannxi, China.
- Jin-Ji Li, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 190–196, Athens, Greece.
- Jae-Hee Lee, Seung-Wook Lee, Gumwon Hong, Young-Sook Hwang, Sang-Bum Kim, and Hae-Chang Rim. 2010. A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 623–629, Beijing, China.
- Jonghoon Lee, Donghyeon Lee, and Gary Geunbae Lee. 2006. Improving phrase-based Korean-English statistical machine translation. In *Proceedings of the Ninth International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania.
- Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of Annual Meeting-association for Computational Linguistic*, pages 304–311, Prague, Czech Republic.
- 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 *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.
- Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, Ohio, USA.
- Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. 2011. A word reordering model for improved machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 486–496, Edinburgh, Scotland, UK.
- Xianchao Wu, Naoaki Okazaki, Takashi Tsunakawa, and Jun'ichi Tsujii. 2008. Improving English-to-Chinese translation for technical terms using morphological information. In *Proceedings of the 8th AMTA Conference*, Hawaii, USA.
- Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508, Geneva, Switzerland.