Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation

Gumwon Hong, Seung-Wook Lee and Hae-Chang Rim
Department of Computer Science & Engineering
Korea University
Seoul 136-713, Korea
{gwhong,swlee,rim}@nlp.korea.ac.kr

Abstract

Often, Statistical Machine Translation (SMT) between English and Korean suffers from null alignment. Previous studies have attempted to resolve this problem by removing unnecessary function words, or by reordering source sentences. However, the removal of function words can cause a serious loss in information. In this paper, we present a possible method of bridging the morpho-syntactic gap for English-Korean SMT. In particular, the proposed method tries to transform a source sentence by inserting pseudo words, and by reordering the sentence in such a way that both sentences have a similar length and word order. The proposed method achieves 2.4 increase in BLEU score over baseline phrase-based system.

1 Introduction

Phrase-based SMT models have performed reasonably well on languages where the syntactic structures are very similar, including languages such as French and English. However, Collins et al. (2005) demonstrated that phrase-based models have limited potential when applied to languages that have a relatively different word order; such is the case between German and English. They proposed a clause restructuring method for reordering German sentences in order to resemble the order of English sentences. By modifying the source sentence structure into the target sentence structure, they argued that they could solve the decoding problem by use of completely monotonic translation.

The translation from English to Korean can be more difficult than the translation of other language pairs for the following reasons: First, Korean is *language isolate*: that is, it has little genealogical relations with other natural languages.[1] Second, the word order in Korean is relatively free because the functional morphemes, case particles and word endings, play the role as a grammatical information marker. Thus, the functional morphemes, rather than the word order, determine whether a word is a subject or an object. Third, Korean is an agglutinative language, in which a word is generally composed of at least one content morpheme and zero or more functional morphemes. Some Korean words are highly synthetic with complex inflections, and this phenomenon produces a very large vocabulary and causes data-sparseness in performing word-based alignment. To mitigate this problem, many systems tokenize Korean sentences by the morpheme unit before training and decoding the sentences.

When analyzing English-Korean translation with MOSES (Koehn et al., 2007), we found high ratio of null alignment. In figure 1, '은(*eun*)', '의(*eui*)', '하(*ha*)', 'ㄴ(*n*)', '지(*ji*)' and '는다(*neunda*)' are not linked to any word in the English sentence. In many cases, these words are function words that are attached to preceding content words. Sometimes they can be linked (incorrectly) to their head's corresponding words, or they can be linked to totally different words with respect to their meaning.

In the preliminary experiment using GIZA++ (Och and Ney, 2003) with grow-diag-final heuristic, we found that about 25% of words in Korean sentences and 21% of English sentences fail to align. This null alignment ratio is relatively high in comparison to the French-English alignment, in which about 9% of French sentences and 6% of English sentences are not aligned. Due to this null alignment, the estimation of translation probabilities for Korean function words may be incomplete; a system would perform mainly based

[1] Some may consider it an Altaic language family.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 233–236,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

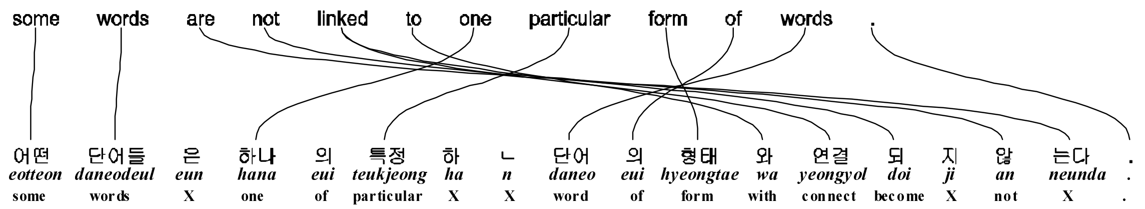
some words are not linked to one particular form of words .
어떤 단어들 은 하나 의 특정 하 ㄴ 단어 의 형태 와 연결 되 지 않 는다 .
eotteon daneodeul eun hana eui teukjeong ha n daneo eui hyeongtae wa yeongyol doi ji an neunda .
some words X one of particular X X word of form with connect become X not X .

Figure 1: An example of null alignment

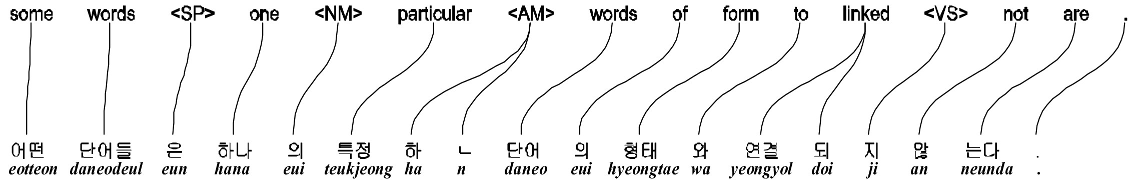
some words <SP> one <NM> particular <AM> words of form to linked <VS> not are .
어떤 단어들 은 하나 의 특정 하 ㄴ 단어 의 형태 와 연결 되 지 않 는다 .
eotteon daneodeul eun hana eui teukjeong ha n daneo eui hyeongtae wa yeongyol doi ji an neunda .

Figure 2: An example of ideal alignment

on content-words, which can deteriorate the performance of candidate generation during decoding. Also, without generating appropriate function words, the quality of the translation will undoubtedly degrade.

In this paper, we present a preprocessing method for both training and decoding in English-Korean SMT. In particular, we transform a source language sentence by inserting pseudo words and syntactically reordering it to form a target sentence structure in hopes of reducing the morpho-syntactic discrepancies between two languages. Ultimately, we expect an ideal alignment, as shown in Figure 2. Our results show that the combined pseudo word insertion and syntactic reordering method reduces null alignment ratio and makes both sentences have similar length. We report results showing that the proposed method can improve the translation quality.

2 Pseudo Word Insertion

Lee et al. (2006) find that function words in Korean sentences are not aligned to any English words, and can simply and easily be removed by referring to their POS information. The unaligned words are case particles, final endings, and auxiliary particles, and they call these words “untranslatable words”.

The method can be effective for Korean-English SMT where target language does not have corresponding function words, but it has a limitation in application to the English-Korean SMT because removing functional morphemes can cause a serious loss in information. Technically, the function words they ignored are not ‘untranslatable’ but are ‘unalignable’. Therefore, instead of removing the function words, we decide to insert some *pseudo words* into an English sentence in order to align them with potential Korean function words and make the length of both sentences similar.

To insert the pseudo words, we need to decide: (1) the kinds of words to insert, and (2) the location to insert the words. Because we expect that a pseudo word corresponds to any Korean function word which decides a syntactic role of its head, it is reasonable to utilize a dependency relation of English. Thus, given an English sentence, the candidate pseudo words are generated by the following methods: First, we parse the English sentence using Stanford dependency parser (de Marneffe et al., 2006). Then, we select appropriate typed dependency relations between pairs of words which are able to generate Korean function words. We found that 21 out of 48 dependency relations can be directly used as pseudo words. Among them, some relations provide very strong cue of case particles when inserted as pseudo words.

For example, from the following sentence, we can select as pseudo words a subjective particle <NS> and an objective particle <DO>, and insert them after the corresponding dependents *Eugene* and *guitar* respectively.

nominal_**s**ubject(play, Eugene)
direct_**o**bject(play, guitar)

Eugene <NS> can 't play the guitar <DO> well .

In a preliminary experiment on word alignment,

nominal_subject	는(*neun*), null, 이(*i*)
direct_object	을(*eul*), null, 를(*reul*)
clausal_subject	는(*neun*), null, 이(*i*)
temporal_modifier	에(*neun*), null, 오늘(*oneul*)
adj_complement	null, 아(*ah*), 하(*ha*)
agent	null, 에(*e*), 가(*ga*)
numeric_modifier	null, 의(*eui*), 개(*gae*)
adj_modifier	null, 에(*e*), 가(*ga*)
particle_modifier	null, ㄴ(*n*), 되(*doe*)

Figure 3: Selected dependency relations and their aligned function words in training data (shown the top 3 results in descending order of alignment probability)

we observe that inserting too many pseudo words can, on the contrary, increase null alignment of English sentence. Thus we filtered some pseudo words according to their respective null alignment probabilities. Figure 3 shows the top 9 selected dependency relations (actually used in the experiment) and the aligned Korean function words.

3 Syntactic Reordering

Many approaches use syntactic reordering in the preprocessing step for SMT systems (Collins et al., 2005; Xia and McCord, 2004; Zwarts and Dras, 2007). Some reordering approaches have given significant improvements in performance for translation from French to English (Xia and McCord, 2004) and from German to English (Collins et al., 2005). However, on the contrary, Lee et al. (2006) reported that the reordering of Korean for Korean-English translation degraded the performance. They presumed that the performance decrease might come from low parsing performance for conversational domain.

We believe that it is very important to consider the structural properties of Korean for reordering English sentences. Though the word order of a Korean sentence is relatively free, Korean generally observes the SOV word order, and it is a head-final language. Consequently, an object precedes a predicate, and all dependents precede their heads.

We use both a structured parse tree and dependency relations to extract following reordering rules.

- ***Verb final***: In any verb phrase, move verbal head to the end of the phrase. Infinitive verbs or verb particles are moved together.

```
He (likes ((to play) (the piano)))(1)
He (likes ((the piano) (to play)))(2)
He (((the piano) (to play)) likes)(3)
```

- ***Adjective final***: In adjective phrase, move adjective head to the end of the phrase especially if followed by PP or S/SBAR.

```
It is ((difficult) to reorder)(1)
It is (to reorder (difficult))(2)
```

- ***Antecedent final***: In noun phrase containing relative clause, move preceding NP to the end of a relative clause.

```
((rules) that are used for reordering)(1)
(that are used for reordering (rules))(2)
```

- ***Negation final***: Move negative markers to directly follow verbal head.

```
(can 't) ((play) the guitar)(1)
(can 't) (the guitar (play))(2)
(the guitar (play)) (can 't)(3)
```

4 Experiments

4.1 Experimental Setup

The baseline of our approach is a statistical phrase-based system which is trained using MOSES (Koehn et al., 2007). We collect bilingual texts from the Web and combine them with the Sejong parallel corpora [2]. About 300K pair of sentences are collected from the major bilingual news broadcasting sites. We also collect around 1M monolingual sentences from the sites to train Korean language models. The best performing language model is 5-gram order with Kneser-Ney smoothing.

For sentence level alignment, we modified the Champollion toolkit for English-Korean pair (Ma, 2006). We randomly selected 5,000 sentence pairs from Sejong corpora, of which 1,500 were used for a tuning set for minimum error rate training, and another 1,500 for development set for analysis experiment. We report testing results on the remaining 2,000 sentence pairs for the evaluation.

Korean sentences are tokenized by the morphological analyzer (Lee and Rim, 2004). For English sentence preprocessing, we use the Stanford parser with output of typed dependency relations. We then applied the pseudo word insertion and four reordering rules described in the previous section to the parse tree of each sentence.

[2] The English-Korean parallel corpora open for research purpose which contain about 60,000 sentence pairs. See http://www.sejong.or.kr/english.php for more information

	BLEU(gain)	Length Ratio
Baseline	18.03(+0.00)	0.78
+*PWI only*	18.62(+0.59)	0.91
+*Reorder only*	19.92(+1.89)	0.78
+*PWI&Reorder*	20.42(+2.39)	0.91

Table 1: BLEU score and sentence length ratio for each method

	Baseline	+*PWI*	+*Reorder*	+*P&R*
src-null	20.5	21.4	19.1	20.9
tgt-null	25.4	22.3	23.4	20.8
all-null	23.3	21.9	21.5	20.8

Table 2: Null alignment ratio (%) for each method (all-null is calculated on the whole training data)

4.2 Experimental Results

The BLEU scores are reported in Table 1. Length ratio indicates the average sentence length ratio between source sentences and target sentences. The largest gain (+2.39) is achieved when the combined pseudo word insertion (*PWI*) and word reordering is performed.

There could be reasons why the proposed approach is effective over baseline approach. Presumably, transforming to similar length and word order contributes to lower the distortion and fertility parameter values. Table 2 analyzes the effect of individual techniques in terms of the null alignment ratio. We discover that the alignment ratio can be a good way to measure the relation between the quality of word alignment and the quality of translation. As shown in Table 2, the BLEU score tends to increase as the all-null ratio decreases. Interestingly, reordering achieves the smallest null alignment ratio for source language.

5 Conclusions

In this paper, we presented a novel approach to preprocessing English-Korean SMT. The morpho-syntactic discrepancy between English and Korean causes a serious null alignment problem.

The main contributions of this paper are the following: 1) we devise a new preprocessing method for English-Korean SMT by transforming a source sentence to be much closer to a target sentence in terms of sentence length and word order. 2) we discover that the proposed method can reduce the null alignment problem, and consequently the null word alignment ratio between two languages can be a good way to measure the quality of translation.

When evaluating the proposed approach using within MOSES, the combined pseudo word insertion and syntactic reordering method outperforms the other methods. The result proves that the proposed method can be used as a useful technique for English-Korean machine translation.

Acknowledgments

This work was supported by Microsoft Research Asia. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the sponsor.

References

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL*.

Marie-Catherine de Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. of LREC*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL Demonstration session*.

Do-Gil Lee and Hae-Chang Rim. 2004. Part-of-speech tagging considering surface form for an agglutinative language. In *Proc. of ACL*.

Jonghoon Lee, Donghyeon Lee, and Gary Geunbae Lee. 2006. Improving phrase-based korean-english statistical machine translation. In *Proc. of Interspeech-ICSLP*.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proc. of LREC*.

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

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proc. of COLING*.

Simon Zwarts and Mark Dras. 2007. Syntax-based word reordering in phrase-based statistical machine translation: Why does it work? In *Proc. of MT-Summit XI*.