A Post-processing Approach to Statistical Word Alignment Reflecting Alignment Tendency between Part-of-speeches

Jae-Hee Lee[1], Seung-Wook Lee[1], Gumwon Hong[1],
Young-Sook Hwang[2], Sang-Bum Kim[2], Hae-Chang Rim[1]
[1]Dept. of Computer and Radio Communications Engineering, Korea University
[2]Institute of Future Technology, SK Telecom
[1]{jlee,swlee,gwhong,rim}@nlp.korea.ac.kr,
[2]{yshwang,sangbum.kim}@sktelecom.com

Abstract

Statistical word alignment often suffers from data sparseness. Part-of-speeches are often incorporated in NLP tasks to reduce data sparseness. In this paper, we attempt to mitigate such problem by reflecting alignment tendency between part-of-speeches to statistical word alignment. Because our approach does not rely on any language-dependent knowledge, it is very simple and purely statistic to be applied to any language pairs. End-to-end evaluation shows that the proposed method can improve not only the quality of statistical word alignment but the performance of statistical machine translation.

1 Introduction

Word alignment is defined as mapping corresponding words in parallel text. A word aligned parallel corpora are very valuable resources in NLP. They can be used in various applications such as word sense disambiguation, automatic construction of bilingual lexicon, and statistical machine translation (SMT). In particular, the initial quality of statistical word alignment dominates the quality of SMT (Och and Ney 2000; Ganchev et al., 2008); almost all current SMT systems basically refer to the information inferred from word alignment result.

One of the widely used approaches to statistical word alignment is based on the IBM models (Brown et al., 1993). IBM models are constructed based on words' co-occurrence and positional information. If sufficient training data are given, IBM models can be successfully applied to any language pairs. However, for minority language pairs such as English-Korean and Swedish-Japanese, it is very difficult to obtain large amounts of parallel corpora. Without sufficient amount of parallel corpus, it is very difficult to learn the correct correspondences between words that infrequently occur in the training data.

Part-of-speeches (POS), which represent morphological classes of words, can give valuable information about individual words and their neighbors. Identifying whether a word is a noun or a verb can let us predict which words are likely to be mapped in word alignment and which words are likely to occur in its vicinity in target sentence generation.

Many studies incorporate POS information in SMT. Some researchers perform POS tagging on their bilingual training data (Lee et al., 2006; Sanchis and Sánchez, 2008). Some of them replace individual words as new words, such as in "word/POS" form, producing new, extended vocabulary. The advantage of this approach is that POS information can help to resolve lexical ambiguity and thus improve translation quality.

On the other hand, Koehn et al. (2007) propose a factored translation model that can incorporate any linguistic factors including POS information in phrase-based SMT. The model provides a generalized representation of a translation model, because it can map multiple source and target factors.

Although all of these approaches are shown to improve SMT performance by utilizing POS information, we observe that the influence is virtually marginal in two ways:

Coling 2010: Poster Volume, pages 623–629,
Beijing, August 2010

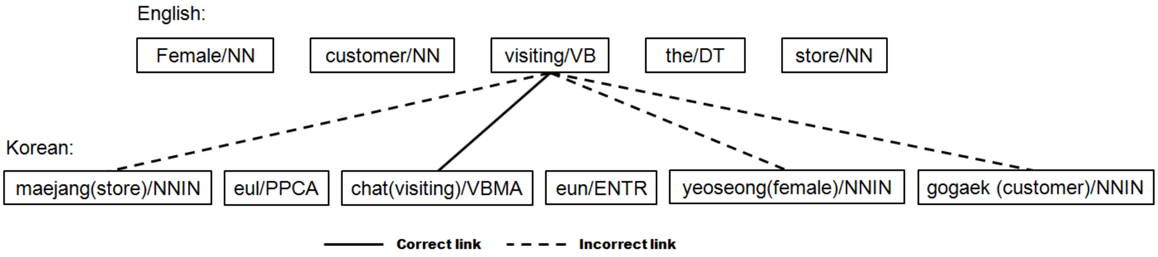

Figure 1. An example of inaccurate word alignment

1) The POS information tagged to each word may help to disambiguate in selecting word correspondences, but the increased vocabulary can also make the training data more sparse.
2) The factored translation model may help to effectively handle out-of-vocabulary (OOV) by incorporating many linguistic factors, but it still crucially relies on the initial quality of word alignment that will dominate the translation probabilities.

This paper focuses on devising a better method for incorporating POS information in word alignment. It attempts to answer the following questions:

1) Can the information regarding POS alignment tendency affect the post-processing of word alignment?
2) Can the result of word alignment affected by such information help improving the quality of SMT?

2 POS Alignment Tendency

Despite the language pairs, words with similar POSs often correspond to each other in statistical word alignment. Similarly, words with different POSs are seldom aligned. For example, Korean proper nouns very often align with English proper nouns very often but seldom align with English adverbs. We believe that this phenomenon occurs not only on English-Korean pairs but also on most of other language pairs.

Thus, in this study we hypothesize that all source language (SL) POSs have some relationship with target language (TL) POSs. Figure 1 exemplifies some results of using the IBM Models in English-Korean word alignment. As can be seen in the figure, the English word "*visiting*" is incorrectly and excessively aligned to four Korean morphemes "*maejang*", "*chat*", "*yeoseong*", and "*gogaek*". One reason for this is the sparseness of the training data; the only correct Korean morpheme "*chat*" does not sufficiently co-occur with "*visiting*" in the training data. However, it is generally believed that an English verb is more likely aligned to a Korean verb rather than a Korean noun. Likewise, we suppose that among many POSs, there are strong relationships between similar POSs and relatively weak relationships between different POSs. We hypothesize that the discovery of such relationships in advance can lead to better word alignment results.

In this paper, we propose a new method to obtain the relationship from word alignment results. The relationships among POSs, henceforth the POS alignment tendency, can be identified by the probability of the given POS pairs' alignment result where the source language POS and the target language POS co-occur in bilingual sentences. We formulate this idea using the maximum likelihood estimation as follows:

$$P\big(align = true\big|pos(f), pos(e)\big) = \frac{count(align = true|pos(f), pos(e))}{\sum_{k\in\{true,false\}} count(align = k, pos(f), pos(e))}$$

where f and e denote source word and target word respectively. *count()* is a function that returns the number of co-occurrence of f and e when they are aligned (or not aligned). Then, we adjust the formula with the existing alignment score between f and e.

$$Score(f, e) = \lambda P_{IBM}(f|e) + (1-\lambda)P\big(align = true\big|pos(f), pos(e)\big)$$

where $P_{IBM}(f\,|\,e)$ indicates the alignment probability estimated by the IBM models. λ is a weighting parameter to interpolate the reliabilities of both alignment factors. In the expe-

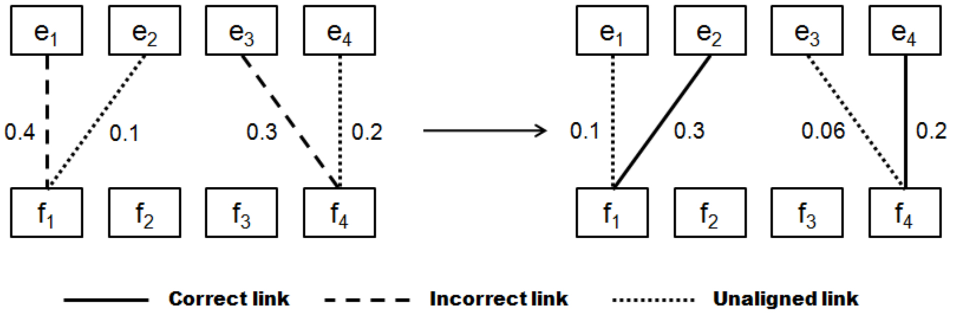

Figure 2. An example of word alignment modification

riment, λ is empirically set to improve the word alignment performance (λ=0.5).

3 Modifying Alignment

Based on the new scoring scheme as introduced in the previous section, we modify the result of the initial word alignment. The modification is performed in the following procedure:

1. For each source word *f* that has out-bound alignment link other than *null*,
2. Find the target word *e* that has the maximum alignment score according to the proposed alignment adjustment measure, and change the alignment result by mapping *f* to *e*.

This modification guarantees that the number of alignment does not change; the algorithm is designed to minimize the risk by maintaining the fertility of a word estimated by the IBM Model. Figure 2 illustrates the result before and after the alignment modification. Incorrectly links from e_1 and e_3 are deleted and missing links from e_2 and e_4 are generated during this alignment modification.

The alignment modification through the reflection of POS alignment tendency is performed on both *e*-to-*f* and *f*-to-*e* bidirectional word alignments. The bidirectional word alignment results are then symmetrized.

4 Experiments

In this paper, we attempt to reflect the POS alignment tendency in improving the word alignment performance. This section provides the experimental setup and the results that demonstrate whether the proposed approach can improve the statistical word alignment performance.

We collected bilingual texts from major bilingual news broadcasting sites. 500K sentence pairs are collected and refined manually to construct correct parallel sentences pairs. The same number of monolingual sentences is also used from the same sites to train Korean language. We also prepared a subset of the bilingual text with the size of 50K to show that the proposed model is very effective when the training set is small.

In order to evaluate the performance of word alignment, we additionally constructed a reference set with 400 sentence pairs. The evaluation is performed using precision, recall, and F-score. We use the GIZA++ toolkit for word alignment as well as four heuristic symmetrizations: intersection, union, grow-diag-final, and grow-diag (Och, 2000).

4.1 Word Alignment

We now evaluate the effectiveness of the proposed word alignment method. Table 1 and 2 report the experimental results by adding POS information to the parallel corpus. “Lexical” denotes the result of conventional word alignment produced by GIZA++. No pre-processing or post-processing is applied in this result. “Lemma/POS” is the result of word alignment with the pre-processing introduced Lee et al. (2006). Compared to the result, lemmatized lexical and POS tags are proven to be useful information for word alignment. “Lemma/POS” consistently outperforms “Lexical” despite the symmetrization heuristics in terms of precision, recall and F-score. We expect this improvement is benefited from the alleviated data sparseness by using lemmatized lexical and POS tags rather than using the lexical itself.

	Alignment heuristic	Precision	Recall	F-score
Lexical	Intersection	94.0%	50.8%	**66.0%**
	Union	53.2%	81.2%	64.3%
	Grow-diag-final	54.6%	80.9%	65.2%
	Grow-diag	60.9%	67.2%	63.9%
Lemma/POS	Intersection	95.8%	55.3%	**70.1%**
	Union	58.1%	83.3%	68.4%
	Grow-diag-final	59.7%	83.0%	69.5%
	Grow-diag	67.0%	71.6%	69.2%
Lemma/POS + POS alignment tendency	Intersection	96.1%	63.5%	76.5%
	Union	67.4%	85.1%	75.2%
	Grow-diag-final	69.8%	84.9%	76.6%
	Grow-diag	80.0%	77.0%	**78.5%**

Table 1. The performance of word alignment using small training set (50k pairs)

Experimental Setup	Alignment heuristic	Precision	Recall	F-score
Lexical	Intersection	96.8%	64.9%	**77.7%**
	Union	66.6%	87.4%	75.6%
	Grow-diag-final	67.8%	87.1%	76.2%
	Grow-diag	74.4%	79.2%	76.7%
Lemma/POS	Intersection	97.3%	66.2%	78.8%
	Union	70.7%	89.0%	78.8%
	Grow-diag-final	72.1%	88.8%	79.6%
	Grow-diag	78.8%	80.5%	**79.7%**
Lemma/POS + POS alignment tendency	Intersection	97.2%	69.3%	80.9%
	Union	73.9%	86.7%	79.8%
	Grow-diag-final	75.6%	86.4%	80.7%
	Grow-diag	85.2%	81.5%	**83.4%**

Table 2. The performance of word alignment using a large training set (500k pairs)

Experimental Setup	Symmetrization Heuristic	BLEU(50k)	BLEU (500k)
Lexical	Intersection	20.1%	29.2%
	Union	18.6%	27.2%
	Grow-diag-final	19.9%	27.7%
	Grow-diag	**20.2**%	**29.4**%
Lemma/POS	Intersection	20.3%	26.4%
	Union	18.5%	27.8%
	Grow-diag-final	20.1%	29.2%
	Grow-diag	**20.4**%	**30.8**%
Factored Model (Lemma, POS)	Intersection	20.5%	30.0%
	Union	18.1%	27.5%
	Grow-diag-final	20.3%	28.2%
	Grow-diag	**20.9**%	**31.1**%
Lemma/POS + POS alignment tendency	Intersection	**21.8**%	**29.3**%
	Union	19.5%	27.2%
	Grow-diag-final	21.3%	28.4%
	Grow-diag	20.8%	29.1%

Table 3. The performance of translation

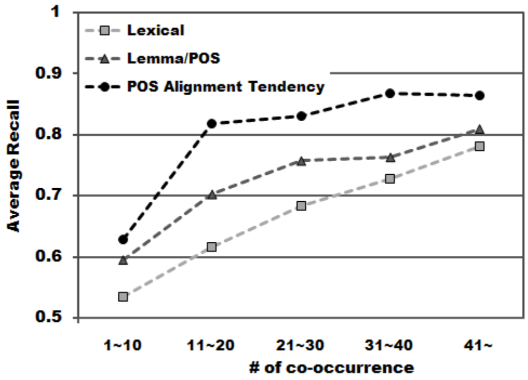

Figure 3. Average recall of word alignment pairs according to the number of their co-occurrence

Since lemmatized lexical and POS tags are shown to be useful, our post-processing method is applied to "Lemma/POS".

The experimental results show that the proposed method consistently improves word alignment in terms of F-score. It is interesting that the proposed method improves the recall of the intersection result and the precision of the union result. Thus, the proposed method achieves the best alignment performance.

As can be seen in Table 1 and 2, our method consistently improves the performance of word alignment despite the size of training data. In a small data set, the improvement of our method is much higher than that in a large set. This implies that our method is more helpful when the training data set is insufficient.

We investigate whether the proposed method actually alleviates the data sparseness problem by analyzing the aligned word pairs of low co-occurrence frequency. There are multiple word pairs that share the same number of co-occurrence in the corpus. For example, let us assume that "report-*bogoha*", "newspaper-*sinmun*" and "China-*jungguk*" pairs are co-occurred 1,000 times. We can calculate the mean of their individual recalls. We refer to this new measurement as average recall. The average recalls of these pairs are relatively higher than those of pairs with low co-occurrence frequency such as "food-*jinji*" and "center-*chojeom*" pairs. These pairs are difficult to be linked, because the word alignment model suffers from data sparseness when estimating their translation probability.

Figure 3 shows the average recall according to the number of co-occurrence. We can observe that the word alignment model tends to link word pairs more correctly if they are more frequently co-occurred. Both "Lemma/POS" and our method consistently show higher average recall throughout all frequencies, and the proposed method shows the best performance. It is also notable that the both "Lemma/POS" and our method achieve much more improvement for low co-occurrence frequencies (e.g., 11~40). This implies that the proposed method incorporates POS information more effectively than the previous method, since the proposed method achieves much higher average recall.

4.2 Statistical Machine Translation

Next, we examine the effect of the improvement of the word alignment on the translation quality. For this, we built some SMT systems with the word alignment results. We use the Moses toolkit for translation (Koehn et al., 2007). Moses is an implementation of phrase-based statistical machine translation model that has shown a state-of-the-art performance in various evaluation sets. We also perform the evaluation of the Factored model (Koehn et al., 2007) using Moses.

To investigate how the improved word alignment affect the quality of machine translation, we calculate the BLEU score for translation results with different word alignment settings as shown in Table 3. First of all, we can easily conclude that the quality of the translation is strongly dominated by the size of the training data. We can also find that the quality of the translation is correlated to the performance of the word alignment.

For a small test set, the proposed method achieved the best performance in terms of BLEU (21.8%). For a larger test set, however, the proposed method could not improve the performance of the translation with better word alignment. It is not feasible to investigate the factors that affect this deterioration, since Moses is a black box module to our system. The training of the phrase-based SMT model involves the extraction of phrases, and the result of word alignment is reflected within this process. When the training data is small, the number of extracted phrases is also apparently small. However, abundant phrases are extracted from a large amount of training data. In this case, we hypothesize that the most plausible

Rank	IBM Model			POS Alignment Tendency		
	translation	$P_{IBM}(f\|e)$	#co-occur	translation	$\mathrm{score}(\mathrm{f},\mathrm{e})$	#co-occur
1	bob/NNP	0.348	83	bob/NNP	0.214	83
2	rice/NN	0.192	73	rice/NN	0.136	73
3	*eat/VB*	0.107	57	meal/NN	0.078	43
4	meal/NN	0.075	43	food/NN	0.062	29
5	food/NN	0.043	29	*eat/VB*	0.061	57
6	bob/NN	0.038	10	bob/NN	0.059	10
7	*feed/VB*	0.010	7	*living/NN*	0.045	4
8	*cook/VB*	0.010	9	dinner/NN	0.044	10
9	*living/NN*	0.008	4	**bread/NN**	0.044	9
10	dinner/NN	0.008	10	**breakfast/NN**	0.043	6

Table 4. Top 10 translations for Korean word "bap" (food).

phrases are already obtained, and the effect of more accurate word alignment seems insignificant. More thorough analysis of this is remained as future work.

4.3 Acquisition of Bilingual Dictionary

One of the most applications of word alignment is the construction of bilingual dictionaries. By using word alignment, we can collect a (ranked) list of bilingual word pairs. Table 4 reports the top 10 translations (the most acceptable target words to align) for Korean word "*bap*" (food). The table contains the probabilities estimated by the IBM Models, the adjusted scores, and the number of co-occurrence, respectively. Italicized translations are in fact incorrect translations. Highlighted ones are new translation candidates that are correct. As can be seen in the table, the proposed approach shows a positive effect of raising new and better candidates for translation. For example, "bread" and "breakfast" have come up to the top 10 translations. This demonstrates that the low co-occurrences of "bap" with "bread" and "breakfast" are not suitably handled by alignments solely based on lexicals. However, the proposed approach ranks them at higher positions by reflecting the alignment tendency of POSs.

5 Conclusion

In this paper, we propose a new method for incorporating the POS alignment tendency to improve traditional word alignment model in post processing step. Experimental results show that the proposed method helps to alleviate the data sparseness problem especially when the training data is insufficient.

It is still difficult to conclude that better word alignment always leads to better translation. We plan on investigating the effectiveness of the proposed method using other translation system, such as Hiero (Chiang et al., 2005). We also plan to incorporate our method into other effective models, such as Factored translation model.

References

David Chiang et al., 2005. *The Hiero machine translation system: Extensions, evaluation, and analysis*. In Proc. of HLT-EMLP:779–786, Oct.

Franz Josef Och. 2000. *Giza++: Training of statistical translation models.* Available at http://www-i6.informatik.rwthaachen.de/ ~och/software/GIZA++.html.

Franz Josef Och & Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics 29 (1):19-51.

G. Sanchis and J.A. Sánchez. *Vocabulary extension via POS information for SMT*. In Mixing Approaches to Machine Translation, 2008.

Jonghoon Lee, Donghyeon Lee and Gary Geunbae Lee. *Improving Phrase-based Korean-English Statistical Machine Translation.* INTERSPEECH 2006.

Kuzman Ganchev, Joao V. Graca and Ben Taskar. 2008. *Better Alignments = Better Translations?* Proceedings of ACL-08: HLT: 986–993.

Peter F. Brown et al.,1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics 9(2): 263-311

Philipp Koehn and Hieu Hoang. *Factored Translation Models.* EMNLP 2007.

Phillipp Koehn et al., 2007. *Moses: Open source toolkit for statistical machine translation.*In Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session.