Transformation from Discontinuous to Continuous Word Alignment Improves Translation Quality

Zhongjun He 1  Hua Wu 1  Haifeng Wang 1  Ting Liu 2
1 Baidu Inc., No. 10, Shangdi 10th Street, Beijing, 100085, China
2 Harbin Institute of Technology, Harbin, China
{hezhongjun, wu_hua, wanghaifeng}@baidu.com
tliu@ir.hit.edu.cn

Abstract

We present a novel approach to improve word alignment for statistical machine translation (SMT). Conventional word alignment methods allow discontinuous alignment, meaning that a source (or target) word links to several target (or source) words whose positions are discontinuous. However, we cannot extract phrase pairs from this kind of alignments as they break the alignment consistency constraint. In this paper, we use a weighted vote method to transform discontinuous word alignment to continuous alignment, which enables SMT systems to extract more phrase pairs. We carry out experiments on large scale Chinese-to-English and German-to-English translation tasks. Experimental results show statistically significant improvements of BLEU score in both cases over the baseline systems. Our method produces a gain of +1.68 BLEU on NIST OpenMT04 for the phrase-based system, and a gain of +1.28 BLEU on NIST OpenMT06 for the hierarchical phrase-based system.

1 Introduction

Word alignment, indicating the correspondence between the source and target words in bilingual sentences, plays an important role in statistical machine translation (SMT). Almost all of the SMT models, not only phrase-based (Koehn et al., 2003), but also syntax-based (Chiang, 2005; Liu et al., 2006; Huang et al., 2006), derive translation knowledge from large amount bilingual text annotated with word alignment. Therefore, the quality of the word alignment has big impact on the quality of translation output.

Word alignments are usually automatically obtained from a large amount of bilingual training corpus. The most widely used toolkit for word alignment in SMT community is GIZA++ (Och and Ney, 2004), which implements the well known IBM models (Brown et al., 1993) and the HMM model (Vogel and Ney, 1996). Koehn et al. (2003) proposed some heuristic methods (e.g. the “grow-diag-final” method) to refine word alignments trained by GIZA++. Another group of word alignment methods (Liu et al., 2005; Moore et al., 2006; Riesa and Marcu, 2010) define feature functions to describe word alignment. They need manually aligned bilingual texts to train the model. However, the manually annotated data is too expensive to be available for all languages. Although these models reported high accuracy, the GIZA++ and “grow-diag-final” method are dominant in practice.

However, automatic word alignments are usually very noisy. The example in Figure 1 shows a Chinese and English sentence pair, with word alignment automatically trained by GIZA++ and the “grow-diag-final” method. We find many errors (dashed links) are caused by discontinuous alignment (formal definition is described in Section 2), a source (or target) word linking to several discontinuous target (or source) words. This kind of errors will result in the loss of many useful phrase pairs that are learned based on bilingual word alignment. Actually, according to the definition of phrases in a standard phrase-based model, we cannot extract phrases from the discontinuous alignment. The reason is that this kind of alignment break the alignment consistency constraint (Koehn et al., 2003). For example, the Chi-
Figure 1: An example of word alignment between a Chinese and English sentence pair. The dashed links are incorrect alignments.

The Chinese word “shi”\(^1\) is aligned to the English words “was” and “that”. However, these two English words are discontinuous, and we cannot extract the phrase pair “(shi, was)”. In this paper, we propose a simple weighed vote method to deal with the discontinuous word alignment. Firstly, we split the discontinuous alignment into several continuous alignment groups, and consider each continuous alignment group as a bucket. Secondly, we vote for each bucket with alignment score measured by word translation probabilities. Finally, we select the bucket with the highest score as the final alignment. The strength of our method is that we refine word alignment without using any external knowledge, as the word translation probabilities can be estimated from the bilingual corpus with the original word alignment.

We notice that the discontinuous alignment is helpful for hierarchical phrase-based model, as the model allows discontinuous phrases. Thus, for the hierarchical phrase-based model, our method may lost some discontinuous phrases. To solve the problem, we keep the original discontinuous alignment in the training corpus.

We carry out experiment with the state-of-the-art phrase-based and hierarchical phrase-based (Chiang, 2005) SMT systems implemented in Moses (Koehn et al., 2007). Experiments on large scale Chinese-to-English and German-to-English translation tasks demonstrate significant improvements in both cases over the baseline systems.

2 The Weighted Vote Method

To refine the discontinuous alignment, we propose a weighted vote method to transform discontinuous alignment to continuous alignment by discarding noisy links. We split discontinuous alignment into several continuous groups, and select the best group with the highest score computed by word translation probabilities as the final alignment.

For further understanding, we first describe some definitions. Given a word-aligned sentence pair \((F^i_1, E^j_1, A)\), an alignment set \(A_{set}(i)\) is the set of target word positions that aligned to the source word \(F^i_1\):

\[
A_{set}(i) = \{j | (i, j) \in A\}
\]  

(1)

For example, in Figure 1, the alignment set for the Chinese word “shaoshu3” is \(A_{set}(3) = \{5, 7, 8, 10\}\). We define an alignment span \(A_{span}(i)\) as \([\min(A_{set}(i)), \max(A_{set}(i))]\). Thus, the alignment span for the Chinese word “shaoshu3” is \(A_{span}(3) = [5, 10]\).

The alignment for \(F^i_1\) is discontinuous if there exist some target words in \(A_{span}(i)\) linking to another source word, i.e. \(\exists (i', j') \in A, \text{ where } i' \neq i, j' \in A_{span}(i)\). Otherwise, the alignment is continuous. According to the definition, the alignment for “shaoshu3” is continuous. Because the target words “the3” and “nations9” in the alignment span link to another Chinese words “de9” and “guojia10”, respectively. For a target word \(E^j_2\), the definition is similar.

If the alignment for \(F^i_1\) is discontinuous, we can split the alignment span \(A_{span}(i) = [j_1, j_2]\) into \(m\) continuous spans \(\{(j_{p1}^k, j_{p2}^k)\}\), where \(k = 1, 2, ..., m, j_{p1}^k, j_{p2}^k \in [j_1, j_2]\). Our goal is to select the best continuous span for the word \(F^i_1\). To do this, we score each continuous span with word translation probabilities:

\[
S((j_{p1}^k, j_{p2}^k)) = \sum_{t=p}^q (Pr(E^j_{t1}|F^i_1) + Pr(F^i_1|E^j_{t2}))
\]  

(2)

where,

\[
Pr(f|e) = \frac{\text{count}(f, e)}{\sum_{f'} \text{count}(f', e)}
\]  

(3)

\(^1\)The subscript denotes the word position.
Figure 2: An example of weighted voted method for selecting the best continuous alignment from the discontinuous alignment. The heavy shading area is selected as the final alignment.

The word translation probabilities can be computed from the bilingual corpus with the initial word alignment. Finally, we select the span with the highest score as the final alignment, and discard all other alignments.

We illustrate our method in Figure 2, which shows the source word “shaoshu” and its alignment in Figure 1. We split the alignments into three continuous alignment spans and compute score for each span. Finally, the span with highest score (heavy shading area) is selected as the final alignment.

We conduct the procedure for each source and target word, the improved alignment (solid links) is shown in Figure 1.

3 Experiment

To demonstrate the effect of the proposed method, we use the state-of-the-art phrase-based system and hierarchical phrase-based system implemented in Moses (Koehn et al., 2007). The phrase-based system uses continuous phrase pair as the main translation knowledge. While the hierarchical phrase-based system uses both continuous and discontinuous phrase pairs, which has an ability to capture long distance phrase reordering.

we carried out experiments on two translation tasks: the Chinese-to-English task comes from the NIST Open MT Evaluation, and the German-to-English task comes from the Workshop on Machine Translation (WMT) shared task.

3.1 Training

The training data we used are listed in Table 1. For the Chinese-English task, the bilingual data are selected from LDC. We used NIST MT03 as the development set and tested our system on NIST MT evaluation sets from 2004 to 2008. For the German-English task, the bilingual data are from the shared translation task 2013. We used WMT08 as the development set and tested our system on WMT test sets from 2010 to 2013.

The baseline systems are trained on the training corpus with initial word alignment, which was obtained via GIZA++ and “grow-diag-final” method. Based on the initial word alignment, we computed word translation probabilities and used the proposed method to obtain a refined word alignment. Then we used the refined word alignment to train our SMT systems.

The translation results are evaluated by case-insensitive BLEU-4 (Papineni et al., 2002). The feature weights of the translation system are tuned with the standard minimum-error-rate-training (Och, 2003) to maximize the systems BLEU score on the development set.

3.2 Results

3.2.1 Phrase-based System

Table 2 shows Chinese-to-English translation quality of the phrase-based system. We observed that our refined method significantly outperformed the baseline word alignment on all test sets. The improvements are ranged from 0.97 to 1.68 BLEU%.

Table 3 shows German-to-English translation quality of the phrase-based system. The improvements are ranged from 0.51 to 0.75 BLEU%.

These results demonstrate that the proposed method improves the translation quality for...
Table 4: Chinese-to-English translation quality of the hierarchical phrase-based system.

<table>
<thead>
<tr>
<th>System</th>
<th>N04</th>
<th>N05</th>
<th>N06</th>
<th>N08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>37.33</td>
<td>34.81</td>
<td>32.20</td>
<td>25.33</td>
</tr>
<tr>
<td>Refined</td>
<td>37.91</td>
<td>35.36</td>
<td>32.75</td>
<td>25.40</td>
</tr>
<tr>
<td>Combined</td>
<td>38.13</td>
<td>35.63</td>
<td>33.48</td>
<td>25.66</td>
</tr>
</tbody>
</table>

Table 5: German-to-English translation quality of the hierarchical phrase-based system.

<table>
<thead>
<tr>
<th>System</th>
<th>W10</th>
<th>W11</th>
<th>W12</th>
<th>W13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.22</td>
<td>19.77</td>
<td>20.53</td>
<td>23.51</td>
</tr>
<tr>
<td>Refined</td>
<td>21.34</td>
<td>20.64</td>
<td>20.88</td>
<td>23.82</td>
</tr>
<tr>
<td>Combined</td>
<td>21.65</td>
<td>20.87</td>
<td>21.16</td>
<td>24.04</td>
</tr>
</tbody>
</table>

phrase-based system. The reason is that by discarding noisy word alignments from the discontinuous alignments, the phrase pairs constrained by the noisy alignments can be extracted. Thus the system utilized more phrase pairs than the baseline did.

3.2.2 Hierarchical Phrase-based System

The hierarchical phrase-based system utilizes discontinuous phrase pairs for long distance phrase reordering. Some of the discontinuous phrase pairs are extracted from the discontinuous alignments. By transforming the discontinuous alignments to continuous alignments, on the one hand, we may lost some discontinuous phrase pairs. On the other hand, we may extract additional continuous and discontinuous phrase pairs as the alignment restriction is loose.

See Figure 3 for illustration. From the initial alignment, we can extract a hierarchical phrase pair "(dang X₁ shi, when X₁)") from the discontinuous alignment of the English word “when”. However, the hierarchical phrase pair cannot be extracted from our refined alignment, because our method discards the link between the Chinese word “dang” and the English word “when”. Instead, we can extract another hierarchical phrase pair “(X₁ shi, when X₁)").

Does our method still obtain improvements on the hierarchical phrase-based system? Table 4 and Table 5 shows Chinese-to-English and German-to-English translation quality of the hierarchical phrase-based system, respectively. For Chinese-to-English translation, the refined alignment obtained improvements ranged from 0.07 to 0.58 BLEU% on the test set (the row “Refined”). While for German-to-English translation, the improvements ranged from 0.12 to 0.59 BLEU% on the test set (the row “Refined”).

We find that the improvements are less than that of the phrase-based system. As discussed above, our method may lost some hierarchical phrase pairs that extracted from the discontinuous alignments. To solve the problem, we combine the initial alignments and the refined alignments to train the SMT system. The results are shown in the row “Combined” in Table 4 and Table 5. For Chinese-to-English translation, we obtained an improvements of 1.28 BLEU% on NIST06 over the baseline. While for German-to-English translation, the greatest improvements is 1.10 BLEU% on WMT11.

4 Analyses

In order to further study the performance of the proposed method, we analyze the word alignment and the phrase table for Chinese-to-English translation. We find that our method improves the quality of word alignment. And as a result, more useful phrase pairs are extracted from the refined word alignment.

4.1 Word Alignment

The Chinese-to-English training corpus contains 4.5M sentence pairs. By applying GIZA++ and the “grow-diag-final” method, we obtained initial alignments. We find that 4.0M (accounting for 89%) sentence pairs contain discontinuous alignments. We then used the proposed method to discard noisy links. By doing this, the total links between words in the training corpus are reduced from 99.6M to 78.9M, indicating that 21% links are discarded.

We do not perform combination for phrase-based system, because the phrase table extracted from the initial alignment is a subset of that extracted from the refined alignment.
<table>
<thead>
<tr>
<th>Alignment</th>
<th>Precision</th>
<th>Recall</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>62.94</td>
<td>89.55</td>
<td>26.07</td>
</tr>
<tr>
<td>Refined</td>
<td>73.43</td>
<td>87.82</td>
<td>20.01</td>
</tr>
</tbody>
</table>

Table 6: Precision, Recall and AER on Chinese-to-English alignment.

<table>
<thead>
<tr>
<th>Alignment</th>
<th>StandPhr</th>
<th>HierPhr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>29M</td>
<td>86M</td>
</tr>
<tr>
<td>Refined</td>
<td>104M</td>
<td>436M</td>
</tr>
</tbody>
</table>

Table 7: The phrase number extracted from the initial and refined alignment for the hierarchical phrase-based system on Chinese-to-English translation. StandPhr is standard phrase, HierPhr is hierarchical phrase.

We evaluated the alignment quality on 200 sentence pairs. Results are shown in Table 6. It is observed that our method improves the precision and decreases the AER, while keeping a high recall. This means that our method effectively discards noisy links in the initial word alignments.

### 4.2 Phrase Table

According to the standard definition of phrase in SMT, phrase pairs cannot be extracted from the discontinuous alignments. By transforming discontinuous alignments into continuous alignment, we can extract more phrase pairs. Table 7 shows the number of standard phrases and hierarchical phrases extracted from the initial and refined word alignments. We find that the number of both phrases and hierarchical phrases grows heavily. This is because that the word alignment constraint for phrase extraction is loosed by removing noisy links. Although the phrase table becomes larger, fortunately, there are some methods (Johnson et al., 2007; He et al., 2009) to prune phrase table without hurting translation quality.

For further illustration, we compare the phrase pairs extracted from the initial alignment and refined alignment in Figure 1. From the initial alignments, we extracted only 3 standard phrase pairs and no hierarchical phrase pairs (Table 8). After discarding noisy alignments (dashed links) by using the proposed method, we extracted 21 standard phrase pairs and 36 hierarchical phrases. Table 9 and Table 10 show selected phrase pairs and hierarchical phrase pairs, respectively.

### 5 Conclusion and Future Work

In this paper, we proposed a novel method to improve word alignment for SMT. The method refines initial word alignments by transforming discontinuous alignment to continuous alignment. As a result, more useful phrase pairs are extracted from the refined word alignment. Our method is simple and efficient, since it uses only the word translation probabilities obtained from the initial alignments to discard noisy links. Our method is independent of languages and can be applied to most SMT models. Experimental results show significantly improvements for the state-of-the-art phrase-based and hierarchical phrase-based systems on all Chinese-to-English and German-to-English translation tasks.

In the future, we will refine the method by considering neighbor words and alignments when discarding noisy links.

**Acknowledgement**

This paper is supported by the 973 program No. 2014CB340505. We would like to thank Xuan Liu and the anonymous reviewers for their insightful comments.
References


Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*.


