Word Alignment Combination over Multiple Word Segmentation

Ning Xi, Guangchao Tang, Boyuan Li, Yinggong Zhao
State Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology,
Nanjing University, Nanjing, 210093, China
{xin,tanggc,liby,zhaoyg}@nlp.nju.edu.cn

Abstract

In this paper, we present a new word alignment combination approach on language pairs where one language has no explicit word boundaries. Instead of combining word alignments of different models (Xiang et al., 2010), we try to combine word alignments over multiple monolingually motivated word segmentation. Our approach is based on link confidence score defined over multiple segmentations, thus the combined alignment is more robust to inappropriate word segmentation. Our combination algorithm is simple, efficient, and easy to implement. In the Chinese-English experiment, our approach effectively improved word alignment quality as well as translation performance on all segmentations simultaneously, which showed that word alignment can benefit from complementary knowledge due to the diversity of multiple and monolingually motivated segmentations.

1 Introduction

Word segmentation is the first step prior to word alignment for building statistical machine translations (SMT) on language pairs without explicit word boundaries such as Chinese-English. Many works have focused on the improvement of word alignment models. (Brown et al., 1993; Haghighi et al., 2009; Liu et al., 2010). Most of the word alignment models take single word segmentation as input. However, for languages such as Chinese, it is necessary to segment sentences into appropriate words for word alignment.

A large amount of works have stressed the impact of word segmentation on word alignment. Xu et al. (2004), Ma et al. (2007), Chang et al. (2008), and Chung et al. (2009) try to learn word segmentation from bilingually motivated point of view; they use an initial alignment to learn word segmentation appropriate for SMT. However, their performance is limited by the quality of the initial alignments, and the processes are time-consuming. Some other methods try to combine multiple word segmentation at SMT decoding step (Xu et al., 2005; Dyer et al., 2008; Zhang et al., 2008; Dyer et al., 2009; Xiao et al., 2010). Different segmentations are yet independently used for word alignment.

Instead of time-consuming segmentation optimization based on alignment or postponing segmentation combination late till SMT decoding phase, we try to combine word alignments over multiple monolingually motivated word segmentation on Chinese-English pair, in order to improve word alignment quality and translation performance for all segmentations. We introduce a tabular structure called word segmentation network (WSN for short) to encode multiple segmentations of a Chinese sentence, and define skeleton links (SL for short) between spans of WSN and words of English sentence. The confidence score of a SL is defined over multiple segmentations. Our combination algorithm picks up potential SLs based on their confidence scores similar to Xiang et al. (2010), and then projects each selected SL to link in all segmentation respectively. Our algorithm is simple, efficient, easy to implement, and can effectively improve word alignment quality on all segmentations simultaneously, and alignment errors caused

Proceedings of the ACL-HLT 2011 Student Session, pages 1–5,
Portland, OR, USA 19-24 June 2011.

by inappropriate segmentations from single segmenter can be substantially reduced.

Two questions will be answered in the paper: 1) how to define the link confidence over multiple segmentations in combination algorithm? 2) According to Xiang et al. (2010), the success of their word alignment combination of different models lies in the complementary information that the candidate alignments contain. In our work, are multiple monolingually motivated segmentations complementary enough to improve the alignments?

The rest of this paper is structured as follows: WSN will be introduced in section 2. Combination algorithm will be presented in section 3. Experiments of word alignment and SMT will be reported in section 4.

2 Word Segmentation Network

We propose a new structure called word segmentation network (WSN) to encode multiple segmentations. Due to space limitation, all definitions are presented by illustration of a running example of a sentence pair:

下雨路滑 (xia-yu-lu-hua)
Road is slippery when raining

We first introduce *skeleton segmentation*. Given two segmentation S_1 and S_2 in Table 1, the word boundaries of their skeleton segmentation is the union of word boundaries (marked by "/") in S_1 and S_2.

	Segmentation
S_1	下 / 雨 / 路滑
S_2	下雨 / 路 / 滑
skeleton	下 / 雨 / 路 / 滑

Table 1: The skeleton segmentation of two segmentations S1 and S2.

The WSN of S_1 and S_2 is shown in Table 2. As is depicted, line 1 and 2 represent words in S_1 and S_2 respectively, line 3 represents skeleton words. Each column, or span, comprises a skeleton word and words of S_1 and S_2 with the skeleton word as their morphemes at that position. The number of columns of a WSN is equal to the number of skeleton words. It should be noted that there may be words covering two or more spans, such as "路滑" in S_1, because the word "路滑" in S_1 is split into two words "路" and "滑" in S_2.

S_1	下 $_1$	雨 $_2$	路滑 $_3$	
S_2	下雨 $_1$		路 $_2$	滑 $_3$
skeleton	下 $_1$	雨 $_2$	路 $_3$	滑 $_4$

Table 2: The WSN of Table 1. Subscripts indicate indexes of words.

The skeleton word can be projected onto words in the same span in S_1 and S_2. For clarity, words in each segmentation are indexed (1-based), for example, "路滑" in S_1 is indexed by 3. We use a projection function $\delta_k(j)$ to denote the index of the word onto which the *j*-th skeleton word is projected in the *k*-th segmentation, for example, $\delta_1(4) = 3$ and $\delta_2(3) = 2$.

In the next, we define the links between spans of the WSN and English words as skeleton links (SL), the subset of all SLs comprise the skeleton alignment (SA). Figure 1 shows an SA of the example.

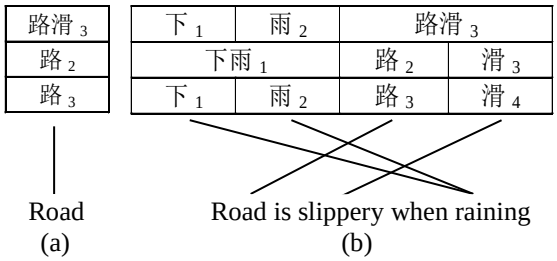

Figure 1: An example alignment between WSN in Table 2 and English sentence "Road is slippery when raining". (a) skeleton link; (b) skeleton alignment.

Each span of the WSN comprises words from different segmentations (Figure 1a), which indicates that the confidence score of a SL can be defined over words in the same span. By projection function, a SL can be projected onto the link for each segmentation. Therefore, the problem of combining word alignment over different segmentations can be transformed into the problem of selecting SLs for SA first, and then project the selected SLs onto links for each segmentation respectively.

3 Combination Algorithm

Given k alignments a_k over segmentations s_k respectively $(k = 1, \dots, K)$, and (C, E) is the pair

of the Chinese WSN and its parallel English sentence. Suppose A_{ij} is the SL between the *j*-th span c_j and *i*-th English word e_i, a_{ij}^k is the link between the *j*-th Chinese word c_j^k in s_k and e_i. Inspired by Huang (2009), we define the confidence score of each SL as follows

$$C(A_{ij}|C,E) = \sum_{k=1}^{K} w_i * c(a_{i\delta_k(j)}^k|C,E) \quad (1)$$

where $c(a_{i\delta_k(j)}^k|C,E)$ is the confidence score of the link $a_{i\delta_k(j)}^k$, defined as

$$c(a_{i\delta_k(j)}^k|C,E) = \sqrt{q_{c2e}(a_{i\delta_k(j)}^k|C,E) * q_{e2c}(a_{i\delta_k(j)}^k|C,E)} \quad (2)$$

where c-to-e link posterior probability is defined as

$$q_{c2e}(a_{i\delta_k(j)}^k|C,E) = \frac{p_k(e_i|c_{\delta_k(j)}^k)}{\sum_{i'=1}^{I} p_k(e_{i'}|c_{\delta_k(j)}^k)} \quad (3)$$

and *I* is the length of E. E-to-c link posterior probability $q_{e2c}(a_{i\delta_k(j)}^k|C,E)$ can be defined similarly,

Our alignment combination algorithm is as follows.

1. Build WSN for Chinese sentence.
2. Compute the confidence score for each SL based on Eq. (1). A SL A_{ij} gets a vote from a_k if $a_{i\delta_k(j)}^k$ appears in a_k $(k = 1, \dots, K)$. Denote the set of all SLs getting at least one vote by B_0.
3. All SLs in B_0 are sorted in descending order and evaluated sequentially. A SL A_{ij} is included if its confidence score is higher than a tunable threshold α, and one of the following is true[1]:
   - Neither c_j nor e_j is aligned so far;
   - c_j is not aligned and its left or right neighboring word is aligned to e_j so far;
   - e_j is not aligned and its left or right neighboring word is aligned to c_j so far.
4. Repeat 3 until no more SLs can be included. All included SLs comprise B_1.
5. Map SLs in B_1 on each s_k to get *k* new alignments a'_k respectively, i.e. $a'_k = \{a_{i\delta_k(j)}^k | A_{ij} \in B_1\}$[2]$(k = 1, \dots, K)$. For each k, we sort all links in a'_k in ascending order and evaluated them sequentially. Compare a'_k and a_k, A link a'^k_{ij} is removed from a'_k if it is not appeared in a_k, and one of the following is true:
   - both c_j^k and e_j are aligned in a'_k;
   - There is a word which is neither left nor right neighboring word of e_j but aligned to c_j^k in a'_k;
   - There is a word which is neither left nor right neighboring word of c_j^k but aligned to e_j in a'_k.

The heuristic in step 3 is similar to Xiang et al. (2010), which avoids adding error-prone links. We apply the similar heuristic again in step 5 in each a'_k $(k = 1, \dots, K)$ to delete error-prone links. The weights in Eq. (1) and α can be tuned in a hand-aligned dataset to maximize word alignment F-score on any a'_k with hill climbing algorithm. Probabilities in Eq. (2) and Eq. (3) can be estimated using GIZA.

4 Experiment

4.1 Data

Our training set contains about 190K Chinese-English sentence pairs from LDC2003E14 corpus. The NIST'06 test set is used as our development set and the NIST'08 test set is used as our test set. The Chinese portions of all the data are preprocessed by three monolingually motived segmenters respectively. These segmenters differ in either training method or specification, including ICTCLAS (I)[3], Stanford segmenters with CTB (C) and PKU (P) specifications[4] respectively. We used a phrase-based MT system similar to (Koehn et al., 2003), and generated two baseline alignments using GIZA++ enhanced by *gdf* heuristics (Koehn et al., 2003) and a linear discriminative word alignment model (DIWA) (Liu et al., 2010) on training set with the three segmentations respectively. A 5-gram language model trained from the Xinhua portion of Gigaword corpus was used. The decoding weights were optimized with Minimum Error Rate Training (MERT) (Och, 2003). We used the hand-aligned set of 491 sentence pairs in Haghighi et al. (2009), the first 250 sentence pairs were used to tune the weights in Eq. (1), and the other 241 were

[1] SLs getting *K* votes are forced to be included without further examination.

[2] Two or more SLs in B_1 may be projected onto one links in a'_k, in this case, we keep only one in a'_k.

[3] http://www.ictclas.org/

[4] http://nlp.stanford.edu/software/segmenter.shtml

Figure 2: Two examples (left and right respectively) of word alignment on segmentation C. Baselines (DIWA) are in the top half, combined alignments are in the bottom half. The solid line represents the correct link while the dashed line represents the bad link. Each word is enclosed in square brackets.

used to measure the word alignment quality. Note that we adapted the Chinese portion of this hand-aligned set to segmentation C.

4.2 Improvement of Word Alignment

We first evaluate our combination approach on the hand-aligned set (on segmentation C). Table 3 shows the precision, recall and F-score of baseline alignments and combined alignments.

As shown in Table 3, the combination alignments outperformed the baselines (setting C) in all settings in both GIZA and DIWA. We notice that the higher F-score is mainly due to the higher precision in GIZA but higher recall in DIWA. In GIZA, the result of C+I and C+P achieve 8.4% and 9.5% higher F-score respectively, and both of them outperformed C+P+I, we speculate it is because GIZA favors recall rather than DIWA, i.e. GIZA may contain more bad links than DIWA, which would lead to more unstable F-score if more alignments produced by GIZA are combined, just as the poor precision (69.68%) indicated. However, DIWA favors precision than recall (this observation is consistent with Liu et al. (2010)), which may explain that the more diversified segmentations lead to better results in DIWA.

	GIZA			DIWA		
setting	**P**	**R**	**F**	**P**	**R**	**F**
C	61.84	84.99	71.59	83.12	78.88	80.94
C+P	80.16	79.80	79.98	84.15	79.41	81.57
C+I	82.96	79.28	81.08	84.41	81.69	83.03
C+I+P	69.68	85.17	77.81	83.38	82.98	83.18

Table 3: Alignment precision, recall and F-score. C: baseline, C+I: Combination of C and I.

Figure 2 gives baseline alignments and combined alignments on two sentence pairs in the training data. As can be seen, alignment errors caused by inappropriate segmentations by single segmenter were substantially reduced. For example, in the second example, the word "香港特别行政区 hksar" appears in segmentation I of the Chinese sentence, which benefits the generation of the three correct links connecting for words "香港" ,"特别", "行政区" respectively in the combined alignment.

4.3 Improvement in MT performance

We then evaluate our combination approach on the SMT training data on all segmentations. For efficiency, we just used the first 50k sentence pairs of the aligned training corpus with the three segmentations to build three SMT systems respectively. Table 4 shows the BLEU scores of baselines and combined alignment (C+P+I, and then projected onto C, P, I respectively). Our approach achieves improvement over baseline alignments on all segmentations consistently, without using any lattice decoding techniques as Dyer et al. (2009). The gain of translation performance purely comes from improvements of word alignment on all segmentations by our proposed word alignment combination.

	GIZA		DIWA	
Segmentation	**B**	**Comb**	**B**	**Comb**
C	19.77	20.9	20.18	20.71
P	20.5	21.16	20.41	21.14
I	20.11	21.14	20.46	21.30

Table 4: Improvement in BLEU scores. B:Baseline alignment, Comb: Combined alignment.

5 Conclusion

We evaluated our word alignment combination over three monolingually motivated segmentations on Chinese-English pair. We showed that the combined alignment significantly outperforms the baseline alignment with both higher F-score and higher BLEU score on all segmentations. Our work also proved the effectiveness of link confidence score in combining different word alignment models (Xiang et al., 2010), and extend it to combine word alignments over different segmentations.

Xu et al. (2005) and Dyer et al. (2009) combine different segmentations for SMT. They aim to achieve better translation but not higher alignment quality of all segmentations. They combine multiple segmentations at SMT decoding step, while we combine segmentation alternatives at word alignment step. We believe that we can further improve the performance by combining these two kinds of works. We also believe that combining word alignments over both monolingually motivated and bilingually motivated segmentations (Ma et al., 2009) can achieve higher performance.

In the future, we will investigate combining word alignments on language pairs where both languages have no explicit word boundaries such as Chinese-Japanese.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61003112, and the National Fundamental Research Program of China (2010CB327903). We would like to thank Xiuyi Jia and Shujie Liu for useful discussions and the anonymous reviewers for their constructive comments.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Peitra, Robert L. Mercer. 1993. *The Mathematics of statistical machine translation: parameter estimation*. *Computational Linguistics*, 19(2):263-311.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. *Optimizing Chinese word segmentation for machine translation performance*. In *Proceedings of* third workshop on SMT, Pages:224-232.

Tagyoung Chung and Daniel Gildea. 2009. *Unsupervised tokenization for machine translation*. In *Proceedings of* EMNLP, Pages:718-726.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. *Generalizing word lattice translation*. In *Proceedings of* ACL, Pages:1012-1020.

Christopher Dyer. 2009. *Using a maximum entropy model to build segmentation lattices for mt*. In *Proceedings of* NAACL, Pages:406-414.

Franz Josef Och. 2003. *Minimum error rate training in statistical machine translation*. In *Proceedings of* ACL, Pages:440-447.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. *Better word alignments with supervised ITG models*. In *Proceedings of* ACL, Pages: 923-931.

Fei Huang. 2009. Confidence measure for word alignment. In *Proceedings of* ACL, Pages:932-940.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. *Statistical phrase-based translation.* In *Proceedings of* HLT-NAACL, Pages:48-54.

Yang Liu, Qun Liu, Shouxun Lin. 2010. *Discriminative word alignment by linear modeling*. Computational Linguistics, 36(3):303-339.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. *Bootstrapping word alignment via word packing*. In *Proceedings of* ACL, Pages:304-311.

Yanjun Ma and Andy Way. 2009. *Bilingually motivated domain-adapted word segmentation for statistical machine translation.* In *Proceedings of* EACL, Pages:549-557.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010. *Diversify and combine: improving word alignment for machine translation on low-resource languages.* In *Proceedings of* ACL, Pages:932-940.

Xinyan Xiao, Yang Liu, Young-Sook Hwang, Qun Liu, Shouxun Lin. 2010. *Joint tokenization and translation*. In *Proceedings of* COLING, Pages:1200-1208.

Jia Xu, Richard Zens, and Hermann Ney. 2004. *Do we need Chinese word segmentation for statistical machine translation?* In *Proceedings of* the ACL SIGHAN Workshop, Pages: 122-128.

Jia Xu, Evgeny Matusov, Richard Zens, and Hermann Ney. 2005. *Integrated Chinese word segmentation in statistical machine translation*. In *Proceedings of* IWSLT.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. *Improved statistical machine translation by multiple Chinese word segmentation.* In *Proceedings of* the Third Workshop on SMT, Pages:216-223.