Diversify and Combine: Improving Word Alignment for Machine Translation on Low-Resource Languages

Bing Xiang, Yonggang Deng, and Bowen Zhou
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
{bxiang,ydeng,zhou}@us.ibm.com

Abstract

We present a novel method to improve word alignment quality and eventually the translation performance by producing and combining complementary word alignments for low-resource languages. Instead of focusing on the improvement of a single set of word alignments, we generate multiple sets of diversified alignments based on different motivations, such as linguistic knowledge, morphology and heuristics. We demonstrate this approach on an English-to-Pashto translation task by combining the alignments obtained from syntactic reordering, stemming, and partial words. The combined alignment outperforms the baseline alignment, with significantly higher F-scores and better translation performance.

1 Introduction

Word alignment usually serves as the starting point and foundation for a statistical machine translation (SMT) system. It has received a significant amount of research over the years, notably in (Brown et al., 1993; Ittycheriah and Roukos, 2005; Fraser and Marcu, 2007; Hermjakob, 2009). They all focused on the improvement of word alignment models. In this work, we leverage existing aligners and generate multiple sets of word alignments based on complementary information, then combine them to get the final alignment for phrase training. The resource required for this approach is little, compared to what is needed to build a reasonable discriminative alignment model, for example. This makes the approach especially appealing for SMT on low-resource languages.

Most of the research on alignment combination in the past has focused on how to combine the alignments from two different directions, source-to-target and target-to-source. Usually people start from the intersection of two sets of alignments, and gradually add links in the union based on certain heuristics, as in (Koehn et al., 2003), to achieve a better balance compared to using either intersection (high precision) or union (high recall). In (Ayan and Dorr, 2006) a maximum entropy approach was proposed to combine multiple alignments based on a set of linguistic and alignment features. A different approach was presented in (Deng and Zhou, 2009), which again concentrated on the combination of two sets of alignments, but with a different criterion. It tries to maximize the number of phrases that can be extracted in the combined alignments. A greedy search method was utilized and it achieved higher translation performance than the baseline.

More recently, an alignment selection approach was proposed in (Huang, 2009), which computes confidence scores for each link and prunes the links from multiple sets of alignments using a hand-picked threshold. The alignments used in that work were generated from different aligners (HMM, block model, and maximum entropy model). In this work, we use soft voting with weighted confidence scores, where the weights can be tuned with a specific objective function. There is no need for a pre-determined threshold as used in (Huang, 2009). Also, we utilize various knowledge sources to enrich the alignments instead of using different aligners. Our strategy is to diversify and then combine in order to catch any complementary information captured in the word alignments for low-resource languages.

The rest of the paper is organized as follows.

Proceedings of the ACL 2010 Conference Short Papers, pages 22–26,
Uppsala, Sweden, 11-16 July 2010.

We present three different sets of alignments in Section 2 for an English-to-Pashto MT task. In Section 3, we propose the alignment combination algorithm. The experimental results are reported in Section 4. We conclude the paper in Section 5.

2 Diversified Word Alignments

We take an English-to-Pashto MT task as an example and create three sets of additional alignments on top of the baseline alignment.

2.1 Syntactic Reordering

Pashto is a subject-object-verb (SOV) language, which puts verbs after objects. People have proposed different syntactic rules to pre-reorder SOV languages, either based on a constituent parse tree (Drábek and Yarowsky, 2004; Wang et al., 2007) or dependency parse tree (Xu et al., 2009). In this work, we apply syntactic reordering for verb phrases (VP) based on the English constituent parse. The VP-based reordering rule we apply in the work is:

- $VP(VB*, *) \rightarrow VP(*, VB*)$

where $VB*$ represents VB, VBD, VBG, VBN, VBP and VBZ.

In Figure 1, we show the reference alignment between an English sentence and the corresponding Pashto translation, where E is the original English sentence, P is the Pashto sentence (in romanized text), and E' is the English sentence after reordering. As we can see, after the VP-based reordering, the alignment between the two sentences becomes monotone, which makes it easier for the aligner to get the alignment correct. During the reordering of English sentences, we store the index changes for the English words. After getting the alignment trained on the reordered English and original Pashto sentence pairs, we map the English words back to the original order, along with the learned alignment links. In this way, the alignment is ready to be combined with the baseline alignment and any other alternatives.

2.2 Stemming

Pashto is one of the morphologically rich languages. In addition to the linguistic knowledge applied in the syntactic reordering described above, we also utilize morphological analysis by applying stemming on both the English and Pashto sides. For English, we use Porter stemming (Porter, 1980), a widely applied algorithm to remove the common morphological and inflexional endings from words in English. For Pashto, we utilize a morphological decompostion algorithm that has been shown to be effective for Arabic speech recognition (Xiang et al., 2006). We start from a fixed set of affixes with 8 prefixes and 21 suffixes. The prefixes and suffixes are stripped off from the Pashto words under the two constraints:(1) Longest matched affixes first; (2) Remaining stem must be at least two characters long.

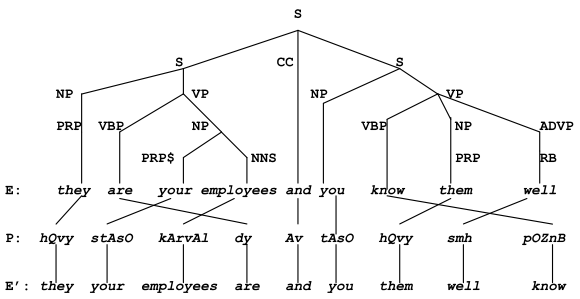

Figure 1: Alignment before/after VP-based reordering.

2.3 Partial Word

For low-resource languages, we usually suffer from the data sparsity issue. Recently, a simple method was presented in (Chiang et al., 2009), which keeps partial English and Urdu words in the training data for alignment training. This is similar to the stemming method, but is more heuristics-based, and does not rely on a set of available affixes. With the same motivation, we keep the first 4 characters of each English and Pashto word to generate one more alternative for the word alignment.

3 Confidence-Based Alignment Combination

Now we describe the algorithm to combine multiple sets of word alignments based on weighted confidence scores. Suppose a_{ijk} is an alignment link in the i-th set of alignments between the j-th source word and k-th target word in sentence pair (S,T). Similar to (Huang, 2009), we define the confidence of a_{ijk} as

$$c(a_{ijk}|S,T) = \sqrt{q_{s2t}(a_{ijk}|S,T)q_{t2s}(a_{ijk}|T,S)}, \quad (1)$$

where the source-to-target link posterior probability

$$q_{s2t}(a_{ijk}|S,T) = \frac{p_i(t_k|s_j)}{\sum_{k'=1}^{K} p_i(t_{k'}|s_j)}, \quad (2)$$

and the target-to-source link posterior probability $q_{t2s}(a_{ijk}|T,S)$ is defined similarly. $p_i(t_k|s_j)$ is the lexical translation probability between source word s_j and target word t_k in the i-th set of alignments.

Our alignment combination algorithm is as follows.

1. Each candidate link a_{jk} gets soft votes from N sets of alignments via weighted confidence scores:

$$v(a_{jk}|S,T) = \sum_{i=1}^{N} w_i * c(a_{ijk}|S,T), \quad (3)$$

   where the weight w_i for each set of alignment can be optimized under various criteria. In this work, we tune it on a hand-aligned development set to maximize the alignment F-score.

2. All candidates are sorted by soft votes in descending order and evaluated sequentially. A candidate link a_{jk} is included if one of the following is true:
   - Neither s_j nor t_k is aligned so far;
   - s_j is not aligned and its left or right neighboring word is aligned to t_k so far;
   - t_k is not aligned and its left or right neighboring word is aligned to s_j so far.

3. Repeat scanning all candidate links until no more links can be added.

In this way, those alignment links with higher confidence scores have higher priority to be included in the combined alignment.

4 Experiments

4.1 Baseline

Our training data contains around 70K English-Pashto sentence pairs released under the DARPA TRANSTAC project, with about 900K words on the English side. The baseline is a phrase-based MT system similar to (Koehn et al., 2003). We use GIZA++ (Och and Ney, 2000) to generate the baseline alignment for each direction and then apply grow-diagonal-final (*gdf*). The decoding weights are optimized with minimum error rate training (MERT) (Och, 2003) to maximize BLEU scores (Papineni et al., 2002). There are 2028 sentences in the tuning set and 1019 sentences in the test set, both with one reference. We use another 150 sentence pairs as a heldout hand-aligned set to measure the word alignment quality. The three sets of alignments described in Section 2 are generated on the same training data separately with GIZA++ and enhanced by *gdf* as for the baseline alignment. The English parse tree used for the syntactic reordering was produced by a maximum entropy based parser (Ratnaparkhi, 1997).

4.2 Improvement in Word Alignment

In Table 1 we show the precision, recall and F-score of each set of word alignments for the 150-sentence set. Using partial word provides the highest F-score among all individual alignments. The F-score is 5% higher than for the baseline alignment. The VP-based reordering itself does not improve the F-score, which could be due to the parse errors on the conversational training data. We experiment with three options (c_0, c_1, c_2) when combining the baseline and reordering-based alignments. In c_0, the weights w_i and confidence scores $c(a_{ijk}|S,T)$ in Eq. (3) are all set to 1. In c_1, we set confidence scores to 1, while tuning the weights with hill climbing to maximize the F-score on a hand-aligned tuning set. In c_2, we compute the confidence scores as in Eq. (1) and tune the weights as in c_1. The numbers in Table 1 show the effectiveness of having both weights and confidence scores during the combination.

Similarly, we combine the baseline with each of the other sets of alignments using c_2. They all result in significantly higher F-scores. We also generate alignments on VP-reordered partial words (X in Table 1) and compared $B + X$ and $B + V + P$. The better results with $B + V + P$ show the benefit of keeping the alignments as diversified as possible before the combination. Finally, we compare the proposed alignment combination c_2 with the heuristics-based method (*gdf*), where the latter starts from the intersection of all 4 sets of alignments and then applies grow-diagonal-final (Koehn et al., 2003) based on the links in the union. The proposed combination approach on $B + V + S + P$ results in close to 7% higher F-scores than the baseline and also 2% higher than

gdf. We also notice that its higher F-score is mainly due to the higher precision, which should result from the consideration of confidence scores.

Alignment	Comb	P	R	F
Baseline		0.6923	0.6414	0.6659
V		0.6934	0.6388	0.6650
S		0.7376	0.6495	0.6907
P		0.7665	0.6643	0.7118
X		0.7615	0.6641	0.7095
B+V	c_0	0.7639	0.6312	0.6913
B+V	c_1	0.7645	0.6373	0.6951
B+V	c_2	0.7895	0.6505	0.7133
B+S	c_2	0.7942	0.6553	0.7181
B+P	c_2	0.8006	0.6612	0.7242
B+X	c_2	0.7827	0.6670	0.7202
B+V+P	c_2	0.7912	0.6755	0.7288
B+V+S+P	*gdf*	0.7238	0.7042	0.7138
B+V+S+P	c_2	0.7906	0.6852	**0.7342**

Table 1: Alignment precision, recall and F-score (B: baseline; V: VP-based reordering; S: stemming; P: partial word; X: VP-reordered partial word).

4.3 Improvement in MT Performance

In Table 2, we show the corresponding BLEU scores on the test set for the systems built on each set of word alignment in Table 1. Similar to the observation from Table 1, c_2 outperforms c_0 and c_1, and $B + V + S + P$ with c_2 outperforms $B + V + S + P$ with *gdf*. We also ran one experiment in which we concatenated all 4 sets of alignments into one big set (shown as *cat*). Overall, the BLEU score with confidence-based combination was increased by 1 point compared to the baseline, 0.6 compared to *gdf*, and 0.7 compared to *cat*. All results are statistically significant with $p < 0.05$ using the sign-test described in (Collins et al., 2005).

5 Conclusions

In this work, we have presented a word alignment combination method that improves both the alignment quality and the translation performance. We generated multiple sets of diversified alignments based on linguistics, morphology, and heuristics, and demonstrated the effectiveness of combination on the English-to-Pashto translation task. We showed that the combined alignment significantly outperforms the baseline alignment with both higher F-score and higher BLEU score. The combination approach itself is not limited to any specific alignment. It provides a general framework that can take advantage of as many alignments as possible, which could differ in preprocessing, alignment modeling, or any other aspect.

Alignment	Comb	Links	Phrase	BLEU
Baseline		963K	565K	12.67
V		965K	624K	12.82
S		915K	692K	13.04
P		906K	716K	13.30
X		911K	689K	13.00
B+V	c_0	870K	890K	13.20
B+V	c_1	865K	899K	13.32
B+V	c_2	874K	879K	13.60
B+S	c_2	864K	948K	13.41
B+P	c_2	863K	942K	13.40
B+X	c_2	871K	905K	13.37
B+V+P	c_2	880K	914K	13.60
B+V+S+P	*cat*	3749K	1258K	13.01
B+V+S+P	*gdf*	1021K	653K	13.14
B+V+S+P	c_2	907K	771K	**13.73**

Table 2: Improvement in BLEU scores (B: baseline; V: VP-based reordering; S: stemming; P: partial word; X: VP-reordered partial word).

Acknowledgments

This work was supported by the DARPA TRANSTAC program. We would like to thank Upendra Chaudhari, Sameer Maskey and Xiaoqiang Luo for providing useful resources and the anonymous reviewers for their constructive comments.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In *Proc. HLT/NAACL*, June.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Chiang, Kevin Knight, Samad Echihabi, et al. 2009. Isi/language weaver nist 2009 systems. In *Presentation at NIST MT 2009 Workshop*, August.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL*, pages 531–540.

Yonggang Deng and Bowen Zhou. 2009. Optimizing word alignment combination for phrase table training. In *Proc. ACL*, pages 229–232, August.

Elliott Franco Drábek and David Yarowsky. 2004. Improving bitext word alignments via syntax-based reordering of english. In *Proc. ACL*.

Alexander Fraser and Daniel Marcu. 2007. Getting the structure right for word alignment: Leaf. In *Proc. of EMNLP*, pages 51–60, June.

Ulf Hermjakob. 2009. Improved word alignment with statistics and linguistic heuristics. In *Proc. EMNLP*, pages 229–237, August.

Fei Huang. 2009. Confidence measure for word alignment. In *Proc. ACL*, pages 932–940, August.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proc. of HLT/EMNLP*, pages 89–96, October.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. NAACL/HLT*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. of ACL*, pages 440–447, Hong Kong, China, October.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Martin Porter. 1980. An algorithm for suffix stripping. In *Program*, volume 14, pages 130–137.

Adwait Ratnaparkhi. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proc. of EMNLP*, pages 1–10.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. EMNLP*, pages 737–745.

Bing Xiang, Kham Nguyen, Long Nguyen, Richard Schwartz, and John Makhoul. 2006. Morphological decomposition for arabic broadcast news transcription. In *Proc. ICASSP*.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proc. NAACL/HLT*, pages 245–253, June.