

An Empirical Study on Improving Hierarchical Phrase-based Translation Using Alignment Features

Songfang Huang, Bowen Zhou

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

{shuang, zhou}@us.ibm.com

Abstract

In this paper, we empirically investigate three new features from word alignments to improve speech-to-speech translation on mobile devices for low-resource languages. The three features include one feature about alignment for boundary words of the target side phrase, one about the balance of terminal words between the source and the target side, and another about the number of unaligned words. We carry out experiments on both directions (E2F and F2E) for Pashto and Dari, two official languages of Afghanistan. By using the proposed alignment features, we can obtain improvements (up to 1% BLEU score) on the test sets for both Pashto and Dari.

Index Terms: hierarchical phrase-based translation, alignment, feature, empirical

1. Introduction

In recent years, speech-to-speech (S2S) translation has been an interesting application in spoken language process community. It benefits from the rapid advances in technologies for automatic speech recognition (ASR), statistical machine translation (SMT), and speech synthesis. Speech-to-speech translation on mobile devices is a more challenging task, which requires the algorithms and models to be more efficient in terms of time and space.

Hierarchical phrase-based translation [1], one example of syntax-based translation models, takes advantage of both phrase-based translation [2, 3] to handle local fluency, and the synchronous context-free grammars (SCFG) to handle non-local phrase reordering. It has been demonstrated to produce better performance than traditional phrase-based translation systems. The SCFG grammars are automatically extracted from parallel corpora based on hierarchical structures of natural language, without using any explicit linguistic knowledge or annotations. There are several important steps of the pipeline to translate from input E to its translation F in hierarchical phrase-based translation, including word alignment, phrase pair extraction, rule extraction, rule scoring, and decoding.

The task of phrase pair extraction is to extract those phrase pairs that are consistent with the word alignment. Due to the limit amount of available training data, however, there are normally some errors in the word alignment, for example, there exist some unaligned words. To compensate the limitation of word alignment, we in this paper empirically investigate three new features derived from word alignment to augment the quality of phrase translation table and consequently improve the translation performance. The three features include one feature about alignment for boundary words of the target side phrase, one about the balance of terminal words between the source and the

target side, and another about the number of unaligned words. Adding new features is a simple approach to incorporating additional knowledge into hierarchical phrase-based translation. We can also easily apply the augmented models on speech-to-speech translation on mobile devices using these low-cost alignment features.

Previous work by [4] studies the relationship between word alignment and phrase pair extraction. They compared different word alignments to the human generated alignment, and found that the number of unaligned words has a large impact on the quality of phrase translation table. They therefore introduced two new features, the number of unaligned words for the source side and the target side respectively, in a phrase-based translation system, and obtained an improvement in BLEU [5] scores.

Compared to previous work in [4], our work in this paper has some improvements and extensions. Firstly, we extract features from word alignment and apply them on hierarchical phrase-based translation instead of phrase-based translation. We will introduce how we deal with features for SCFG rules in Sec. 3. Secondly, we use more alignment-derived features, besides the number of unaligned words in [4]. We can see in Sec. 4 that the feature about the number of unaligned words alone does not work well in our experiments. Thirdly, we further verify the effectiveness of alignment features for hierarchical phrase-based translation on different language pairs. We carry out experiments on both directions (E2F and F2E) for Pashto and Dari, two official languages of Afghanistan. By using the proposed alignment features, we can obtain improvements (up to 1% BLEU score) on the test sets for both Pashto and Dari.

2. Hierarchical Phrase-based Translation

In this section, we briefly introduce some essential steps of the pipeline for a hierarchical phrase-based translation system [1].

2.1. Word Alignment

The goal of word alignment is to identify word-to-word links between parallel sentences. An alignment is a tuple (E, F, A) , where $E = e_1 \dots e_n$ and $F = f_1 \dots f_m$ are strings, and A is a set of links (x, y) between words where $1 \leq x \leq n$ and $1 \leq y \leq m$. The widely used techniques to automatically generate word alignments for parallel sentences include IBM Model 1 to Model 4 [6], and the approaches based on hidden Markov models (HMM) [7]. Figure 1 shows an example of a many-to-many alignment for a sentence pair and its matrix representation, with tight phrase pairs marked.

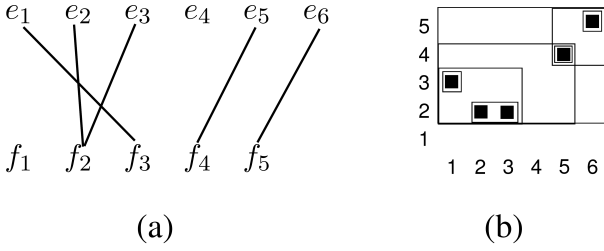


Figure 1: An example of (a) a many-to-many word alignment, and (b) the same alignment as a matrix representation, with its tight phrase pairs marked.

2.2. Phrase Pair Extraction

A phrase in SMT is a contiguous sequence of words. A pair of substring from E and F respectively, i.e., $[s, t] \subset E$ and $[u, v] \subset F$, is a phrase pair if and only if the subset of links emitted from $[s, t]$ in E is equal to the subset of links emitted from $[u, v]$ in F , and both are nonempty [2, 3, 8]. That is, all words of a phrase pair are aligned only within the phrases and not to any word outside. We call this phrase pair is *consistent* with the word alignment. The process of phrase pair extraction collects all phrase pairs that are consistent with word alignments, and forms the phrase translation table. In practice, we normally enforce a maximum phrase length during phrase extraction to reduce the number of phrases extracted from word alignment and keep phrase translation table manageable. For example, the following phrase pairs can be extracted from the alignment shown in Figure 1:

$$\begin{aligned}
 &([1, 1], [3, 3]), ([2, 3], [2, 2]), \\
 &([1, 3], [2, 3]), ([5, 5], [4, 4]), \\
 &([6, 6], [5, 5]), ([5, 6], [4, 5]), \\
 &([1, 5], [2, 4]), ([1, 6], [2, 5]), \\
 &([2, 3], [1, 2]), ([1, 3], [1, 3]), \\
 &([1, 5], [1, 4]), ([1, 6], [1, 5]), \\
 &([1, 4], [2, 3]), ([2, 4], [2, 2]), \\
 &([2, 4], [1, 2]), ([4, 5], [4, 4]), \\
 &([4, 6], [4, 5]).
 \end{aligned}$$

It is possible that some words of a phrase do not have alignment points (i.e., unaligned), e.g., e_4 and f_1 of the example in Figure 1. Having unaligned words in a phrase pair cannot violate the condition for consistency. Hence, unaligned words may occur within and even at the boundaries of a source/target phrase, e.g., $([2, 3], [1, 2])$. A phrase pair $([s, t], [u, v])$ is *tight* if none of e_s, e_t, f_u and f_v is unaligned [8]. Only the first eight phrase pairs shown in the above example are tight. We can see that allowing unaligned words in the alignment largely increases the number of phrase pairs extracted from that. It has been shown in [4] that an increased number of unaligned words can cause degraded translation quality. However, it is also not wise to prune out all those non-tight phrase pairs. This motivates us to introduce some new features to set some soft constraints to compensate the limitation (e.g., unaligned words) of word alignments.

2.3. Rule Extraction

An SCFG is a synchronous rewriting system generating source and target side string pairs simultaneously based on a context-free grammar. Each synchronous production (i.e., rule) rewrites

a non-terminal into a pair of strings, γ and α , with both terminals and non-terminals in both languages. In particular, hierarchical phrase-based models [1] explore hierarchical structures of natural language and utilize only a unified non-terminal symbol X in the grammar:

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle \quad (1)$$

where $\sim$ is the one-to-one correspondence between X 's in γ and α .

The phrase pairs extracted from word alignments are regarded as *phrasal rules* (i.e., without non-terminal), being part of the whole SCFG rules. Next, we loop through each phrase pair and generalize the sub-phrase pair it contains with co-indexed non-terminal symbols. We thereby obtain a new rule. There are several constraints described in [1] to reduce the rule size and spurious ambiguity, e.g., there are at most two non-terminals in a rule. We will hereafter refer to rules with non-terminal symbols as *abstract rules*.

2.4. Decoding

Translation using SCFG grammars for an input sentence E is casted as to find the optimal derivation on the source and target sides (as the grammar is synchronous, the derivations on source and target sides are identical). That is, the goal of decoding is to search for the best derivation D that maximizes the following log-linear model over all possible derivations:

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times \prod_i \prod_{X \rightarrow \langle \gamma, \alpha \rangle \in D} \phi_i(X \rightarrow \langle \gamma, \alpha \rangle)^{\lambda_i}, \quad (2)$$

where the set of $\phi_i(X \rightarrow \langle \gamma, \alpha \rangle)$ are features defined over given production rule, and $P_{LM}(e)$ is the language model score on hypothesized output, the λ_i is the feature weight. A bottom-up CKY parser is widely used for decoding.

3. Alignment Features

In this section, we introduce our proposed new features from word alignment for hierarchical phrase-based translation.

3.1. Features for Phrasal Rules

We derive three new features from word alignments for phrasal rules. Since we typically calculate Equation 2 in the logarithmic scale, we interpret these features as *cost* for SCFG rules. The three alignment features (costs) are as follows:

- C1: cost about the alignment for boundary words on the *target* side of phrase pairs. We use this feature to evaluate the *tight vs non-tight* status of a phrase pair. For a tight phrase pair, we assign a cost value of 0 to it. For a non-tight phrase pair, we further divide into two cases: by assigning a cost of 0.5 to a phrase pair with *either* one of the boundary words of the target side phrase unaligned, and a cost of 1.0 to a phrase pair with *both* boundary words of the target side phrase unaligned. Intuitively we are in favor of tight phrase pairs. By assigning a cost to those non-tight phrase pairs, we enforce a soft constraint over them rather than removing them entirely.
- C2: cost about the the balance of terminal words between the source and the target side. Mathematically, $C2 = |N_s - N_t|$ where N_s and N_t are the numbers of

terminals for the source side and the target side, respectively. The intuition is that we want to penalize those phrase pairs with unbalanced number of terminals between phrases of the source side and the target side, e.g., $([1, 1], [1, 9])$.

- $\mathbb{C}3$: cost about the total number of unaligned terminal words in both the source side and the target side. Previous work also used features about unaligned words for phrase pairs and obtained some improvements in BLEU [4]. The difference is that there are two separate features used in [4], one for the source side and the other for the target side, respectively.

3.2. Features for Abstract Rules

It is straightforward to derive the three new features from word alignment for abstract rules. For $\mathbb{C}1$, we assign a cost value to an abstract rule according to the phrase pair from which the abstract rule is extracted. Since we only extract abstract rules from those tight phrase pairs that contain other tight phrase pairs [1], all abstract rules will have a cost value of 0 for $\mathbb{C}1$. For $\mathbb{C}2$, given that an abstract rule always has co-indexed non-terminal symbols, it is equal to $|L_s - L_t|$ where L_s and L_t are the length for the source side and the target side of an abstract rule. For $\mathbb{C}3$, we can easily count the total number of unaligned terminal words and assign costs by referring to word alignments during rule extraction.

We note that a same phrasal rule can be extracted from different parallel sentences, and a same abstract rule can be extracted from different phrase pairs too. When combining these to get a distinct rule, we select the maximum value as the final cost for the rule.

The discriminative translation framework expressed in Equation 2 makes it relatively straightforward to incorporate the three new features into a hierarchical phrase-based translation system.

4. Experiments and Results

The baseline translation system we used in this paper is a state-of-the-art hierarchical phrase-based translation system, as described in detail in [9]. The system is designed and optimized for conversational speech-to-speech translation tasks on mobile devices. We used interpolated 4-gram language models with modified Kneser-Ney smoothing [10] for all the following experiments. Minimum-error-rate training (MERT) [11] was carried out to optimize the feature weights on the development set. We conducted experiments on both directions for Pashto and Dari. All development and test tests in the experiments have only one reference. Unless explicitly stated otherwise, the maximum length of phrase pairs is set to 8, and the baseline model has seven basic features, including four rule and lexicalized translation probabilities plus three discounting features.

4.1. Experiments on Pashto

By duplicating some in-domain components of the Pashto corpus, we have a training set with 200k parallel sentences. There are 28k unique words for English and 51k unique words for Pashto.

For E2F (i.e., English-to-Pashto) direction, we have a held-out development set and a test set, consisting of 2k and 1k sentences respectively. The baseline model has four basic features, such as rule and lexicalized translation probabilities.

Table 1 shows the BLEU results on E2F for Pashto. In this experiment, we evaluated the effects of each individual feature and the combinations, respectively. From the results, we can see that the feature $\mathbb{C}3$ (the total number of unaligned words) alone performs worse than $\mathbb{C}1$ and $\mathbb{C}2$, and $\mathbb{C}2$ produces the best BLEU score on this test set. For the combinations, it seems that $\mathbb{C}2+\mathbb{C}3$ does not work well on this test set. We note that the improvement on the development set is less significant than that on the test set. Overall, we can improve the baseline model on the test set by using some specific alignment features.

Table 1: The BLEU results (%) of E2F (i.e., English-to-Pashto) on the Pashto corpus.

Model	Dev	Test	Δ Test
Baseline	17.95	18.19	-
Baseline+ $\mathbb{C}1$	18.31	18.81	0.62
Baseline+ $\mathbb{C}2$	18.24	18.92	0.73
Baseline+ $\mathbb{C}3$	18.08	18.40	0.21
Baseline+ $\mathbb{C}1+\mathbb{C}2$	18.36	18.81	0.62
Baseline+ $\mathbb{C}1+\mathbb{C}3$	17.99	18.69	0.50
Baseline+ $\mathbb{C}2+\mathbb{C}3$	18.26	18.16	-0.03
Baseline+ $\mathbb{C}1+\mathbb{C}2+\mathbb{C}3$	17.92	18.25	0.06

For F2E (i.e., Pashto-to-English) direction, we have a held-out development set and a test set, consisting of 3.1k and 1.6k sentences respectively.

Table 2 shows the BLEU results on F2E for Pashto. By using the combinations of the three features, we are able to obtain consistent improvements on the test set, ranging from 0.42% to 0.64%. We note that the effect of $\mathbb{C}2+\mathbb{C}3$ on F2E (+0.64%) is different from that on E2F (-0.03%), although they are not directly comparable given the different training, tuning, and test sets.

Table 2: The BLEU results (%) of F2E (i.e., Pashto-to-English) on the Pashto corpus

Model	Dev	Test	Δ Test
Baseline	20.88	22.98	-
Baseline+ $\mathbb{C}1+\mathbb{C}2$	21.24	23.40	0.42
Baseline+ $\mathbb{C}1+\mathbb{C}3$	20.92	23.40	0.42
Baseline+ $\mathbb{C}2+\mathbb{C}3$	21.18	23.62	0.64
Baseline+ $\mathbb{C}1+\mathbb{C}2+\mathbb{C}3$	21.08	23.43	0.45

4.2. Experiments on Dari

We have a training corpus consisting of 150k parallel sentences for Dari experiments, after duplicating some in-domain components of the Dari corpus. The numbers of unique words for English and Dari are 26k and 48k, respectively.

For E2F (i.e., English-to-Dari) direction, we have 1.6k sentences for both the held-out development set and the test set.

Table 3 shows the BLEU results on E2F for Dari. In this experiments, $\mathbb{C}1+\mathbb{C}2+\mathbb{C}3$ produces the best BLEU score on the test set, while it has a worse result on the development set. We can also see from Table 1 and 2 that $\mathbb{C}1+\mathbb{C}2+\mathbb{C}3$ generally has lower BLEU scores on the development set. This may indicate that it becomes more difficult to tune the features weights as the number of features to tune increases. Although we have worse BLEU scores on the development set in some cases, the

improvements on the test set are all positive on E2F for Dari across the four feature combinations.

Table 3: The BLEU results (%) of E2F (i.e., English-to-Dari) on the Dari corpus.

Model	Dev	Test	Δ Test
Baseline	15.40	14.99	-
Baseline+C1+C2	15.18	15.12	0.13
Baseline+C1+C3	15.51	15.47	0.48
Baseline+C2+C3	15.54	15.20	0.21
Baseline+C1+C2+C3	15.32	15.73	0.74

For F2E (i.e., Dari-to-English) direction, we have 1.3k sentences for both the held-out development set and the test set. The maximum length of phrase pairs is set to 15.

Table 4 shows the BLEU results on F2E for Dari. In this experiment, we observe an improvement with up to 1% BLEU score on the test set using C1+C3. The other three feature combinations are also improving the baseline model on both the development set and the test set.

Table 4: The BLEU results (%) of F2E (i.e., Dari-to-English) on the Dari corpus.

Model	Dev	Test	Δ Test
Baseline	17.72	17.68	-
Baseline+C1+C2	18.08	17.98	0.30
Baseline+C1+C3	18.27	18.67	0.99
Baseline+C2+C3	17.94	17.86	0.18
Baseline+C1+C2+C3	18.40	18.53	0.85

5. Discussion

For low-resource languages such as Pashto and Dari, the relatively small size of the training corpora will increase the likelihood of incorrect word alignments, which will consequently decrease the quality of the phrase translation tables. The alignment features we introduce in this paper can, to some extent, alleviate this problem. This has been demonstrated by the experimental results on Pashto and Dari.

We note that adding new alignment features to a hierarchical translation system is a low-cost process. We can combine the features with the tuned weights, and store one single pre-computed feature value for each SCFG rule in the model. This makes it possible to apply the augmented models to speech-to-speech translation on mobile devices.

From the experiments, we observe smaller improvements on the development set than those on the test set, which suggests that the new alignment features may somewhat prevent the translation model from over-fitting the tuning set. This is the same observation as that in [4]. Another reason for smaller improvements on the development set may be that it becomes more difficult to tune the weights using MERT [11], as we add more features.

We can not identify from the experimental results in Sec. 4 a common feature, or a feature combination, that works best across both directions (E2F and F2E) of both language pairs (English $\leftrightarrow$ Pashto, and English $\leftrightarrow$ Dari). We believe this is a common observation given the fact that a log-linear model is used for SMT and the tuning of feature weights is a complex

process. We also need to take into account the differences of languages, the amount of training data, etc. However, we can always find an alignment feature, or a feature combination, that improves the baseline model, for all the experimental conditions we tried in Sec.4.

6. Conclusion

The conclusions we make in this paper are as follows. Firstly, the quality of word alignments is directly affected by the amount of training data in SMT, especially for those low-resource languages that have only small amount of data for training. Secondly, unaligned words occur in word alignments, which directly affects the quality of extracted phrase pairs. Thirdly, we can improve the performance of a state-of-the-art hierarchical phrase-based translation system, by using some features derived from word alignments as soft constraints for phrase pairs.

7. Acknowledgements

This work is partially supported by the DARPA TRANSTAC program under the contract number NBCH2030007. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. References

- [1] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.
- [2] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," in *North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, 2003.
- [3] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.
- [4] F. Guzman, Q. Gao, and S. Vogel, "Reassessment of the role of phrase extraction in PBSMT," in *Proc. of MT Summit XII*, Ottawa, Canada, 2009.
- [5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.
- [6] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [7] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proceedings of the 16th conference on Computational linguistics - Volume 2*, ser. COLING '96, 1996, pp. 836–841.
- [8] H. Zhang, D. Gildea, and D. Chiang, "Extracting synchronous grammar rules from word-level alignments in linear time," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08, 2008, pp. 1081–1088.
- [9] B. Zhou, B. Xiang, X. Zhu, and Y. Gao, "Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels," in *Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, 2008.
- [10] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.
- [11] F. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, 2003, pp. 160–167.