Interlocking Phrases in Phrase-based Statistical Machine Translation

Ye Kyaw Thu, Andrew Finch and **Eiichiro Sumita**
Multilingual Translation Lab.,
Advanced Speech Translation Research and Development Promotion Center,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, JAPAN
{yekyawthu, andrew.finch, eiichiro.sumita}@nict.go.jp

Abstract

This paper presents an study of the use of interlocking phrases in phrase-based statistical machine translation. We examine the effect on translation quality when the translation units used in the translation hypotheses are allowed to overlap on the source side, on the target side and on both sides. A large-scale evaluation on 380 language pairs was conducted. Our results show that overall the use of overlapping phrases improved translation quality by 0.3 BLEU points on average. Further analysis revealed that language pairs requiring a larger amount of re-ordering benefited the most from our approach. When the evaluation was restricted to such pairs, the average improvement increased to up to 0.75 BLEU points with over 97% of the pairs improving. Our approach requires only a simple modification to the decoding algorithm and we believe it should be generally applicable to improve the performance of phrase-based decoders.

1 Introduction

In this paper we examine the effect on machine translation quality of using interlocking phrases to during the decoding process in phrase-based statistical machine translation (PBSMT). The motivation for this is two-fold.

Firstly, during the phrase-pair extraction process that occurs in the training of a typical PBSMT system, all possible alternative phrase-pairs are extracted that are consistent with a set of alignment points. As a consequence, the source and target sides of these extracted phrase pairs may overlap. However, in contrast to this, the decoding process traditionally proceeds by concatenating disjoint translation units; the process relies on the language model to eliminate awkward hypotheses with repeated words produced by sequences of translation units that overlap.

Secondly, the transduction process in PBSMT is carried out by generating hypotheses that are composed of sequences of translation units. These sequences are normally generated independently, as modeling the dependencies between them is difficult due to the data sparseness issues arising from modeling with word sequences. The process of interlocking is a way of introducing a form of dependency between translation units, effectively producing larger units from pairs of compatible units.

2 Related Work

(Karimova et al., 2014) presented a method to extract overlapping phrases offline for hierarchical phrase based SMT. They used the CDEC SMT decoder (Dyer et al., 2010) that offers several learners for discriminative tuning of weights for the new phrases. Their results showed improvements of 0.3 to 0.6 BLEU points over discriminatively trained hierarchical phrase-based SMT systems on two datasets for German-to-English translation. (Tribble and et al., 2003) proposed a method to generate longer new phrases by merging existing phrase-level alignments that have overlaping words on both source and target sides. Their experiments on translating Arabic-English text from the news domain were encouraging.

Proceedings of NAACL-HLT 2016, pages 1076–1081,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

(Roth and McCallum, 2010) proposed a conditional-random-field approach to discriminatively train phrase based machine translation in which training and decoding are both cast in a sampling framework. Different with traditional PBSMT decoding that infers both a Viterbi alignment and the target sentence, their approach produced a rich overlapping phrase alignment. Their approach leveraged arbitrary features of the entire source sentence, target sentence and alignment. (Kääriäinen, 2009) proposed a novel phrase-based conditional exponential family translation model for SMT. The model operates on a feature representation in which sentence level translations are represented by enumerating all the known phrase level translations that occur inside them. The model automatically takes into account information provided by phrase overlaps. Although both of the latter two approaches were innovative the translation performance was lower than tranditional PBSMT baselines.

Our proposed approach is most similar to that of (Tribble and et al., 2003). Our approach differs in the interlocking process is less constrained; phrase pairs can interlock independently on source and target sides, and the interlocking process performed during the decoding process itself, rather than by augmenting the phrase-table.

3 Methodology

3.1 Target Interlocking

In the decoding process for PBSMT, the target is generated from left-to-right phrase-by-phrase. The process of interlocking the phrases is illustrated in Figure 1. The s_i are the source tokens, the t_j are the target tokens, the lower target token sequence on the left represents the partial translation hypothesis, and the upper target phrase is the target side of a translation unit $(s_3 s_4, t_3 t_4 t_5)$ being used to extend the hypothesis. An interlock of length k is can occur if the last k tokens of the partial translation match the first k tokens of the target side of the translation unit being used to extend the hypothesis. In this case the decoder may create an extended hypothesis with the target side of the translation unit interlocked with the target word sequence generated so far. In order to do this, the k interlocked words are not inserted a second time into the target token sequence and the word penalty (if pre-calculated) is adjusted to reflect this. In the example given in Figure 1, the translation resulting from extending the search with the interlocking translation unit will be $t_1 t_2 t_3 t_4 t_5$.

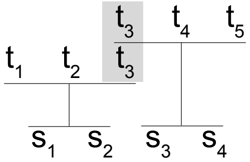

Figure 1: Interlocking target phrases.

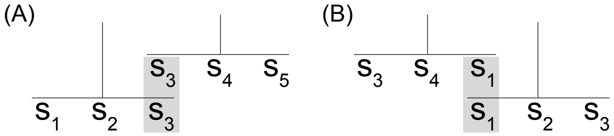

Figure 2: Interlocking source phrases.

3.2 Source Interlocking

The interlocking of source phrases can occur in three different ways as shown in Figure 2. In Figure 2 (A), a source phrase is interlocking with source words to the left; in (B) a source phrase is interlocking with source words to the right; and in (C) a source phrase is interlocking on both sides. The interlocking process is handled before the search process begins, at the time the set of translation options used in the search is created. Additional interlocking translation options are created in which the source side phrase is permitted to overlap with the surrounding source context, however, later during the search this translation unit will only be used to translate (cover) the sequence of non-interlocking words. In this way, the decoder's search algorithm can be used without modification, when dealing with interlocking source phrases.

4 Experiments

4.1 Corpora

We used twenty languages from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions (Kikui et al., 2003). The languages were Arabic (ar), Danish (da), German (de), English (en), Spenish (es), French (fr), Italian (it), Dutch (nl), Portugese (pt), Russian (ru), Tagalog (tl), Indonesian (id), Malaysian (ms), Vietnamese (vi), Thai (th), Hindi (hi), Chinese (zh), Japanese (ja), Korean (ko) and Myanmar (my). 155,121 sentences were used for training, 5,000 sentences for development and 2,000 sentences for evaluation.

In addition, we ran experiments on two language pairs from the Europarl corpus (Koehn, 2005). The language pairs were English-German, German-English, English-Spanish and Spanish-English. The corpus statistics are given in Table 2.

4.2 Experimental Methodology

We used a modified version of our in-house phrase based SMT system which operates similarly to Moses (Koehn and Haddow, 2009). GIZA++ (Och and Ney, 2000) was used for word alignment, together with the grow-diag-final-and heuristics (Koehn et al., 2003). A lexicalized reordering model was trained with the msd-bidirectional-fe option (Tillmann, 2004). We used the SRILM toolkit to create 5-gram language models with interpolated modified Kneser-Ney discounting (Stolcke, 2002; Chen and Goodman, 1996). The weights for the log-linear models were tuned using the MERT procedure (Och, 2003). The translation performance was evaluated using the BLEU score (Papineni et al., 2001).

We ran three sets of experiments; (1) target interlocking, (2) source interlocking and (3) both source and target interlocking for all possible combinations of languages (i.e. 380 language pairs). We studied two methods for accomplishing (3). In the first, interlocking as defined in Sections 3.1 and 3.2 are permitted freely. In the second, the target is allowed to interlock if and only if the source is also interlocked. This was similar to the method proposed by (Tribble and et al., 2003).

4.3 Results

In this section, we will first present the results of the experiments on the BTEC corpora and then report the results from the experiments from the Europarl corpus.

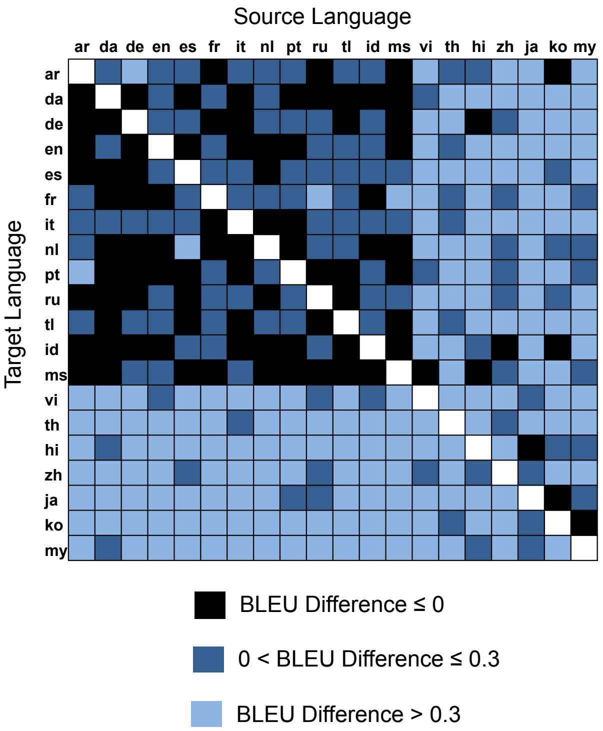

Figure 3: BLEU difference by language pair.

4.3.1 BLEU Differences

The difference in BLEU between a baseline system, a standard phrase-based SMT system without interlocking, and the proposed systems in which interlocking phrases where permitted, was calculated, and the average taken over all 380 language pairs. The results show that interlocking the phrases generally improves translation quality, and that the system gained slightly more from interlocking the target phrases, than from interlocking the source phrases. The average BLEU difference was 0.22 from interlocking target phrases, 0.14 from overlapping source phrases and 0.33 from interlocking both source and target. When the interlocking was constrained to ensure that both source and target phrases were interlocked, the average BLEU difference dropped to 0.27 BLEU. In the cases where both source and target phrases were allowed to interlock freely, 77% of the experiments showed an improvement in BLEU score.

A sequence of experiments were run on the baseline system with increasing stack size from 100, to 1000 in increments of 100. These experiments showed an increase of 0.07 BLEU points from stack

Src-Trg	**Corpus Statistics (sentences)**			**Baseline**	**Interlocking**		
	train	develop	test	No-interlock	Source	Target	Src & Trg
en-de	1500,000	3,000	3,000	18.70	**21.03**	18.83	18.90
de-en	1500,000	3,000	3,000	**26.08**	24.86	26.05	25.44
en-es	1500,000	3,000	3,000	33.55	**35.44**	33.68	33.89
es-en	1500,000	3,000	3,000	33.81	**36.42**	33.90	33.92

Table 2: BLEU scores for the Europarl corpora.

Kendall's Tau Distance	**Avg BLEU Difference**	**% Expts Showing Gain**
[0, 1.00]	0.33	77.3
[0, 0.95]	0.34	78.4
[0, 0.90]	0.38	80.5
[0, 0.85]	0.51	90.5
[0, 0.80]	0.58	93.5
[0, 0.75]	0.63	97.3
[0, 0.70]	0.68	97.3
[0, 0.65]	0.71	97.7
[0, 0.60]	0.72	97.2
[0, 0.55]	0.74	97.3
[0, 0.50]	0.75	100.0

Table 1: Filtering the Set of Language Pairs.

size 100 to stack size 200, followed by a sequence of scores that did not vary more than 0.01. Therefore, we conclude that the gains we obtained through interlocking the phrases, could not have been obtained by simply increasing the amount of searching performed by the baseline system. In other words, the interlocking method is introducing novel and useful search steps into the search space.

4.3.2 Results by Language Pair

Figure 3 shows how the gains and losses in BLEU score were distributed over the set of language pairs. Lighter cells in the figure represent gains in BLEU, the black cells represent losses. The order of the language pairs has been arranged to show the is a clear pattern. The languages on the left hand side and upper part of the figure are mostly European languages with similar word orders, whereas the remaining languages are typically Asian languages with different word orders. The language pairs that gained the most from interlocking on both source and target were: th-hi, hi-it, ko-de, ko-tl, th-ja, my-ru, my-th, my-ar, th-my, and it-ja. The languages that lost the most in BLEU score were: id-pt, ko-my, fr-id, ms-pt, id-it, da-ar, id-ko, da-es, id-ar, and de-it.

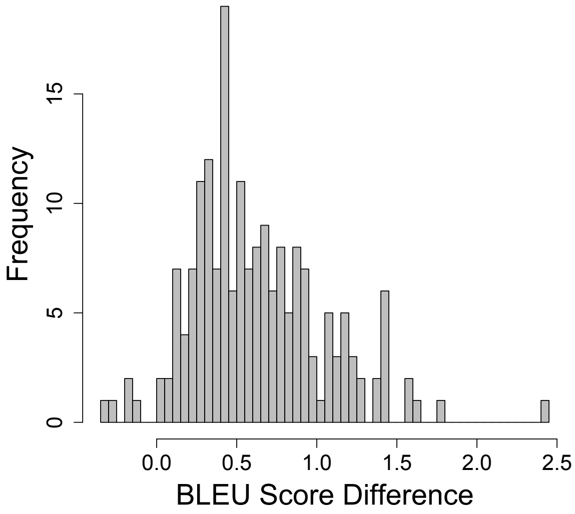

Figure 4: The Distribution of differences in BLEU score.

It is clear from Figure 3 that translation among the group of similar languages does not benefit from our approach, but the dissimilar languages do. This observation motivated further analysis of the data in order to develop a method for selecting language pairs

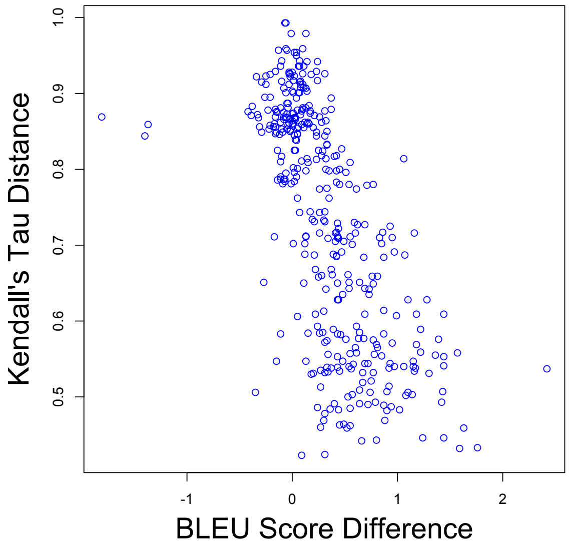

Figure 5: Plot of the Kendall's tau distance difference against BLEU difference.

suitable for our approach.

4.3.3 Kendall's Tau Distance

Kendall's tau distance is the minimum number of transpositions of adjacent symbols necessary to transform one permutation into another (Kendall, 1938; Birch, 2011), and is one method to gauge the amount of re-ordering that would be required during the translation process between two languages.

Figure 5 shows a scatter plot of all of the experiments, plotting BLEU difference against Kendall's tau. The points show a strong negative correlation (coefficient: -0.7). Therefore, we propose to use Kendall's tau as a means of selecting appropriate language pairs to be used with our method.

Table 1 shows the effect of filtering the set of language pairs by Kendall's tau. The effectiveness of the proposed method increases as languages with higher Kendall's tau distance are removed from the experimental set. When language pairs are selected according to Kendall's tau in the range $0 \leq \tau \leq 0.75$, the average BLEU gain of the set increases to 0.6 BLEU while still retaining approximately half of the language pairs in the set. Moreover, the proportion of experiments showing an in improvement in BLEU increases to over 97%. Figure 4 shows the distribution of BLEU differences for this subset of language pairs.

4.3.4 Europarl Corpus

The results of the previous sections were all based on experiments on the BTEC corpus. This corpus is unusual in that the sentences are short and the training data size is also small. In order to establish that our approach has more general application, we applied it to four language pairs from the much larger Europarl corpus. The results on the Europarl corpus are shown in Table 2. For three of the language pairs we observed increases in BLEU scores over the baseline for all interlocking methods with substantial gains of 1.9 to 2.6 BLEU points coming from the source interlocking technique. However, the German to English pair gave a negative result. The results from the Europarl corpus are generally very encouraging but the negative result motivates further study on more language pairs from different domain in the future.

5 Conclusion

In this paper we propose and evaluate a simple technique for improving the performance of phrase-based statistical machine translation decoders, that can be implemented with only minor modifications to the decoder. In the proposed method phrases are allowed to interlock freely on both the source and target side during decoding. The experimental results, based on a large-scale study involving 380 language pairs provide strong evidence that our approach is genuinely effective in improving the machine translation quality. The translation quality improved for 77% of the language pairs tested, and this was increased to over 97% when the set of language pairs was filtered according to Kendall's tau distance. The translation quality improved by an average of up to 0.75 BLEU points on this subset. This value represents a lower bound on what is possible with this technique and in future work we intend to study the introduction of additional features into the log-linear model to encourage or discourage the use of interlocking phrases during decoding, and investigate the effect of increasing the number of interlocked words.

References

Alexandra Birch. 2011. *Reordering Metrics for Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Stanley F Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL*.

Matti Kääriäinen. 2009. Sinuhe – statistical machine translation using a globally trained conditional exponential family translation model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1036, Singapore, August. Association for Computational Linguistics.

Sariya Karimova, Patrick Simianer, and Stefan Riezler. 2014. Offline extraction of overlapping phrases for hierarchical phrase-based translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 236–243.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proceedings of EUROSPEECH-03*, pages 381–384.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164.

Philipp Koehn, Franz Josef Och, , and Daniel Marcu. 2003. Statistical phrase-based translation. In *In Proceedings of the Human Language Technology Conference*, Edmonton, Canada.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hong Kong, China.

Franz J. Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2001. *Bleu: a Method for Automatic Evaluation of Machine Translation*. IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center.

Benjamin Roth and Andrew McCallum. 2010. Machine translation using overlapping alignments and samplerank. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas, AMTA2010*.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alicia Tribble and et al. 2003. Overlapping phrase-level translation rules in an smt engine.