Overlapping phrase-level translation rules in an SMT engine

OVERLAPPING PHRASE-LEVEL TRANSLATION RULES IN AN SMT ENGINE

Alicia Tribble, Stephan Vogel and Alex Waibel

Interactive Systems Laboratory
Carnegie Mellon University
Pittsburgh, PA
{atribble, stephan.vogel, ahw}@cs.cmu.edu

ABSTRACT

In this paper we explore a technique for adding longer phrase translation pairs to a Statistical Machine Translation (SMT) system. New phrases are generated by merging existing phrase-level alignments that have overlapping words on both the source and target sides. The effect on translation quality is reported for an Arabic-English system in the news domain.

1. INTRODUCTION

In this paper we explore a technique for adding longer phrase translation pairs to a Statistical Machine Translation (SMT) system. The current system is based on the work of [1], and relies on word- and phrase-level alignments in the form of stochastic transducer rules. In some cases, the phrase-level rules overlap in their source and target coverage and could be merged to create new, longer rules. In this paper we explore the effects of generating these merged rules, or overlapping phrases (OPs), and we demonstrate that augmenting the rule base through such merges makes the translation output more accurate.

The rest of this section gives some background, discussing phrase-level translation in SMT systems and presenting a motivating example from another direct MT approach, Example-Based Machine Translation. In section 2 we describe the application of Overlapping Phrases in our SMT system. Section 3 gives some translation results using the new phrases and section 4 concludes the paper.

1.1. SMT and Phrase-level Translation

Finding and using phrase-level alignment information has long been a challenge for statistical MT systems. Traditional systems built along the source-channel model presented by [2] rely on word-level translation models that give the probability of a source word, f, given a target word, e. Sentence-level translations are assigned probabilities based on the combination of these translation model scores with a language model score and other scores modeling features like position or fertility.

This dependence on word alignment models contributes to the translation errors that statistical MT systems typically make: the content words may be present, but long sequences of fluent-sounding text are rare.

Solutions proposed by several groups including [3] and [4] focus on building phrase-level aligment models where alignment scores for entire phrases that appear in the training data are calculated. Translating with M-N alignments gives significant improvements in translation quality because the phrase-level rules can capture word reordering and other multiword phenomena that are difficult or impossible for a word translation model.

The motivation for phrase-level alignment mod-

Table 1. Example Transducer Rule

@Phrase	#	APAr Alwzyr Aly An	#	minister pointed out that	#	0.0730747
@Phrase	#	An AlSyn stwASl	#	that china will continue	#	8.63074e-07
@NewPhrase	#	APAr Alwzyr Aly An AlSyn stwASl	#	minister pointed out that china will continue	#	3.0991e-08
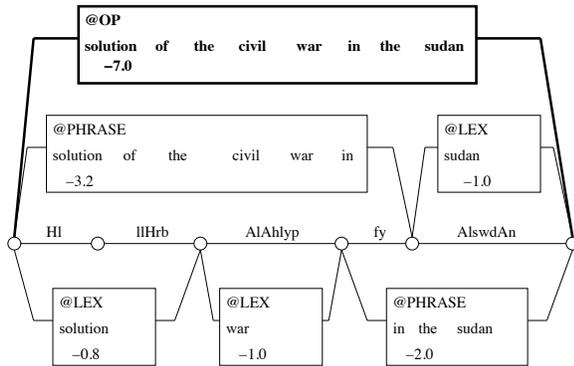
Fig. 1. OP in a Translation Lattice

els applies to the Overlapping Phrases experiments, as well: by generating merged rules as described here, we are able to leverage even more of the power of phrase-to-phrase alignments and capture longer, more fluent subsentential translation sequences.

Figure 1 gives an example of why this is so. Our baseline system works by generating non-overlapping paths through a translation lattice and comparing the total translation cost of each complete path. The un-translated input sentence is represented by the flat path *Hl llHrb AlAhlyp fy AlswdAn*. Although the phrase translations "in the sudan" and "solution of civil war in" were both generated as phrase translation candidates, they lie on different translation paths because they overlap. The new edge shown in bold in Figure 1 demonstrates how allowing overlap can generate useful merges of existing phrase translations.

1.2. Overlap Experiments in EBMT

Our experiments with overlapping phrases were also motivated by the success of recent work in Example-Based Machine Translation (EBMT). EBMT, like Sta-

tistical Machine Translation, is a direct translation approach. The EBMT engine stores and indexes a parallel training corpus and builds up translations for test sentences based on patterns or literal examples found in the training text.

An example of such a system is described in [5]. In the baseline system, a source sentence can only be translated by sub-phrases covering strictly sequential parts of the text; overlapping sub-phrases can not be used to translate a sentence.

In [6], the authors modify the EBMT engine to use phrase candidates that overlap on both source and target side. The modification is made in the search algorithm for an optimal combination of sub-phrases to cover an entire input sentence. In experiments on the French-English Hansard Corpus, their approach showed as much as a 15% increase in translation quality as measured by the BLEU and NIST metrics for automatic MT evaluation [7, 8].

2. OVERLAPPING PHRASES IN THE SMT ENGINE

2.1. Generating New Rules

Our translation system, described in [9], stores alignment model scores in the form of weighted transducer rules. Table 1 shows example transducer rules that overlap in similar fashion to the edges of the lattice in Figure 1.

This example comes from a transducer trained on transliterated Arabic news text and its English counterpart. The table shows how new rules are generated from pairs of existing rules that overlap by at least one word on both the source and target sides. In our experiments, the length of the overlap is allowed to vary

from 1 to 4 words, with different lengths in this range allowed on source and target sides.

To allow these OPs to be used by the decoder for translation, we first generate offline all of the Overlapping Phrases that appear in an existing transducer. After adding translation probabilities to the new phrases, we provide this file to the translation program as an additional transducer. This implementation has allowed us to make a series of experiments quickly and to assess the effect of OPs on translation quality. It will be followed in the near future by an implementation which is integrated with the translation decoder.

When generating OPs from a file of phrase translation pairs, we index every pair in the file according to its source-side prefix and suffix. Next we examine all pairs of rules where the prefix of one matches the suffix of the other. For every such rule pair, the source sides are merged and the resulting phrase is checked for usability against a list of valid source N-grams (taken from the expected test source).

If the merged source phrase is usable, then the the target sides of the phrase pair are checked for overlap of 1 to 4 words (taking the maximal in case of more than one possible overlap length). When the target sides of such a phrase pair also overlap, then we merge the two targets and generate a new OP translating the merged source into the merged target.

This process may generate a large number of new alignment rules, depending on the number and length of the original rule base and on the vocabulary size. Section 3 gives some statistics of the new phrases generated during our experiments.

2.2. Unseen Translation Rules

An important feature of Overlapping Phrases is that they add generalization power to the translation rule base. They allow phrase translation pairs to be created for phrases that were never seen in training but may be applicable to new test data. Again, Table 1 provides an example. Although *minister pointed out that* and *that china will continue* appear in the original transducer, the full phrase *minister pointed out that china*

will continue does not. By generating the OP transducer we are able to store a translation for this long phrase and apply it when its source appears in the test data.

2.3. Assigning New Rule Probabilities

We assign the probabilities for the new candidates with the help of a word-level alignment model trained in the style of IBM model 1.

The alignment score for source phrase $\tilde{f}$ and target phrase $\tilde{e}$ is given by

$$p(\tilde{f}|\tilde{e}) = \prod_j \sum_i p(f_j|e_i)$$

where $p(f_j|e_i)$ for source word f_j and target word e_i is given by the IBM model.

Previous work in phrase translation has preferred to assign these probabilities using the relative frequencies of the phrase pairs themselves, as in [4] for example. Our choice of word-model probabilities is motivated by two factors: First, the phrase pairs are typically not seen frequently enough in training to give reliable counts. The unseen translation pairs discussed above provide an example. Because they are never seen during training, their probability according to relative frequency would be the same default value assigned to all unknown or out-of-vocabulary items. Scores assigned by the word model described above are much more meaningful and help to differentiate the new rules from each other.

Second, since word-to-word lexica trained in the IBM style are used along with phrase rules during translation, assigning phrase translation probabilities according to word models makes these two components fairly comparable. A slight advantage is still given to the phrase-level translations as a result of the summation over all alignments in $p(\tilde{f}|\tilde{e})$. Translating the same phrase at decoding time using the lexicon alone would give only the Viterbi Alignment score. Hence, preference is given to phrase-level translation while still allowing the lexica to play a role.

3. EXPERIMENTS

3.1. Data and Phrase Translation Methods

All of these experiments were conducted on Arabic news-domain test data. The test data contains 203 sentences, with a vocabulary size of 2,273 types and a total of 4,906 tokens. Four reference translations were used for automatic scoring of this data. The training corpus was composed of approximately 6 million words of parallel news text in Arabic and English, covering 95.6% of the test tokens and 98.4% of the types.

In order to demonstrate the usefulness of the OP technique independent of the generating model for the original phrase translations, we applied Overlapping Phrases to transducers generated by two different alignment methods: HMM Alignment (HMM), and Integrated Segmentation-Alignment (ISA).

The algorithm used to create the HMM Viterbi Alignment was introduced in [10]. Word-level probabilities along with "Jump" probabilities representing the likelihood of deviations from monotonic alignment are trained over several iterations. We use the Viterbi Alignment based on these probabilities to extract phrase-translation transducer rules and refer to a collection of such rules as an HMM transducer.

The ISA approach was introduced in [11] and is based not only on a bilingual word alignment but on monolingual co-occurrence information as well. In this technique, the Point-Wise Mutual Information (MI) between each of the source and target words in the training corpus is calculated. Phrases are identified as contiguous areas of high MI with a maximum length of three words on the source side and three words on the target side. Phrases generated by this method are thus constrained to be short but generally have high accuracy.

3.2. Number of new rules

Table 2 shows the number of new phrase pairs generated during our experiments. These numbers reflect rules which are unique to a single transducer, with no overlap between original and OP rules.

Table 2. New Phrases Generated by OP

	Orig. Rules	Orig. Vocab	New OP Rules
HMM	135,003	140,163	551,375
ISA	46,641	61,103	19,072

Table 3. Distribution of Overlap Lengths

Src/Tgt	1	2	3	4
1	**408,425**	3,635	76	4
2	98,205	**15,364**	819	28
3	11,217	5,575	**3,122**	489
4	1,819	835	1,172	**440**

Table 3 shows the distribution of overlap lengths for the same two transducers. Numbers shown in bold represent merges that overlapped by the same number of tokens on source and target side.

3.3. Usefulness of new rules

Given that we produce rules never seen in training, we would like to verify whether these new phrases are valid translations. We say that a rule is *applicable* if it is useful for the test set: the source side appears somewhere in the test data, and the target side is a valid translation. On the source side, newly generated rules were pruned according to a list of N-grams from the test source, guaranteeing that only applicable rule sources were generated.

To verify the applicability of the target sides we generated, we compared the new rule targets to a list of N-grams occurring in the 4 human reference translations provided for our test set. Table 4 gives the number of applicable target sides in the original HMM transducer and in the new HMM-OP transducer (new rules only). Although the percentage of applicable rules was smaller for the new OP transducer, over one thousand applicable new rules were generated. Table 5 shows that the number of applicable target phrases of length 4-7 went up among the new rules when compared to the original transducer.

Table 4. Applicability of Target Side Phrase Rules

	Unique Targets	Seen in Reference
HMM	135,003	4,639
HMM-OP	464,792	1,844
ISA	20,236	2,828
ISA-OP	14,213	1,319

Table 5. Length of Applicable Target Side Rules

Tgt phrase length	2	3	4	5-7
HMM	2,759	1,354	407	119
HMM-OP	42	1,227	451	124
ISA	951	141	–	–
ISA-OP	822	195	18	–

3.4. Translation Results

Translation results are given in Tables 6 and 7. Two methods for automatic evaluation were used, the NIST score [8] and the Bleu score [7]. Both of these metrics assign values to a translation hypothesis based on the number of matching N-grams between the hypothesis and a set of human reference translations. To put the evaluation scores in perspective, scoring the reference translations in round-robin fashion against each other gave an average Bleu Score of 0.463 and an average NIST score of 9.19.

3.5. OP with HMM Alignment

Our translation experiments are divided into three tests, Overlapping Phrases with the HMM Alignments alone, with the ISA Alignments alone, and finally with a full state-of-the-art system including both HMM and ISA components.

In our overlap experiments with HMM Alignment, we generated overlapping phrase rules from an HMM

Table 6. Isolated Translation Scores

	NIST	Bleu
Baseline HMM	8.26	0.349
Baseline HMM + HMM-OP	8.33	0.354
Baseline ISA	5.97	0.216
Baseline ISA + ISA-OP	6.70	0.239
Human References (avg)	9.20	0.463

transducer and compared the translation performance using these new rules plus the original transducer to using the original rules only. Both configurations used a statistical lexicon as well. In Table 6 the result of adding HMM-OP rules to the baseline HMM system appears in the first two rows.

3.6. OP with ISA Alignment

The lower half of Table 6 shows the improvement after adding ISA-OP rules to a baseline ISA system. The Bleu score increases from 0.216 to 0.239, or 10% over the original ISA phrases. Both configurations used the same statistical lexicon in addition to the ISA phrase translations.

3.7. Full-System Translation

In addition to showing an improvement in isolated component tests, we experimented with the Overlapping Phrases procedure in our fully developed system using the ISA-OP and HMM-OP together.

Table 7 gives the translation results for the Full-system experiments. The Bleu score consistently rises as we add OP rules to the configuration, and both NIST and Bleu scores reflect an overall improvement over the baseline for this test. There is a dip in translation quality according to the NIST score for the Full-system + ISA-OP configuration, but we attribute this to the emphasis on unigram precision in the NIST metric. Overlapping Phrases increase the number of long phrases correctly translated by the SMT engine. A slight fall in single-word precision as a result is balanced out by these longer phrases in the Bleu score but not in the NIST score.

Table 7. Full-system Translation Scores

	NIST	Bleu
Full-system SMT	8.59	0.385
Full + ISA-OP	8.58	0.402
Full + HMM-OP	8.73	0.412
Full + ISA-OP + HMM-OP	8.78	0.425
Human References (avg)	9.20	0.463

3.8. OP rule selection frequency

Analysis of the phrase rules selected by the decoder in the Full-system experiment (Full-system + ISA-OP + HMM-OP; see Table 7) shows that the OP rules play a strong role once they are added to the translation process. In addition, analysis of the decoder log files for this experiment showed that 20% of the transducer rules applied during translation were new rules generated by the OP technique. In addition, the average length of phrases used during translation went up from 1.3 tokens per phrase in the original Full SMT system to 1.4 tokens per phrase in the final +OP system, an increase of 7%.

4. CONCLUSIONS

In conclusion, we found that adding overlapping phrases to the SMT engine had a favorable impact on translation quality. The average phrase length used during translation increased, new phrase pairs that were not seen in the original transducers could be applied during decoding, and translation quality increased according to the Bleu and NIST evaluation metrics.

Remaining issues and future work in this direction include interaction of Overlapping Phrases with other multiword techniques like reordering algorithims. It would also be interesting to test iterative applications of the augmentation process, generating longer and longer phrases into 10- or even 20-word sequences. At some point the rule base may no longer contain any mergeable or applicable candidates. Testing these limits will be a point for future investigations.

5. REFERENCES

[1] Stephan Vogel and Hermann Ney, "Translation with cascaded finite state transducers," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hongkong, China, October 2000, pp. 23–30.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The mathematics of statistical machine translation: Parameter estimat ion," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] Daniel Marcu and William Wong, "A phrase-based, joint probability model for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-02*, Philadelphia, PA, July 2002.

[4] Richard Zens, Franz Josef Och, and Hermann Ney, "Phrase-based statistical machine translation," in *KI - 2002: Advances in Artificial Intelligence 25th Annual German Conference on AI*, Springer Verlag, September 2002, vol. LNAI 2479, pp. 18–32.

[5] Ralf D. Brown, "Example-based machine translation in the Pangloss system," in *16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, August 1996, pp. 169–174.

[6] Rebecca Hutchinson, Paul N. Bennett, Jaime G. Carbonell, Peter Jansen, and Ralf Brown, "Maximal lattice overlap in example-based machine translation," in *MT-Summit 2003*, New Orleans, LA, September 2003.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," Tech. Rep. RC22176(W0109-022), September 17, 2001, IBM, 2001.

[8] George Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of HLT 2002*, San Diego, March 2002, pp. 128–132.

[9] Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venogupal, Bing Zhao, and Alex

Waibel, "The CMU statistical translation system," in *MT-Summit 2003*, New Orleans, LA, September 2003.

[10] Stephan Vogel, Hermann Ney, and Christoph Tillmann, "HMM-based word alignment in statistical translation," in *COLING96*, Copenhagen, August 1996, pp. 836–841.

[11] Ying Zhang, Stephan Vogel, and Alex Waibel, "Integrated phrase segmentation and alignment model for statistical machine translation," in *to appear in Proceedings of the Int'l Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.