Selective Phrase Pair Extraction for Improved Statistical Machine Translation

Luke S. Zettlemoyer
MIT CSAIL
Cambridge, MA 02139
lsz@csail.mit.edu

Robert C. Moore
Microsoft Research
One Microsoft Way
Redmond, WA 98052
bobmoore@microsoft.com

Abstract

Phrase-based statistical machine translation systems depend heavily on the knowledge represented in their phrase translation tables. However, the phrase pairs included in these tables are typically selected using simple heuristics that potentially leave much room for improvement. In this paper, we present a technique for selecting the phrase pairs to include in phrase translation tables based on their estimated quality according to a translation model. This method not only reduces the size of the phrase translation table, but also improves translation quality as measured by the BLEU metric.

1 Introduction

Phrase translation tables are the heart of phrase-based statistical machine translation (SMT) systems. They provide pairs of phrases that are used to construct a large set of potential translations for each input sentence, along with feature values associated with each phrase pair that are used to select the best translation from this set.[1]

The most widely used method for building phrase translation tables (Koehn et al., 2003) selects, from a word alignment of a parallel bilingual training corpus, all pairs of phrases (up to a given length) that are consistent with the alignment. This procedure typically generates many phrase pairs that are not remotely reasonable translation candidates.[2] To avoid creating translations that use these pairs, a set of features is computed for each pair. These features are used to train a translation model, and phrase pairs that produce low scoring translations are avoided. In practice, it is often assumed that current translation models are good enough to avoid building translations with these unreasonable phrase pairs.

In this paper, we question this assumption by investigating methods for pruning low quality phrase pairs. We present a simple procedure that reduces the overall phrase translation table size while increasing translation quality. The basic idea is to initially gather the phrase pairs and train an translation model as usual, but to then select a subset of the overall phrases that performs the best, prune the others, and retrain the translation model. In experiments, this approach reduced the size of the phrase tranlsation table by half, and improved the BLEU score of the resulting translations by up to 1.5 points.

2 Background

As a baseline, we present a relatively standard SMT approach, following Koehn et al. (2003). Potential translations are scored using a linear model where the best translation is computed as

$$\arg\max_{t,a} \sum_{i=1}^{n} \lambda_i f_i(s, a, t)$$

where s is the input sentence, t is the output sentence, and a is a phrasal alignment that specifies how

[1] A "phrase" in this sense can be any contiguous sequence of words, and need not be a complete linguistic constituent.

[2] In one experiment, we managed to generate more than 117,000 English phrases for the the French word "de".

Proceedings of NAACL HLT 2007, Companion Volume, pages 209–212,
Rochester, NY, April 2007.

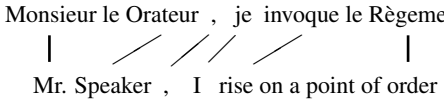

Figure 1: A word aligned sentence pair.

t is constructed from s. The weights λ_i associated with each feature f_i are tuned to maximize the quality of the translations.

The training procedure starts by computing a word alignment for each sentence pair in the training corpus. A word alignment is a relation between the words in two sentences where, intuitively, words are aligned to their translation in the other language. In this work, we use a discriminatively trained word aligner (Moore et al., 2006) that has state of the art performance. Figure 1 presents a high quality alignment produced by this aligner.

Given a word aligned corpus, the second step is to extract a phrase translation table. Each entry in this table contains a source language phrase s, a target language phrase t, and a list of feature values $\phi(s,t)$. It is usual to extract every phrase pair, up to a certain phrase length, that is consistent with the word alignment that is annotated in the corpus. Each consistent pair must have at least one word alignment between words within the phrases and no words in either phrase can be aligned any words outside of the phrases. For example, Figure 2 shows some of the phrase pairs that would be extracted from the word-aligned sentence pair in Figure 1. A full list using phrases of up to three words would include 28 pairs.

For each extracted phrase pair (s,t), feature values $\phi(s,t) = \langle \log p(s|t), \log p(t|s), \log l(s,t) \rangle$ are computed. The first two features, the log translation and inverse translation probabilities, are estimated by counting phrase cooccurrences, following Koehn et al. (2003). The third feature is the logarithm of a lexical score $l(s,t)$ that provides a simple form of smoothing by weighting a phrase pair based on how likely individual words within the phrases are to be translations of each other. We use a version from Foster et al. (2006), modified from (Koehn et al., 2003), which is an average of pairwise word translation probabilities.

In phrase-based SMT, the decoder produces translations by dividing the source sentence into a sequence of phrases, choosing a target language phrase

#	Source Lang. Phrase	Target Lang. Phrase
1	Monsieur	Mr.
2	Monsieur le	Mr.
3	Monsieur le Orateur	Mr. Speaker
4	le Orateur	Speaker
5	Orateur	Speaker
...	...	...
23	le Règlement	point of order
24	le Règlement	of order
25	le Règlement	order
26	Règlement	point of order
27	Règlement	of order
28	Règlement	order

Figure 2: Phrase pairs consistent with the word alignment in Figure 1.

as a translation for each source language phrase, and ordering the target language phrases to build the final translated sentence. Each potential translation is scored according to a weighted linear model. We use the three features from the phrase translation table, summing their values for each phrase pair used in the translation. We also use four additional features: a target language model, a distortion penalty, the target sentence word count, and the phrase pair count, all computed as described in (Koehn, 2004). For all of the experiments in this paper, we used the Pharaoh beam-search decoder (Koehn, 2004) with the features described above.

Finally, to estimate the parameters λ_i of the weighted linear model, we adopt the popular minimum error rate training procedure (Och, 2003) which directly optimizes translation quality as measured by the BLEU metric.

3 Selective Phrase Pair Extraction

In order to improve performance, it is important to select high quality phrase pairs for the phrase translation table. We use two key ideas to guide selection:

- **Preferential Scoring:** Phrase pairs are selected using a function $q(s,t)$ that returns a high score for source, target phrase pairs (s,t) that lead to high quality translations.
- **Redundancy Constraints:** Our intuition is that each occurrence of a source or target language phrase really has at most one translation for that sentence pair. Redundancy constraints minimize the number of possible translations that are extracted for each phrase occurrence.

Selecting phrases that a translation model prefers and eliminating at least some of the ambiguity that comes with extracting multiple translations for a single phrase occurrence creates a smaller phrase translation table with higher quality entries.

The ideal scoring metric would give high scores to phrase pairs that lead to high-quality translations and low scores to those that would decrease translation quality. The best such metric we have available is provided by the overall translation model. Our scoring metric $q(s,t)$ is therefore computed by first extracting a full phrase translation table, then training a full translation model, and finally using a subpart of the model to score individual phrase pairs in isolation. Because the scoring is tied to a model that is optimized to maximize translation quality, more desirable phrase pairs should be given higher scores.

More specifically, $q(s,t) = \phi(s,t) \cdot \lambda$ where $\phi(s,t)$ is the length three vector that contains the feature values stored with the phrase pair (s,t) in the phrase translation table, and λ is a vector of the three parameter values that were learned for these features by the full translation model. The rest of the features are ignored because they are either constant or depend on the target language sentence which is fixed during phrase extraction. In essence, we are using the subpart of a full translation model that looks at phrase pair identity and scoring the pair based on how the full model would like it.

This scoring metric is used in a phrase pair selection algorithm inspired by competitive linking for word alignment (Melamed, 2000). *Local competitive linking* extracts high scoring phrase pairs while enforcing a redundancy constraint that minimizes the number of phrase pairs that share a common phrase. For each sentence pair in the training set, this algorithm marks the highest scoring phrase pair, according to $q(s,t)$, containing each source language phrase and the highest scoring phrase pair containing each target language phrase. Each of these marked phrase pairs is selected and the phrase translation table is rebuilt. This is a soft redundancy constraint because a phrase pair will only be excluded if there is a higher scoring pair that shares its source language phrase and a higher scoring pair that shares its target language phrase. For example, consider again the phrase pairs in Figure 2 and assume they are sorted by their scores. Local competitive linking will select every phrase pair except for 27 and 28. All other pairs are the highest scoring options for at least one of their phrases.

Selective phrase extraction with competitive linking can be seen as a Viterbi reestimation algorithm. Because we are extracting fewer phrase pairs, the features associated with each phrase pair will differ. If the removed phrases were not real translations of each other in the first place, the translation features $p(s|t)$ and $p(t|s)$ should be better estimates because the high quality phrases that remain will be given the probability mass that was assigned to the pruned phrase pairs. Although we are running it in a purely discriminative setting, it has a similar feel to an EM algorithm. First, a full phrase translation table and parameter estimate is computed. Then, based on that estimate, a subset of the phrases is selected which, in turn, supplies a new estimate for the parameters. One question is how many times to run this reestimation procedure. We found, on the development set, that it never helped to run more than one iteration. Perhaps because of the hard nature of the algorithm, repeated iterations caused slight decreases in phrase translation table size and overall performance.

4 Experiments

In this section, we report experiments conducted with Canadian Hansards data from the 2003 HLT-NAACL word-alignment workshop (Mihalcea and Pedersen, 2003). Phrase pairs are extracted from 500,000 word-aligned French-English sentence pairs. Translation quality is evaluated according to the BLEU metric (with one reference translation). Three additional disjoint data sets (from the same source) were used, one with 500 sentence pairs for minimum error rate training, another with 1000 pairs for development testing, and a final set of 2000 sentence pairs for the final test. For each experiment, we trained the full translation model as described in Section 2. Each trial varied only in the phrase translation table that was used.[3]

One important question is what the maximum phrase length should be for extraction. To investigate this issue, we ran experiments on the devel-

[3]These experiments also used the default pruning from the Pharaoh decoder, allowing only the 10 best output phrases to be considered for each input phrase. This simple global pruning cannot be substituted for the competitive linking described here.

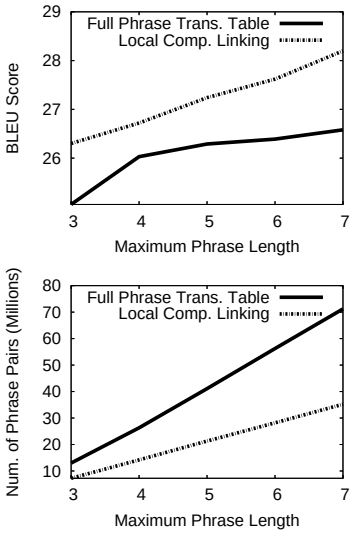

Figure 3: Scaling the maximum phrase length.

opment set. Figure 3 shows a comparison of the full phrase table to local competitive linking as the maximum phrase length is varied. Local competitive linking consistently outperforms the full table and the difference in BLEU score seems to increase with the length. The growth in the size of the phrase translation table seems to be linear with maximum phrase length in both cases, with the table size growing at a slower rate under local competitive linking.

To verify these results, we tested the model trained with the full phrase translation table against the model trained with the table selected by local competitive linking on the heldout test data. Both tables included phrases up to length 7 and the models were tested on a set of 2000 unseen sentence pairs. The results matched the development experiments. The full system scored 26.78 while the local linking achieved 28.30, a difference of 1.52 BLEU points.

5 Discussion

The most closely related work attempts to create higher quality phrase translation tables by learning a generative model that directly incorporates phrase pair selection. The original approach (Marcu and Wong, 2002) was limited due to computational constraints but recent work (DeNero et al., 2006; Birch et al., 2006) has improved the efficiency by using word alignments as constraints on the set of possible phrase pairs. The best results from this line of work allow for a significantly smaller phrase translation table, but never improve translation performance.

In this paper, we presented an algorithm that improves translation quality by selecting a smaller phrase translation table. We hope that this work highlights the need to think carefully about the quality of the phrase translation table, which is the central knowledge source for most modern statistical machine translation systems. The methods used in the experiments are so simple that we believe that there is significant potential for improvement by using better methods for scoring phrase pairs and selecting phrase pairs based those scores.

References

Alexandra Birch, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, join probability statistical translation model. In *Proceedings of the Workshop on Stastical Machine Translation*.

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of the Workshop on Stastical Machine Translation*.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for stastical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Stastical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of The Sixth Conference of the Association for Machine Translation in the Americas*.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

I. Dan Melamed. 2000. Models of translation equivalence amoung words. *Computational Linguistics*, 26(2):221–249.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop, Building and Using Parallel Texts: Data Driven Machine Translation and Beyond.*

Robert C. Moore, Wen-tau Yih, and Andreas Bode. 2006. Improved discriminative bilingual word alignment. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.