Constraining the Phrase-Based, Joint Probability Statistical Translation Model

Alexandra Birch Chris Callison-Burch Miles Osborne Philipp Koehn
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
a.c.birch-mayne@sms.ed.ac.uk

Abstract

The joint probability model proposed by Marcu and Wong (2002) provides a strong probabilistic framework for phrase-based statistical machine translation (SMT). The model's usefulness is, however, limited by the computational complexity of estimating parameters at the phrase level. We present the first model to use word alignments for constraining the space of phrasal alignments searched during Expectation Maximization (EM) training. Constraining the joint model improves performance, showing results that are very close to state-of-the-art phrase-based models. It also allows it to scale up to larger corpora and therefore be more widely applicable.

1 Introduction

Machine translation is a hard problem because of the highly complex, irregular and diverse nature of natural languages. It is impossible to accurately model all the linguistic rules that shape the translation process, and therefore a principled approach uses statistical methods to make optimal decisions given incomplete data.

The original IBM Models (Brown et al., 1993) learn word-to-word alignment probabilities which makes it computationally feasible to estimate model parameters from large amounts of training data. Phrase-based SMT models, such as the alignment template model (Och, 2003), improve on word-based models because phrases provide local context which leads to better lexical choice and more reliable local reordering. However, most phrase-based models extract their phrase pairs from previously word-aligned corpora using ad-hoc heuristics. These models perform no search for optimal phrasal alignments. Even though this is an efficient strategy, it is a departure from the rigorous statistical framework of the IBM Models.

Marcu and Wong (2002) proposed the joint probability model which directly estimates the phrase translation probabilities from the corpus in a theoretically governed way. This model neither relies on potentially sub-optimal word alignments nor on heuristics for phrase extraction. Instead, it searches the phrasal alignment space, simultaneously learning translation lexicons for both words and phrases. The joint model has been shown to outperform standard models on restricted data sets such as the small data track for Chinese-English in the 2004 NIST MT Evaluation (Przybocki, 2004).

However, considering all possible phrases and all their possible alignments vastly increases the computational complexity of the joint model when compared to its word-based counterpart. In this paper, we propose a method of constraining the search space of the joint model to areas where most of the unpromising phrasal alignments are eliminated and yet as many potentially useful alignments as possible are still explored. The joint model is constrained to phrasal alignments which do not contradict a set high confidence word alignments for each sentence. These high confidence alignments could incorporate information from both statistical and linguistic sources. In this paper we use the points of high confidence from the intersection of the bi-directional Viterbi word alignments to constrain the model, increasing performance and decreasing complexity.

Proceedings of the Workshop on Statistical Machine Translation, pages 154–157,
New York City, June 2006.

2 Translation Models

2.1 Standard Phrase-based Model

Most phrase-based translation models (Och, 2003; Koehn et al., 2003; Vogel et al., 2003) rely on a pre-existing set of word-based alignments from which they induce their parameters. In this project we use the model described by Koehn et al. (2003) which extracts its phrase alignments from a corpus that has been word aligned. From now on we refer to this phrase-based translation model as the standard model. The standard model decomposes the foreign input sentence F into a sequence of I phrases $\overline{f}_1, \ldots, \overline{f}_I$. Each foreign phrase $\overline{f_i}$ is translated to an English phrase $\overline{e_i}$ using the probability distribution $\theta(\overline{f}_i|\overline{e}_i)$. English phrases may be reordered using a relative distortion probability.

This model performs no search for optimal phrase pairs. Instead, it extracts phrase pairs $(\overline{f}_i, \overline{e}_i)$ in the following manner. First, it uses the IBM Models to learn the most likely word-level Viterbi alignments for English to Foreign and Foreign to English. It then uses a heuristic to reconcile the two alignments, starting from the points of high confidence in the intersection of the two Viterbi alignments and growing towards the points in the union. Points from the union are selected if they are adjacent to points from the intersection and their words are previously unaligned.

Phrases are then extracted by selecting phrase pairs which are 'consistent' with the symmetrized alignment, which means that all words within the source language phrase are only aligned to the words of the target language phrase and vice versa. Finally the phrase translation probability distribution is estimated using the relative frequencies of the extracted phrase pairs.

This approach to phrase extraction means that phrasal alignments are locked into the symmetrized alignment. This is problematic because the symmetrization process will grow an alignment based on arbitrary decisions about adjacent words and because word alignments inadequately represent the real dependencies between translations.

2.2 Joint Probability Model

The joint model (Marcu and Wong, 2002), does not rely on a pre-existing set of word-level alignments. Like the IBM Models, it uses EM to align and estimate the probabilities for sub-sentential units in a parallel corpus. Unlike the IBM Models, it does not constrain the alignments to being single words.

The joint model creates phrases from words and commonly occurring sequences of words. A concept, c_i, is defined as a pair of aligned phrases $< \overline{e}_i, \overline{f}_i >$. A set of concepts which completely covers the sentence pair is denoted by C. Phrases are restricted to being sequences of words which occur above a certain frequency in the corpus. Commonly occurring phrases are more likely to lead to the creation of useful phrase pairs, and without this restriction the search space would be much larger.

The probability of a sentence and its translation is the sum of all possible alignments $\mathcal{C}$, each of which is defined as the product of the probability of all individual concepts:

$$p(F, E) = \sum_{C \in \mathcal{C}} \prod_{<\overline{e}_i, \overline{f}_i > \in C} p(< \overline{e}_i, \overline{f}_i >) \quad (1)$$

The model is trained by initializing the translation table using Stirling numbers of the second kind to efficiently estimate $p(< \overline{e}_i, \overline{f}_i >)$ by calculating the proportion of alignments which contain $p(< \overline{e}_i, \overline{f}_i >)$ compared to the total number of alignments in the sentence (Marcu and Wong, 2002). EM is then performed by first discovering an initial phrasal alignments using a greedy algorithm similar to the competitive linking algorithm (Melamed, 1997). The highest probability phrase pairs are iteratively selected until all phrases are are linked. Then hill-climbing is performed by searching once for each iteration for all merges, splits, moves and swaps that improve the probability of the initial phrasal alignment. Fractional counts are collected for all alignments visited.

Training the IBM models is computationally challenging, but the joint model is much more demanding. Considering all possible segmentations of phrases and all their possible alignments vastly increases the number of possible alignments that can be formed between two sentences. This number is exponential with relation to the length of the shorter sentence.

3 Constraining the Joint Model

The joint model requires a strategy for restricting the search for phrasal alignments to areas of the alignment space which contain most of the probability mass. We propose a method which examines

phrase pairs that are consistent with a set of high confidence word alignments defined for the sentence. The set of alignments are taken from the intersection of the bi-directional Viterbi alignments.

This strategy for extracting phrase pairs is similar to that of the standard phrase-based model and the definition of 'consistent' is the same. However, the constrained joint model does not lock the search into a heuristically derived symmetrized alignment. Joint model phrases must also occur above a certain frequency in the corpus to be considered.

The constraints on the model are binding during the initialization phase of training. During EM, inconsistent phrase pairs are given a small, nonzero probability and are thus not considered unless unaligned words remain after linking together high probability phrase pairs. All words must be aligned, there is no NULL alignment like in the IBM models.

By using the IBM Models to constrain the joint model, we are searching areas in the phrasal alignment space where both models overlap. We combine the advantage of prior knowledge about likely word alignments with the ability to perform a probabilistic search around them.

4 Experiments

All data and software used was from the NAACL 2006 Statistical Machine Translation workshop unless otherwise indicated.

4.1 Constraints

The unconstrained joint model becomes intractable with very small amounts of training data. On a machine with 2 Gb of memory, we were only able to train 10,000 sentences of the German-English Europarl corpora. Beyond this, pruning is required to keep the model in memory during EM. Table 1 shows that the application of the word constraints considerably reduces the size of the space of phrasal alignments that is searched. It also improves the BLEU score of the model, by guiding it to explore the more promising areas of the search space.

4.2 Scalability

Even though the constrained joint model reduces complexity, pruning is still needed in order to scale up to larger corpora. After the initialization phase of the training, all phrase pairs with counts less than 10 million times that of the phrase pair with the highest count, are pruned from the phrase table. The model is also parallelized in order to speed up training.

	Unconstrained	Constrained
No. Concepts	6,178k	1,457k
BLEU	19.93	22.13
Time(min)	299	169

Table 1. The impact of constraining the joint model trained on 10,000 sentences of the German-English Europarl corpora and tested with the Europarl test set used in Koehn et al. (2003)

The translation models are included within a log-linear model (Och and Ney, 2002) which allows a weighted combination of features functions. For the comparison of the basic systems in Table 2 only three features were used for both the joint and the standard model: $p(e|f)$, $p(f|e)$ and the language model, and they were given equal weights.

The results in Table 2 show that the joint model is capable of training on large data sets, with a reasonable performance compared to the standard model. However, here it seems that the standard model has a slight advantage. This is almost certainly related to the fact that the joint model results in a much smaller phrase table. Pruning eliminates many phrase pairs, but further investigations indicate that this has little impact on BLEU scores.

	BLEU	Size
Joint Model	25.49	2.28
Standard Model	26.15	19.04

Table 2. Basic system comparisons: BLEU scores and model size in millions of phrase pairs for Spanish-English

The results in Table 3 compare the joint and the standard model with more features. Apart from including all Pharaoh's default features, we use two new features for both the standard and joint models: a 5-gram language model and a lexicalized reordering model as described in Koehn et al. (2005). The weights of the feature functions, or model components, are set by minimum error rate training provided by David Chiang from the University of Maryland.

On smaller data sets (Koehn et al., 2003) the joint model shows performance comparable to the standard model, however the joint model does not reach the level of performance of the stan-

	EN-ES	ES-EN
Joint		
3-gram, dl4	20.51	26.64
5-gram, dl6	26.34	27.17
+ lex. reordering	26.82	27.80
Standard Model		
5-gram, dl6		
+ lex. reordering	31.18	31.86

Table 3. Bleu scores for the joint model and the standard model showing the effect of the 5-gram language model, distortion length of 6 (dl) and the addition of lexical reordering for the English-Spanish and Spanish-English tasks.

dard model for this larger data set. This could be due to the fact that the joint model results in a much smaller phrase table. During EM only phrase pairs that occur in an alignment visited during hill-climbing are retained. Only a very small proportion of the alignment space can be searched and this reduces the chances of finding optimum parameters. The small number of alignments visited would lead to data sparseness and over-fitting. Another factor could be efficiency trade-offs like the fast but not optimal competitive linking search for phrasal alignments.

4.3 German-English submission

We also submitted a German-English system using the standard approach to phrase extraction. The purpose of this submission was to validate the syntactic reordering method that we previously proposed (Collins et al., 2005). We parse the German training and test corpus and reorder it according to a set of manually devised rules. Then, we use our phrase-based system with standard phrase-extraction, lexicalized reordering, lexical scoring, 5-gram LM, and the Pharaoh decoder.

On the development test set, the syntactic reordering improved performance from 26.86 to 27.70. The best submission in last year's shared task achieved a score of 24.77 on this set.

5 Conclusion

We presented the first attempt at creating a systematic framework which uses word alignment constraints to guide phrase-based EM training. This shows competitive results, to within 0.66 BLEU points for the basic systems, suggesting that a rigorous probabilistic framework is preferable to heuristics for extracting phrase pairs and their probabilities.

By introducing constraints to the alignment space we can reduce the complexity of the joint model and increase its performance, allowing it to train on larger corpora and making the model more widely applicable.

For the future, the joint model would benefit from lexical weighting like that used in the standard model (Koehn et al., 2003). Using IBM Model 1 to extract a lexical alignment weight for each phrase pair would decrease the impact of data sparseness, and other kinds smoothing techniques will be investigated. Better search algorithms for Viterbi phrasal alignments during EM would increase the number and quality of model parameters.

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

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

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 127–133.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, and Chris Callison-Burch. 2005. Edinburgh system description. In *IWSLT Speech Translation Evaluation*.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP*.

Dan Melamed. 1997. A word-to-word model of translational equivalence. In *Proceedings of ACL*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*.

Franz Josef Och. 2003. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen Department of Computer Science, Aachen, Germany.

Mark Przybocki. 2004. NIST 2004 machine translation evaluation results. Confidential e-mail to workshop participants, May.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Machine Translation Summit*.