Leave-One-Out Phrase Model Training for Large-Scale Deployment

Joern Wuebker
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Germany
wuebker@cs.rwth-aachen.de

Mei-Yuh Hwang, Chris Quirk
Microsoft Corporation
Redmond, WA, USA
{mehwang,chrisq}@microsoft.com

Abstract

Training the phrase table by force-aligning (FA) the training data with the reference translation has been shown to improve the phrasal translation quality while significantly reducing the phrase table size on medium sized tasks. We apply this procedure to several large-scale tasks, with the primary goal of reducing model sizes without sacrificing translation quality. To deal with the noise in the automatically crawled parallel training data, we introduce on-demand word deletions, insertions, and backoffs to achieve over 99% successful alignment rate. We also add heuristics to avoid any increase in OOV rates. We are able to reduce already heavily pruned baseline phrase tables by more than 50% with little to no degradation in quality and occasionally slight improvement, without any increase in OOVs. We further introduce two global scaling factors for re-estimation of the phrase table via posterior phrase alignment probabilities and a modified absolute discounting method that can be applied to fractional counts.

Index Terms: phrasal machine translation, phrase training, phrase table pruning

1 Introduction

Extracting phrases from large amounts of noisy word-aligned training data for statistical machine translation (SMT) generally has the disadvantage of producing many unnecessary phrases (Johnson et al., 2007). These can include poor quality phrases, composite phrases that are concatenations of shorter ones, or phrases that are assigned very low probabilities, so that they have no realistic chance when competing against higher scoring phrase pairs. The goal of this work is two-fold: (i) investigating forced alignment training as a phrase table pruning method for large-scale commercial SMT systems and (ii) proposing several extensions to the training procedure to deal with practical issues and stimulate further research.

Generative phrase translation models have the inherent problem of over-fitting to the training data (Koehn et al., 2003; DeNero et al., 2006). (Wuebker et al., 2010) introduce a leave-one-out procedure which is shown to counteract over-fitting effects. The authors report significant improvements on the German-English Europarl data with the additional benefit of a severely reduced phrase table size. This paper investigates its impact on a number of commercial large-scale systems and presents several extensions.

The first extension is to deal with the highly noisy training data, which is automatically crawled and sentence aligned. The noise and the baseline pruning of the phrase table lead to low success rates when aligning the source sentence with the target sentence. We introduce on-demand word deletions, insertions, and backoff phrases to increase the success rate so that we can cover essentially the entire training data. Secondly, phrase table pruning makes out-of-vocabulary (OOV) issues even more pronounced. To avoid an increased OOV rate, we retrieve single-word translations from the baseline phrase table. Lastly, we propose two global scaling

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 460–467,
Montréal, Canada, June 7-8, 2012.

factors to allow fine-tuning of the phrase counts in an attempt to re-estimate the translation probabilities and a modification of absolute discounting that can be applied to fractional counts.

Our main contribution is applying forced-alignment on the training data to prune the phrase table. The rationale behind this is that by decoding the training data, we can identify the phrases that are actually used by the decoder. Further, we present preliminary experiments on re-estimating the channel models in the phrase table based on counts extracted from the force-aligned data.

This work is organized as follows. We discuss related work in Section 2, describe our decoder and training procedure in Section 3 and the experiments in Section 4. A conclusion and discussion of future work is given in Section 5.

2 Related Work

Force-aligning bilingual data has been explored as a means of model training in previous work. Liang et al. (2006) use it for their *bold updating* strategy to update discriminative feature weights. Utilizing force-aligned data to train a unigram phrase segmentation model is proposed by Shen et al. (2008). Wuebker et al. (2010) apply forced alignment to train the phrase table in an EM-like fashion. They report a significant reduction in phrase table size.

In this work we apply forced alignment training as a pure phrase table pruning technique. Johnson et al. (2007) successfully investigate a number of pruning methods for the phrase inventory based on significance testing. While their approach is more straightforward and less elaborate, we argue that our method is directly tailored to the decoding process and works on top of an already heavily pruned baseline phrase table.

We further experiment with applying the (scaled) phrase alignment posteriors to train the phrase table. A similar idea has been addressed in previous work, e.g. (Venugopal et al., 2003; de Gispert et al., 2010), where word alignment posterior probabilities are leveraged for grammar extraction.

Finally, a number of papers describe extending real phrase training to the hierarchical machine translation paradigm (Blunsom et al., 2008; Cmejrek et al., 2009; Mylonakis and Sima'an, 2010).

3 Phrase Training

3.1 Decoder

Our translation decoder is similar to the open-source toolkit Moses (Koehn et al., 2007). It models translation as a log-linear combination of two phrasal and two lexical channel models, an n-gram language model (LM), phrase, word and distortion penalties and a lexicalized reordering model. The decoding can be summarized as finding the best scoring target sentence T^* given a source sentence S:

$$T^* = \underset{T}{\operatorname{argmax}} \sum_i \lambda_i \log\ g_i(S,T) \qquad (1)$$

where each g_i represents one feature (the channel models, n-gram, phrase count, etc.). The model weights λ_i are usually discriminatively learned on a development data set via minimum error rate training (MERT) (Och, 2003).

Constraining the decoder to a fixed target sentence is straightforward. Each partial hypothesis is compared to the reference and discarded if it does not match. The language model feature can be dropped since all hypotheses lead to the same target sentence. The training data is divided into subsets for parallel alignment. A bilingual phrase matching is applied to the phrase table to extract only the subset of entries that are pertinent to each subset of training data, for memory efficiency. For forced alignment training, we set the distortion limit Δ to be larger than in regular translation decoding. As unlimited distortion leads to very long training times, we compromise on the following heuristic. The distortion limit is set to be the maximum of 10, twice that of the baseline setting, and 1.5 times the maximum phrase length:

$$\begin{aligned} \Delta \;=\; & \max\{10, \\ & 2 * (\text{baseline distortion}), \\ & 1.5 * (\text{max phrase length})\} \end{aligned} \qquad (2)$$

To avoid over-fitting, we employ the same leave-one-out procedure as (Wuebker et al., 2010) for training. Here, it is applied on top of the Good-Turing (GT) smoothed phrase table (Foster et al.,

2006). Our phrase table stores the channel probabilites and marginal counts for each phrase pair, but not the discounts applied. Therefore, for each sentence, if the phrase pair (s,t) has a joint count $c(s,t)$ computed from the entire training data, and occurs $c_1(s,t)$ times in the current sentence, the leave-one-out probability $p'(t|s)$ for the current sentence will be:

$$\begin{aligned} p'(t|s) &= \frac{c'(s,t)-d}{c'(s)} \\ &= \frac{c(s,t)-c_1(s,t)-d}{c(s)-c_1(s)} \\ &= \frac{p(t|s)c(s)-c_1(s,t)}{c(s)-c_1(s)} \end{aligned} \quad (3)$$

since $p(t|s)c(s) = c(s,t)-d$, where d is the GT discount value. In the case where $c(s,t) = c_1(s,t)$ (i.e. (s,t) occurs exclusively in one sentence pair), we use a very low probability as the floor value. We apply leave-one-out discounting to the forward and backward translation models only, not to the lexical channel models.

Our baseline phrase extraction applies some heuristic-based pruning strategies. For example, it prunes offensive translations and many-words to many-words singletons (i.e. a joint count of 1 and both source phrase and target phrase contain multiple words)*. Finally the forward and backward translation probabilities are smoothed with Good-Turing discounting.

3.2 Weak Lambda Training with High Distortion

Our leave-one-out training flowchart can be illustrated in Figure 1. To force-align the training data with good quality, we need a set of trained lambda weights, as shown in Equation 1. We can use the lambda weights learned from the baseline system for that purpose. However, ideally we want the lambda values to be learned under a similar configuration as the forced alignment. Therefore, for this purpose we run MERT with the larger distortion limit given in Equation 2.

*The pruned entries are nevertheless used in computing joint counts and marginal counts.

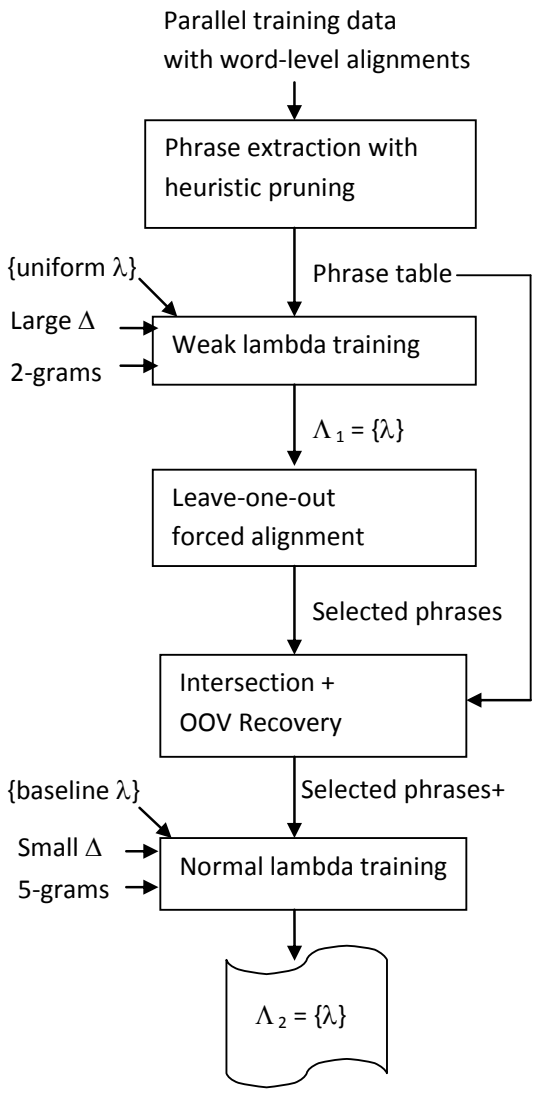

Figure 1: Flowchart of forced-alignment phrase training.

Additionally, since forced alignment does not use the language model, we propose to use a weaker language model for training the lambdas (Λ_1) to be used in the forced alignment decoding.

Using a weaker language model also speeds up the lambda training process, especially when we are using a distortion limit Δ at least twice as high as in the baseline system. In our experiments, the baseline system uses an English 5-gram language model trained on a large amount of monolingual data. The lambda values used for forced alignment are learned using the bigram LM trained on the target side of the

parallel corpus for each system.

We compared a number of systems using different degrees of weak models and found out the impact on the final system was minimal. However, using a small bigram LM with large distortion yielded a stable performance in terms of BLEU, and was 25% faster than using a large 5-gram with the baseline distortion. Because of the speed improvement and its stability, this paper adopts the weak bigram lambda training.

3.3 On-demand Word Insertions and Deletions

For many training sentences the translation decoder is not able to find a phrasal alignment. We identified the following main reasons for failed alignments:

- Incorrect sentence alignment or sentence segmentation by the data crawler,
- OOVs due to initial pruning in the phrase extraction phase,
- Faulty word alignments,
- Strongly reordered sentence structure. That is, the distortion limit during forced alignment is too restrictive.

For some of these cases, discarding the sentence pairs can be seen as implicit data cleaning. For others, there do exist valid sub-sentences that are aligned properly. We would like to be able to leverage those sub-sentences, effectively allowing us to do partial sentence removal. Therefore, we introduce on-demand word insertions and deletions. Whenever a partial hypothesis can not be expanded to the next target word t_j, with the given phrase table, we allow the decoder to artificially introduce a phrase pair $(null, t_j)$ to insert the target word into the hypothesis without consuming any source word. These artificial phrase pairs are introduced with a high penalty and are ignored when creating the output phrase table. We can also introduce *backoff* phrase pairs (s_i, t_j) for all source words s_i that are not covered so far, also with a fixed penalty.

After we reach the end of the target sentence, if there are any uncovered source words s_i, we artificially add the deletion phrase pairs $(s_i, null)$ with a high penalty. Introducing on-demand word insertions and deletions increases the data coverage to at least 99% of the training sentences on all tasks we have worked on. Due to the success of insertion/deletion phrases, we have not conducted experiments using backoff phrases within the scope of this work, but leave this to future work.

3.4 Phrase Training as Pruning

This work concentrates on practical issues with large and noisy training data. Our main goal is to apply phrase training to reduce phrase table size without sacrificing quality. We do this by dumping n-best alignments of the training data, where n ranges from 100-200. We prune the baseline phrase table to only contain phrases that appear in any of the n-best phrase alignments, leaving the channel probabilities unchanged. That is, the model scores are still estimated from the original counts. We can control the size of the final phrase table by adjusting the size of the n-best list. Based on the amount of memory we can afford, we can thus keep the most important entries in the phrase table.

3.5 OOV retrieval

When performing phrase table pruning as described in Section 3.4, OOV rates tend to increase. This effect is even more pronounced when deletion/insertion phrases are not used, due to the low alignment success rate. For commercial applications, untranslated words are a major concern for end users, although it rarely has any impact on BLEU scores. Therefore, for the final phrase table after forced alignment training, we check the translations for single words in the baseline phrase table. If any single word has no translation in the new table, we recover the top x translations from the baseline table. In practice, we set $x = 3$.

3.6 Fractional Counts and Model Re-estimation

As mentioned in Section 3.4, for each training sentence pair we produce the n-best phrasal alignments. If we interpret the model score of an alignment as its log likelihood, we can weight the count for each phrase by its posterior probability. However, as the

log-linear model weights are trained in a discriminative fashion, they do not directly correspond to probabilities. In order to leverage the model scores, we introduce two scaling factors ϑ and ρ that allow us to shape the count distribution according to our needs. For one sentence pair, the count for the phrase pair (s,t) is defined as

$$c(s,t) = \left(\sum_{i=1}^{n} c(s,t|h_i) \cdot \frac{exp(\vartheta \cdot \phi(h_i))}{\sum_{j=1}^{n} exp(\vartheta \cdot \phi(h_j))} \right)^{\rho} , \quad (4)$$

where h_i is the i-th hypothesis of the n-best list, $\phi(h_i)$ the log-linear model score of the alignment hypothesis h_i and $c(s,t|h_i)$ the count of (s,t) within h_i. If $\vartheta = 0$, all alignments within the n-best list are weighted equally. Setting $\rho = 0$ means that all phrases that are used anywhere in the n-best list receive a count of 1.

Absolute discounting is a popular smoothing method for relative frequencies (Foster et al., 2006). Its application, however, is somewhat difficult, if counts are not required to be integer numbers and can in fact reach arbitrarily small values. We propose a minor modification, where the discount parameter d is added to the denominator, rather than subtracting it from the numerator. The discounted relative frequency for a phrase pair (s,t) is computed as

$$p(s|t) = \frac{c(s,t)}{d + \sum_{s'} c(s',t)} \quad (5)$$

3.7 Round-Two Lambda Training

After the phrase table is pruned with forced alignment (either re-estimating the channel probabilities or not), we recommend a few more iterations of lambda training to ensure our lambda values are robust with respect to the new phrase table. In our experiments, we start from the baseline lambdas and train at most 5 more iterations using the baseline distortion and the 5-gram English language model. The settings have to be consistent with the final decoding; therefore we are not using weak lambda training here.

system	parallel corpus (sent. pairs)	Dev	Test1	WMT
it-en	13.0M	2000	5000	3027
pt-en	16.9M	2448	5000	1000
nl-en	15.0M	499	4996	1000
et-en	3.5M	1317	1500	995

Table 1: Data sizes of the four systems Italian, Portuguese, Dutch and Estonian to English. All numbers refer to sentence pairs.

Empirically we found the final lambdas (Λ_2) made a very small improvement over the baseline lambdas. However, we decided to keep this second round of lambda training to guarantee its stability across all language pairs.

4 Experiments

In this section, we describe our experiments on large-scale training data. First, we prune the original phrase table without re-estimation of the models. We conducted experiments on many language pairs. But due to the limited space here, we chose to present two high traffic systems and the two worst systems so that readers can set the correct expectation with the worst-case scenario. The four systems are: Italian (it), Portuguese (pt), Dutch (nl) and Estonian (et), all translating to English (en).

4.1 Corpora

The amount of data for the four systems is shown in Table 1. There are two test sets: Test1 and WMT. Test1 is our internal data set, containing web page translations among others. WMT is sampled from the English side of the benchmark test sets of the *Workshop on Statistical Machine Translation*[†]. The sampled English sentences are then manually translated into other languages, as the input to test X-to-English translation. WMT tends to contain news-like and longer sentences. The development set (for learning lambdas) is from our internal data set. We make sure that there is no overlap among the development set, test sets, and the training set.

[†] www.statmt.org/wmt09

	baseline	FA w/ del.	FA w/o del.
it-en			
suc.rate	–	99.5%	61.2%
Test1	42.27	42.05	42.31
WMT	30.16	30.19	30.19
pt-en			
suc.rate	–	99.5%	66.9%
Test1	47.55	47.47	47.24
WMT	40.74	41.36	41.01
nl-en			
suc.rate	–	99.6%	79.9%
Test1	32.39	31.87	31.18
WMT	43.37	43.06	43.38
et-en			
suc.rate	–	99.1%	73.1%
Test1	46.14	46.35	45.77
WMT	20.08	19.60	19.83

Table 2: BLEU scores of forced-alignment-based phrase-table pruning using weak lambda training. n-best size is 100 except for nl-en, where it is 160. We contrast forced alignment with and without on-demand insertion/deletion phrases. With the on-demand artificial phrases, FA success rate is over 99%.

4.2 Insertion/Deletion Phrases

Unless explicitly stated, all experiments here used the weak bigram LMs to obtain the lambdas used for forced alignment, and on-demand insertion/deletion phrases are applied. For the size of n-best, we use $n = 100$. The only exception is the nl-en language pair, for which we set $n = 160$ because its phrase distortion setting is higher than the others and for its higher number of morphological variations. Table 2 shows the BLEU performance of the four systems, in the baseline setting and in the forced-alignment setting with insertion/deletion phrases and without insertion/deletion phrases. Whether partial sentences should be kept or not (via insertion/deletion phrases) depends on the quality of the training data. One would have to run both settings to decide which is better for each system. In all cases, there is little or no degradation in quality after the table is sufficiently pruned.

Table 3 shows that our main goal of reducing the phrase table size is achieved. On all four language pairs, we are able to prune over 50% of the phrase table. Without on-demand insertions/deletions, the size reduction is even stronger. Notice the size reduction here is relative to the already heavily pruned baseline phrase table.

	PT size reduction	
	w/o del.	w/ del.
it-en	65.4%	54.0%
pt-en	68.5%	61.3%
nl-en	64.1%	56.9%
et-en	63.6%	58.5%

Table 3: % Phrase table size reduction compared with the baseline phrase table

With such a successful size cut, we expected a significant increase in decoding speed in the final system. In practice we experienced 3% to 12% of speedup across all the systems we tested. Both our baseline and the reduced systems use a tight beam width of 20 hypotheses per stack. We assume that with a wider beam, the speed improvement would be more pronounced.

We also did human evaluation on all 8 system outputs (four language pairs, with two test sets per language pair) and all came back positive (more improvements than regressions), even on those that had minor BLEU degradation. We conclude that the size cut in the phrase table is indeed harmless, and therefore we declare our initial goal of phrase table pruning without sacrificing quality is achieved.

In (Wuebker et al., 2010) it was observed, that phrase training reduces the average phrase length. The longer phrases, which are unlikely to generalize, are dropped. We can confirm this obersvation for the it-en and pt-en language pairs in Table 4. However, for nl-en and et-en the average source phrase length is not significantly affected by phrase training, especially with the insertion/deletion phrases. When these artificial phrases are added during forced alignment, they tend to encourage long target phrases as uncovered single target words can be consumed by the insertion phrases. However, these insertion phrases are not dumped into the final phrase table and hence cannot help in reducing the average phrase length of the final phrase table.

	avg. src phrase length		
	baseline	w/o del.	w/ del.
it-en	3.1	2.4	2.4
pt-en	3.7	3.0	3.0
nl-en	3.1	3.0	3.0
et-en	2.9	2.8	3.0

Table 4: Comparison of average source phrase length in the phrase table.

nl-en	Test1	WMT	PT size reduction
baseline	32.29	43.37	–
n=100	31.45	42.90	66.0%
n=160	31.87	43.06	64.1%

et-en	Test1	WMT	PT size reduction
baseline	46.14	20.08	–
n=100	46.35	19.60	63.6%
n=200	46.34	19.88	58.4%

Table 5: BLEU scores of different n-best sizes for the highly inflected Dutch system and the noisy Estonian system.

Table 5 illustrates how the n-best size affects BLEU scores and model sizes for the nl-en and et-en systems.

4.3 Phrase Model Re-estimation

This section conducts a preliminary evaluation of the techniques introduced in Section 3.6. For fast turnaround, these experiments were conducted on approximately $1/3$ of the Italian-English training data. Training is performed with and without insertion/deletion phrases and both with (*FaTrain*) and without (*FaPrune*) re-training of the forward and backward phrase translation probabilities. Table 6 shows the BLEU scores with different settings of the global scaling factor ρ and the inverse discount d. The second global scaling factor is fixed to $\vartheta = 0$. The preliminary results seem to be invariant of the settings. We conclude that using forced alignment posteriors as a feature training method seems to be less effective than using competing hypotheses from free decoding as in (He and Deng, 2012).

				BLEU	
	ins/del	ρ	d	Test1	WMT
baseline	-	-	-	40.6	28.9
FaPrune	no	-	-	40.7	29.1
FaTrain	no	0	0	40.4	28.9
		0.5	0	40.2	28.9
FaPrune	yes	-	-	40.6	28.9
FaTrain	yes	0	0	40.1	28.6
		0.5	0	40.5	29.1
		0.5	0.2	40.5	29.0
		0.5	0.4	40.5	29.0

Table 6: Phrase pruning (*FaPrune*) vs. further model re-estimation after pruning (*FaTrain*) on 1/3 it-en training data, both with and without on-demand insertions/deletions.

5 Conclusion and Outlook

We applied forced alignment on parallel training data with leave-one-out on four large-scale commercial systems. In this way, we were able to reduce the size of our already heavily pruned phrase tables by at least 54%, with almost no loss in translation quality, and with a small improvement in speed performance. We show that for language pairs with strong reordering, the n-best list size needs to be increased to account for the larger search space.

We introduced several extensions to the training procedure. On-demand word insertions and deletions can increase the data coverage to nearly 100%. We plan to extend our work to use backoff translations (the target word that can not be extended given the input phrase table will be aligned to any uncovered single source word) to provide more alignment varieties, and hence hopefully to be able to keep more good phrase pairs. To avoid higher OOV rates after pruning, we retrieved single-word translations from the baseline phrase table.

We would like to emphasize that this leave-one-out pruning technique is not restricted to phrasal translators, even though all experiments presented in this paper are on phrasal translators. It is possible to extend the principle of forced alignment guided pruning to hierarchical decoders, treelet decoders, or syntax-based decoders, to prune redundant or useless phrase mappings or translation rules.

Re-estimating phrase translation probabilities using forced alignment posterior scores did not yield any noticable BLEU improvement so far. Instead, we propose to apply discriminative training similar to (He and Deng, 2012) after forced-alignment-based pruning as future work.

References

[Blunsom et al.2008] Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proceedings of the 46th Annual Conference of the Association for Computational Linguistics: Human Language Technologies (ACL-08:HLT)*, pages 200–208, Columbus, Ohio, June. Association for Computational Linguistics.

[Cmejrek et al.2009] Martin Cmejrek, Bowen Zhou, and Bing Xiang. 2009. Enriching SCFG Rules Directly From Efficient Bilingual Chart Parsing. In *Proc. of the International Workshop on Spoken Language Translation*, pages 136–143, Tokyo, Japan.

[de Gispert et al.2010] Adriá de Gispert, Juan Pino, and William Byrne. 2010. Hierarchical Phrase-based Translation Grammars Extracted from Alignment Posterior Probabilities. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 545–554, MIT, Massachusetts, U.S.A., October.

[DeNero et al.2006] John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why Generative Phrase Models Underperform Surface Heuristics. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 31–38, New York City, June.

[Foster et al.2006] George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 53–61, Sydney, Australia, July.

[He and Deng2012] Xiaodong He and Li Deng. 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, page to appear, Jeju, Republic of Korea, Jul.

[Johnson et al.2007] J Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975, Prague, June.

[Koehn et al.2003] P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta.

[Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantine, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic, June.

[Liang et al.2006] Percy Liang, Alexandre Buchard-Côté, Dan Klein, and Ben Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 761–768, Sydney, Australia.

[Mylonakis and Sima'an2010] Markos Mylonakis and Khalil Sima'an. 2010. Learning Probabilistic Synchronous CFGs for Phrase-based Translation. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 117–, Uppsala,Sweden, July.

[Och2003] Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

[Shen et al.2008] Wade Shen, Brian Delaney, Tim Anderson, and Ray Slyh. 2008. The MIT-LL/AFRL IWSLT-2008 MT System. In *Proceedings of IWSLT 2008*, pages 69–76, Hawaii, U.S.A., October.

[Venugopal et al.2003] Ashish Venugopal, Stephan Vogel, and Alex Waibel. 2003. Effective Phrase Translation Extraction from Alignment Models. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 319–326, Sapporo, Japan, July.

[Wuebker et al.2010] Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden, July.