A Systematic Comparison of Phrase Table Pruning Techniques

Richard Zens and Daisy Stanton and Peng Xu

Google Inc.
{zens,daisy,xp}@google.com

Abstract

When trained on very large parallel corpora, the phrase table component of a machine translation system grows to consume vast computational resources. In this paper, we introduce a novel pruning criterion that places phrase table pruning on a sound theoretical foundation. Systematic experiments on four language pairs under various data conditions show that our principled approach is superior to existing ad hoc pruning methods.

1 Introduction

Over the last years, statistical machine translation has become the dominant approach to machine translation. This is not only due to improved modeling, but also due to a significant increase in the availability of monolingual and bilingual data. Here are just two examples of very large data resources that are publicly available:

- The Google Web 1T 5-gram corpus available from the Linguistic Data Consortium consisting of the 5-gram counts of about one trillion words of web data.[1]
- The 10^9-French-English bilingual corpus with about one billion tokens from the Workshop on Statistical Machine Translation (WMT).[2]

These enormous data sets yield translation models that are expensive to store and process. Even with modern computers, these large models lead to a long experiment cycle that hinders progress. The situation is even more severe if computational resources are limited, for instance when translating on handheld devices. Then, reducing the model size is of the utmost importance.

The most resource-intensive components of a statistical machine translation system are the language model and the phrase table. Recently, compact representations of the language model have attracted the attention of the research community, for instance in Talbot and Osborne (2007), Brants et al. (2007), Pauls and Klein (2011) or Heafield (2011), to name a few. In this paper, we address the other problem of any statistical machine translation system: large phrase tables.

Johnson et al. (2007) has shown that large portions of the phrase table can be removed without loss in translation quality. This motivated us to perform a systematic comparison of different pruning methods. However, we found that many existing methods employ ad-hoc heuristics without theoretical foundation.

The pruning criterion introduced in this work is inspired by the very successful and still state-of-the-art language model pruning criterion based on entropy measures (Stolcke, 1998). We motivate its derivation by stating the desiderata for a good phrase table pruning criterion:

- **Soundness**: The criterion should optimize some well-understood information-theoretic measure of translation model quality.

[1] LDC catalog No. LDC2006T13

[2] http://www.statmt.org/wmt11/translation-task.html

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 972–983, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

- **Efficiency**: Pruning should be fast, i. e., run linearly in the size of the phrase table.
- **Self-containedness**: As a practical consideration, we want to prune phrases from an existing phrase table. This means pruning should use only information contained in the model itself.
- **Good empirical behavior**: We would like to be able to prune large parts of the phrase table without significant loss in translation quality.

Analyzing existing pruning techniques based on these objectives, we found that they are commonly deficient in at least one of them. We thus designed a novel pruning criterion that not only meets these objectives, it also performs very well in empirical evaluations.

The **novel contributions** of this paper are:

1. a systematic description of existing phrase table pruning methods.
2. a new, theoretically sound phrase table pruning criterion.
3. an experimental comparison of several pruning methods for several language pairs.

2 Related Work

The most basic pruning methods rely on probability and count cutoffs. We will cover the techniques that are implemented in the Moses toolkit (Koehn et al., 2007) and the Pharaoh decoder (Koehn, 2004) in Section 3. We are not aware of any work that analyzes their efficacy in a systematic way. It is thus not surprising that some of them perform poorly, as our experimental results will show.

The work of Johnson et al. (2007) is promising as it shows that large parts of the phrase table can be removed without affecting translation quality. Their pruning criterion relies on statistical significance tests. However, it is unclear how this significance-based pruning criterion is related to translation model quality. Furthermore, a comparison to other methods is missing. Here we close this gap and perform a systematic comparison. The same idea of significance-based pruning was exploited in (Yang and Zheng, 2009; Tomeh et al., 2009) for hierarchical statistical machine translation.

A different approach to phrase table pruning was undertaken by Eck et al. (2007a; 2007b). They rely on usage statistics from translating sample data, so it is not self-contained. However, it could be combined with the methods proposed here.

Another approach to phrase table pruning is triangulation (Chen et al., 2008; Chen et al., 2009). This requires additional bilingual corpora, namely from the source language as well as from the target language to a third bridge language. In many situations this does not exist or would be costly to generate.

Duan et al. (2011), Sanchis-Trilles et al. (2011) and Tomeh et al. (2011) modify the phrase extraction methods in order to reduce the phrase table size. The work in this paper is independent of the way the phrase extraction is done, so those approaches are complementary to our work.

3 Pruning Using Simple Statistics

In this section, we will review existing pruning methods based on simple phrase table statistics. There are two common classes of these methods: absolute phrase table pruning and relative phrase table pruning.

3.1 Absolute pruning

Absolute pruning methods rely only on the statistics of a single phrase pair $(\tilde{f}, \tilde{e})$. Hence, they are independent of other phrases in the phrase table. As opposed to relative pruning methods (Section 3.2), they may prune all translations of a source phrase. Their application is easy and efficient.

- **Count-based pruning.** This method prunes a phrase pair $(\tilde{f}, \tilde{e})$ if its observation count $N(\tilde{f}, \tilde{e})$ is below a threshold τ_c:

$$N(\tilde{f}, \tilde{e}) < \tau_c \tag{1}$$

- **Probability-based pruning.** This method prunes a phrase pair $(\tilde{f}, \tilde{e})$ if its probability is below a threshold τ_p:

$$p(\tilde{e}|\tilde{f}) < \tau_p \tag{2}$$

Here the probability $p(\tilde{e}|\tilde{f})$ is estimated via relative frequencies.

3.2 Relative pruning

A potential problem with the absolute pruning methods is that it can prune all occurrences of a source phrase $\tilde{f}$.[3] Relative pruning methods avoid this by considering the full set of target phrases for a specific source phrase $\tilde{f}$.

- **Threshold pruning.** This method discards those phrases that are far worse than the best target phrase for a given source phrase $\tilde{f}$. Given a pruning threshold τ_t, a phrase pair $(\tilde{f}, \tilde{e})$ is discarded if:

$$p(\tilde{e}|\tilde{f}) < \tau_t \cdot \max_{\tilde{e}} \left\{ p(\tilde{e}|\tilde{f}) \right\} \qquad (3)$$

- **Histogram pruning.** An alternative to threshold pruning is histogram pruning. For each source phrase $\tilde{f}$, this method preserves the K target phrases with highest probability $p(\tilde{e}|\tilde{f})$ or, equivalently, their count $N(\tilde{f}, \tilde{e})$.

Note that, except for count-based pruning, none of the methods take the frequency of the source phrase into account. As we will confirm in the empirical evaluation, this will likely cause drops in translation quality, since frequent source phrases are more useful than the infrequent ones.

4 Significance Pruning

In this section, we briefly review significance pruning following Johnson et al. (2007). The idea of significance pruning is to test whether a source phrase $\tilde{f}$ and a target phrase $\tilde{e}$ co-occur more frequently in a bilingual corpus than they should just by chance. Using some simple statistics derived from the bilingual corpus, namely

- $N(\tilde{f})$ the count of the source phrase $\tilde{f}$
- $N(\tilde{e})$ the count of the target phrase $\tilde{e}$
- $N(\tilde{f}, \tilde{e})$ the co-occurence count of the source phrase $\tilde{f}$ and the target phrase $\tilde{e}$
- N the number of sentences in the bilingual corpus

we can compute the two-by-two contingency table in Table 1.

Following Fisher's exact test, we can calculate the probability of the contingency table via the hypergeometric distribution:

$$p_h(N(\tilde{f}, \tilde{e})) = \frac{\binom{N(\tilde{f})}{N(\tilde{f},\tilde{e})} \cdot \binom{N-N(\tilde{f})}{N(\tilde{e})-N(\tilde{f},\tilde{e})}}{\binom{N}{N(\tilde{e})}} \qquad (4)$$

The p-value is then calculated as the sum of all probabilities that are at least as extreme. The lower the p-value, the less likely this phrase pair occurred with the observed frequency by chance; we thus prune a phrase pair $(\tilde{f}, \tilde{e})$ if:

$$\left(\sum_{k=N(\tilde{f},\tilde{e})}^{\infty} p_h(k) \right) > \tau_F \qquad (5)$$

for some pruning threshold τ_F. More details of this approach can be found in Johnson et al. (2007). The idea of using Fisher's exact test was first explored by Moore (2004) in the context of word alignment.

5 Entropy-based Pruning

In this section, we will derive a novel entropy-based pruning criterion.

5.1 Motivational Example

In general, pruning the phrase table can be considered as selecting a subset of the original phrase table. When doing so, we would like to alter the original translation model distribution as little as possible. This is a key difference to previous approaches: Our goal is to remove *redundant* phrases, whereas previous approaches usually try to remove low-quality or unreliable phrases. We believe this to be an advantage of our method as it is certainly easier to measure the redundancy of phrases than it is to estimate their quality.

In Table 2, we show some example phrases from the learned French-English WMT phrase table, along with their counts and probabilities. For the French phrase *le gouvernement français*, we have, among others, two translations: *the French government* and *the government of France*. If we have to prune one of those translations, we can ask ourselves: how would the translation cost change if the

[3] Note that it has never been systematically investigated whether this is a real problem or just speculation.

$N(\tilde{f},\tilde{e})$	$N(\tilde{f}) - N(\tilde{f},\tilde{e})$	$N(\tilde{f})$
$N(\tilde{e}) - N(\tilde{f},\tilde{e})$	$N - N(\tilde{f}) - N(\tilde{e}) + N(\tilde{f},\tilde{e})$	$N - N(\tilde{f})$
$N(\tilde{e})$	$N - N(\tilde{e})$	N

Table 1: Two-by-two contingency table for a phrase pair $(\tilde{f},\tilde{e})$.

Source Phrase $\tilde{f}$	Target Phrase $\tilde{e}$	$N(\tilde{f},\tilde{e})$	$p(\tilde{e}\|\tilde{f})$
le	the	7.6 M	0.7189
gouvernement	government	245 K	0.4106
français	French	51 K	0.6440
	of France	695	0.0046
le gouvernement français	the French government	148	0.1686
	the government of France	11	0.0128

Table 2: Example phrases from the French-English phrase table (K=thousands, M=millions).

same translation were generated from the remaining, shorter, phrases? Removing the phrase *the government of France* would increase this cost dramatically. Given the shorter phrases from the table, the probability would be $0.7189 \cdot 0.4106 \cdot 0.0046 = 0.0014$[*], which is about an order of a magnitude smaller than the original probability of 0.0128.

On the other hand, composing the phrase *the French government* out of shorter phrases has probability $0.7189 \cdot 0.4106 \cdot 0.6440 = 0.1901$, which is very close to the original probability of 0.1686. This means it is safe to discard the phrase *the French government*, since the translation cost remains essentially unchanged. By contrast, discarding the phrase *the government of France* does not have this effect: it leads to a large change in translation cost.

Note that here the pruning criterion only considers redundancy of the phrases, not the quality. Thus, we are not saying that *the government of France* is a better translation than *the French government*, only that it is less redundant.

5.2 Entropy Criterion

Now, we are going to formalize the notion of redundancy. We would like the pruned model $p'(\tilde{e}|\tilde{f})$ to be as similar as possible to the original model $p(\tilde{e}|\tilde{f})$. We use conditional Kullback-Leibler divergence, also called conditional relative entropy (Cover and Thomas, 2006), to measure the model similarity:

$$D(p(\tilde{e}|\tilde{f})||p'(\tilde{e}|\tilde{f}))$$

$$= \sum_{\tilde{f}} p(\tilde{f}) \sum_{\tilde{e}} p(\tilde{e}|\tilde{f}) \log \left[\frac{p(\tilde{e}|\tilde{f})}{p'(\tilde{e}|\tilde{f})} \right] \quad (6)$$

$$= \sum_{\tilde{f},\tilde{e}} p(\tilde{e},\tilde{f}) \left[\log p(\tilde{e}|\tilde{f}) - \log p'(\tilde{e}|\tilde{f}) \right] \quad (7)$$

Computing the best pruned model of a given size would require optimizing over all subsets with that size. Since that is computationally infeasible, we instead apply the equivalent approximation that Stolcke (1998) uses for language modeling. This assumes that phrase pairs affect the relative entropy roughly independently.

We can then choose a pruning threshold τ_E and prune those phrase pairs with a contribution to the relative entropy below that threshold. Thus, we

[*] We use the assumption that we can simply multiply the probabilities of the shorter phrases.

prune a phrase pair $(\tilde{f}, \tilde{e})$, if

$$p(\tilde{e}, \tilde{f}) \left[\log p(\tilde{e}|\tilde{f}) - \log p'(\tilde{e}|\tilde{f}) \right] < \tau_E \quad (8)$$

We now address how to assign the probability $p'(\tilde{e}|\tilde{f})$ under the pruned model. A phrase-based system selects among different segmentations of the source language sentence into phrases. If a segmentation into longer phrases does not exist, the system has to compose a translation out of shorter phrases. Thus, if a phrase pair $(\tilde{f}, \tilde{e})$ is no longer available, the decoder has to use shorter phrases to produce the same translation. We can therefore decompose the pruned model score $p'(\tilde{e}|\tilde{f})$ by summing over all segmentations s_1^K and all reorderings π_1^K:

$$p'(\tilde{e}|\tilde{f}) = \sum_{s_1^K, \pi_1^K} p(s_1^K, \pi_1^K|\tilde{f}) \cdot p(\tilde{e}|s_1^K, \pi_1^K, \tilde{f}) \quad (9)$$

Here the segmentation s_1^K divides both the source and target phrases into K sub-phrases:

$$\tilde{f} = \bar{f}_{\pi_1} ... \bar{f}_{\pi_K} \text{ and } \tilde{e} = \bar{e}_1 ... \bar{e}_K \quad (10)$$

The permutation π_1^K describes the alignment of those sub-phrases, such that the sub-phrase $\bar{e}_k$ is aligned to $\bar{f}_{\pi_k}$. Using the normal phrase translation model, we obtain:

$$p'(\tilde{e}|\tilde{f}) = \sum_{s_1^K, \pi_1^K} p(s_1^K, \pi_1^K|\tilde{f}) \prod_{k=1}^{K} p(\bar{e}_k|\bar{f}_{\pi_k}) \quad (11)$$

Virtually all phrase-based decoders use the so-called maximum-approximation, i. e. the sum is replaced with the maximum. As we would like the pruning criterion to be similar to the search criterion used during decoding, we do the same and obtain:

$$p'(\tilde{e}|\tilde{f}) \approx \max_{s_1^K, \pi_1^K} \prod_{k=1}^{K} p(\bar{e}_k|\bar{f}_{\pi_k}) \quad (12)$$

Note that we also drop the segmentation probability, as this is not used at decoding time. This leaves the pruning criterion a function only of the model $p(\tilde{e}|\tilde{f})$ as stored in the phrase table. There is no need for a special development or adaptation set. We can determine the best segmentation using dynamic programming, similar to decoding with a phrase-based model. However, here the target side is constrained to the given phrase $\tilde{e}$.

It can happen that a phrase is not compositional, i. e., we cannot find a segmentation into shorter phrases. In these cases, we assign a small, constant probability:

$$p'(\tilde{e}|\tilde{f}) = p_c \quad (13)$$

We found that the value $p_c = e^{-10}$ works well for many language pairs.

5.3 Computation

In our experiments, it was more efficient to vary the pruning threshold τ_E without having to re-compute the entire phrase table. Therefore, we computed the entropy criterion in Equation (8) once for the whole phrase table. This introduces an approximation for the pruned model score $p'(\tilde{e}|\tilde{f})$. It might happen that we prune short phrases that were used as part of the best segmentation of longer phrases. As these shorter phrases should not be available, the pruned model score might be inaccurate. Although we believe this effect is minor, we leave a detailed experimental analysis for future work.

One way to avoid this approximation would be to perform entropy pruning with increasing phrase length. Starting with one-word phrases, which are trivially non-compositional, the entropy criterion would be straightforward to compute. Proceeding to two-word phrases, one would decompose the phrases into sub-phrases by looking up the probabilities of some of the unpruned one-word phrases. Once the set of unpruned two-word phrases was obtained, one would continue with three-word phrases, etc.

6 Experimental Evaluation

6.1 Data Sets

In this section, we describe the data sets used for the experiments. We perform experiments on the publicly available WMT shared translation task for the following four language pairs:

- German-English
- Czech-English
- Spanish-English

	Number of Words	
Language Pair	Foreign	English
German - English	42 M	45 M
Czech - English	56 M	65 M
Spanish - English	232 M	210 M
French - English	962 M	827 M

Table 3: Training data statistics. Number of words in the training data (M=millions).

- French-English

For each pair, we train two separate system, one for each direction. Thus it can happen that a phrase is pruned for X-to-Y, but not for Y-to-X.

These four language pairs represent a nice range of training corpora sizes, as shown in Table 3.

6.2 Baseline System

Pruning experiments were performed on top of the following baseline system. We used a phrase-based statistical machine translation system similar to (Zens et al., 2002; Koehn et al., 2003; Och and Ney, 2004; Zens and Ney, 2008). We trained a 4-gram language model on the target side of the bilingual corpora and a second 4-gram language model on the provided monolingual news data. All language models used Kneser-Ney smoothing.

The baseline system uses the common phrase translation models, such as $p(\tilde{e}|\tilde{f})$ and $p(\tilde{f}|\tilde{e})$, lexical models, word and phrase penalty, distortion penalty as well as a lexicalized reordering model (Zens and Ney, 2006).

The word alignment was trained with six iterations of IBM model 1 (Brown et al., 1993) and 6 iterations of the HMM alignment model (Vogel et al., 1996) using a symmetric lexicon (Zens et al., 2004).

The feature weights were tuned on a development set by applying minimum error rate training (MERT) under the Bleu criterion (Och, 2003; Macherey et al., 2008). We ran MERT once with the full phrase table and then kept the feature weights fixed, i. e., we did *not* rerun MERT after pruning to avoid adding unnecessary noise. We extract phrases up to a length of six words. The baseline system already includes phrase table pruning by removing singletons and keeping up to 30 target language phrases per source phrase. We found that this does not affect translation quality significantly[4]. All pruning experiments are done on top of this.

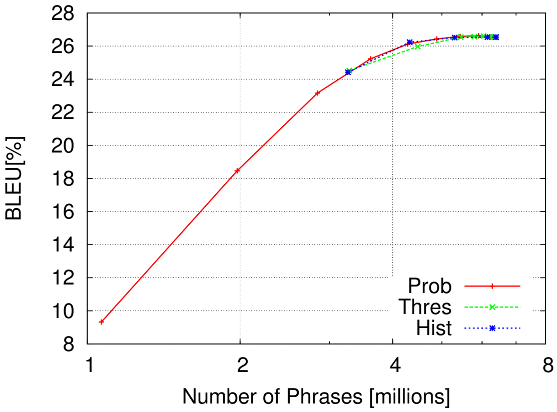

Figure 1: Comparison of probability-based pruning methods for German-English.

6.3 Results

In this section, we present the experimental results. Translation results are reported on the WMT'07 news commentary blind set.

We will show translation quality measured with the Bleu score (Papineni et al., 2002) as a function of the phrase table size (number of phrases). Being in the upper left corner of these figures is desirable.

First, we show a comparison of several probability-based pruning methods in Figure 1. We compare

- **Prob**. Absolute pruning based on Eq. (2).
- **Thres**. Threshold pruning based on Eq. (3).
- **Hist**. Histogram pruning as described in Section 3.2.[5]

We observe that these three methods perform equally well. There is no difference between absolute and relative pruning methods, except that the two relative methods (Thres and Hist) are limited by

[4] The Bleu score drops are as follows: English-French 0.3%, French-English 0.4%, Czech-English 0.3%, all other are less than 0.1%.

[5] Instead of using $p(\tilde{e}|\tilde{f})$ one could use the weighted model score including $p(\tilde{f}|\tilde{e})$, lexical weightings etc.; however, we found that this does not give significantly different results; but it does introduce a undesirable dependance between feature weights and phrase table pruning.

the number of source phrases. Thus, they reach a point where they cannot prune the phrase table any further. The results shown are for German-English; the results for the other languages are very similar. The results that follow use only the absolute pruning method as a representative for probability-based pruning.

In Figures 2 through 5, we show the translation quality as a function of the phrase table size. We vary the pruning thresholds to obtain different phrase table sizes. We compare four pruning methods:

- **Count**. Pruning based on the frequency of a phrase pair, c.f. Equation (1).
- **Prob**. Pruning based on the absolute probability of a phrase pair, c.f. Equation (2).
- **Fisher**. Pruning using significance tests, c.f. Equation (5).
- **Entropy**. Pruning using the novel entropy criterion, c.f. Equation (8).

Note that the x-axis of these figures is on a logarithmic scale, so the differences between the methods can be quite dramatic. For instance, entropy pruning requires less than a quarter of the number of phrases needed by count- or significance-based pruning to achieve a Spanish-English Bleu score of 34 (0.4 million phrases compared to 1.7 million phrases).

These results clearly show how the pruning methods compare:

1. Probability-based pruning performs poorly. It should be used only to prune small fractions of the phrase table.
2. Count-based pruning and significance-based pruning perform equally well. They are much better than probability-based pruning.
3. Entropy pruning consistently outperforms the other methods across translation directions and language pairs.

Figures 6 and 7 show compositionality statistics for the pruned Spanish-English phrase table (we observed similar results for the other language pairs). Each figure shows the composition of the phrase table for a type of pruning for different phrase tables sizes. Along the x-axis, we plotted the phrase table size. These are the same phrase tables used to obtain the Bleu scores in Figure 2 (left). The different shades of grey correspond to different phrase lengths. For instance, in case of the smallest phrase table for count-based pruning, the 1-word phrases account for about 30% of all phrases, the 2-word phrases account for about 35% of all phrases, etc.

Total number of phrases	4 137 M
Compositional	3 970 M
Non-compositional	167 M
of those: one-word phrases	85 M
no segmentation	82 M

Table 4: Statistics of phrase compositionality (M=millions).

With the exception of the probability-based pruning, the plots look comparable. The more aggressive the pruning, the larger the percentage of short phrases. We observe that entropy-based pruning removes many more long phrases than any of the other methods. The plot for probability-based pruning is different in that the percentage of long phrases actually increases with more aggressive pruning (i. e. smaller phrase tables). A possible explanation is that probability-based pruning does *not* take the frequency of the source phrase into account. This difference might explain the poor performance of probability-based pruning.

To analyze how many phrases are compositional, we collect statistics during the computation of the entropy criterion. These are shown in Table 4, accumulated across all language pairs and all phrases, i. e., including singleton phrases. We see that 96% of all phrases are compositional (3 970 million out of 4 137 million phrases). Furthermore, out of the 167 million non-compositional phrases, more than half (85 million phrases), are trivially non-compositional: they consist only of a single source or target language word. The number of non-trivial non-compositional phrases is, with 82 million or 2% of the total number of phrases, very small.

In Figure 8, we show the effect of the constant

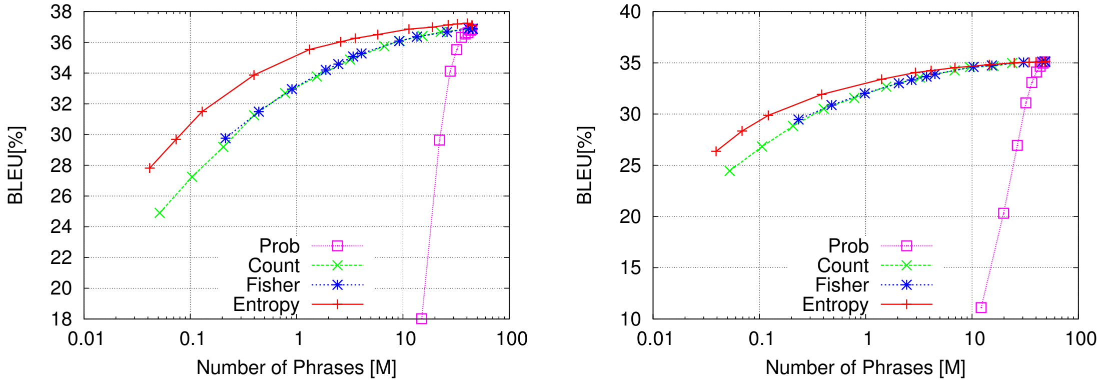

Figure 2: Translation quality as a function of the phrase table size for Spanish-English (left) and English-Spanish (right).

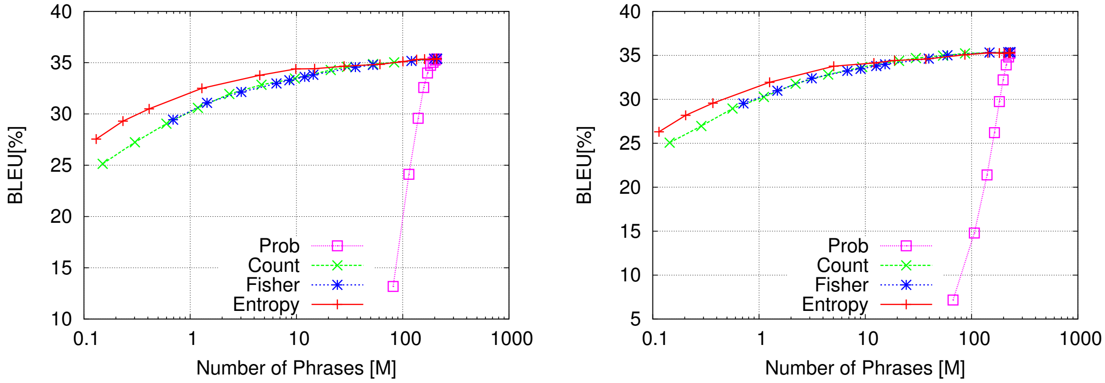

Figure 3: Translation quality as a function of the phrase table size for French-English (left) and English-French (right).

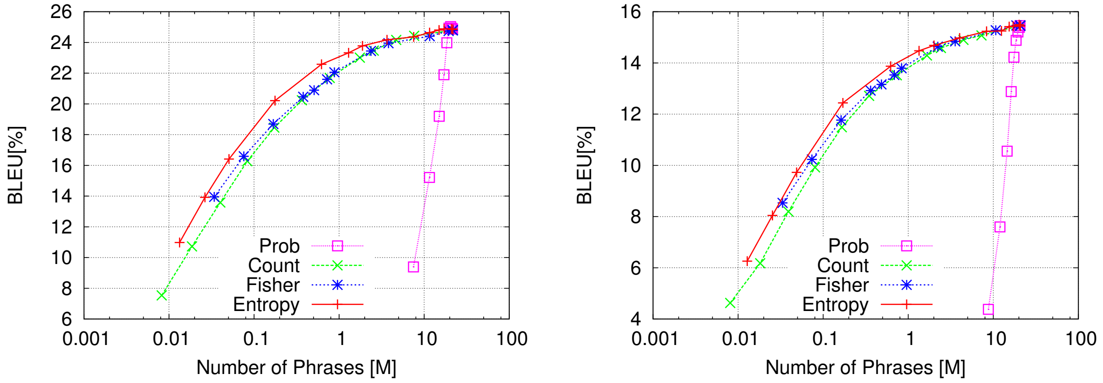

Figure 4: Translation quality as a function of the phrase table size for Czech-English (left) and English-Czech (right).

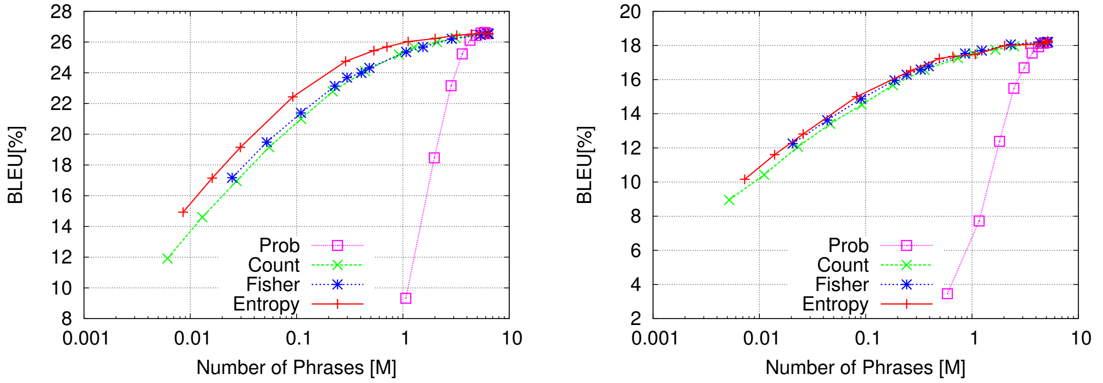

Figure 5: Translation quality as a function of the phrase table size for German-English (left) and English-German (right).

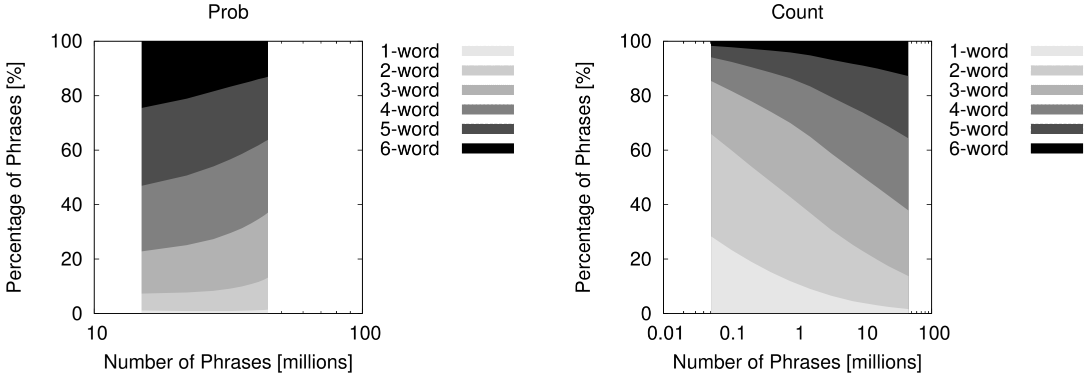

Figure 6: Phrase length statistics for Spanish-English for probability-based (left) and count-based pruning (right).

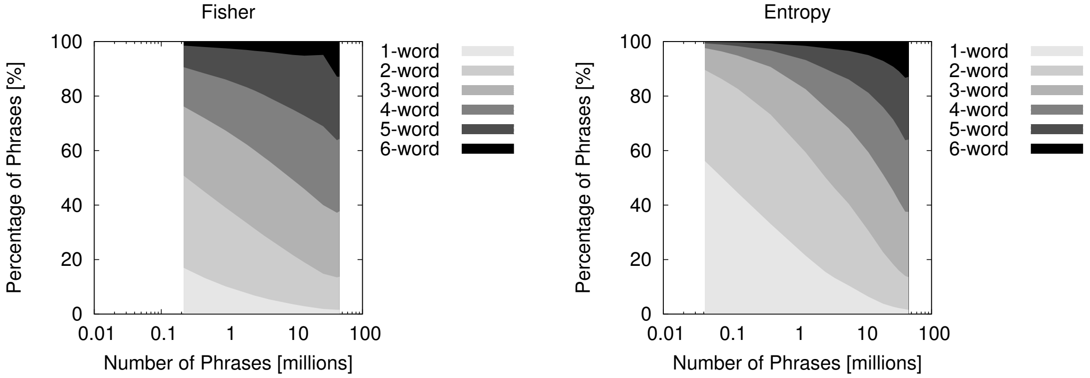

Figure 7: Phrase length statistics for Spanish-English for significance-based (left) and entropy-based pruning (right).

p_c for non-compositional phrases.[6] The results shown are for Spanish-English; additional experiments for the other languages and translation directions showed very similar results. Overall, there is no big difference between the values. Hence, we chose a value of 10 for all experiments.

The results in Figure 2 to Figure 5 show that entropy-based pruning clearly outperforms the alternative pruning methods. However, it is a bit hard to see from the graphs exactly how much additional savings it offers over other methods. In Table 5, we show how much of the phrase table we have to retain under various pruning criteria without losing more than one Bleu point in translation quality. We see that probability-based pruning allows only for marginal savings. Count-based and significance-based pruning results in larger savings between 70% and 90%, albeit with fairly high variability. Entropy-based pruning achieves consistently high savings between 85% and 95% of the phrase table. It always outperforms the other pruning methods and yields significant savings on top of count-based or significance-based pruning methods. Often, we can cut the required phrase table size in half compared to count or significance based pruning.

As a last experiment, we want to confirm that phrase-table pruning methods are actually better than simply reducing the maximum phrase length. In Figure 9, we show a comparison of different pruning methods and a length-based approach for Spanish-English. For the 'Length' curve, we first drop all 6-word phrases, then all 5-word phrases, etc. until we are left with only single-word phrases; the phrase length is measured as the number of source language words. We observe that entropy-based, count-based and significance-based pruning indeed outperform the length-based approach. We obtained similar results for the other languages.

[6] The values are in neg-log-space, i. e., a value of 10 corresponds to $p_c = e^{-10}$.

Method	ES-EN	EN-ES	DE-EN	EN-DE	FR-EN	EN-FR	CS-EN	EN-CS
Prob	77.3 %	82.7 %	61.2 %	67.3 %	84.8 %	94.1 %	85.6 %	86.3 %
Count	24.9 %	11.9 %	19.9 %	14.3 %	11.4 %	9.0 %	20.2 %	10.4 %
Fisher	23.5 %	12.6 %	21.7 %	14.0 %	14.5 %	13.6 %	31.9 %	9.9 %
Entropy	**7.2** %	**6.0** %	**10.2** %	**11.1** %	**7.1** %	**8.1** %	**14.8** %	**6.4** %

Table 5: To what degree can we prune the phrase table without losing more than 1 Bleu point? The table shows percentage of phrases that we have to retain. ES=Spanish, EN=English, FR=French, CS=Czech, DE=German.

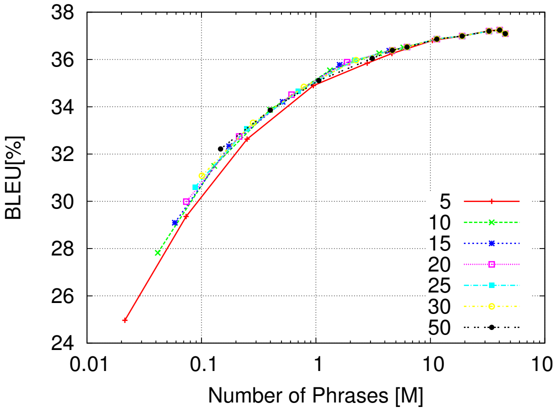

Figure 8: Translation quality (Bleu) as a function of the phrase table size for Spanish-English for entropy pruning with different constants p_c.

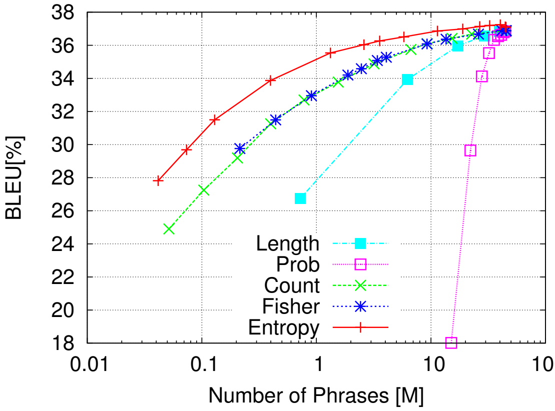

Figure 9: Translation quality (Bleu) as a function of the phrase table size for Spanish-English.

7 Conclusions

Phrase table pruning is often addressed in an ad-hoc way using the heuristics described in Section 3. We have shown that some of those do *not* work well. Choosing the wrong technique can result in significant drops in translation quality without saving much in terms of phrase table size. We introduced a novel entropy-based criterion and put phrase table pruning on a sound theoretical foundation. Furthermore, we performed a systematic experimental comparison of existing methods and the new entropy criterion. The experiments were carried out for four language pairs under small, medium and large data conditions. We can summarize our conclusions as follows:

- Probability-based pruning performs poorly when pruning large parts of the phrase table. This might be because it does not take the frequency of the source phrase into account.
- Count-based pruning performs as well as significance-based pruning.
- Entropy-based pruning gives significantly larger savings in phrase table size than any other pruning method.
- Compared to previous work, the novel entropy-based pruning often achieves the same Bleu score with only half the number of phrases.

8 Future Work

Currently, we take only the model $p(\tilde{e}|\tilde{f})$ into account when looking for the best segmentation. We might obtain a better estimate by also considering the distortion costs, which penalize reordering. We could also include other phrase models such as $p(\tilde{f}|\tilde{e})$ and the language model.

The entropy pruning criterion could be applied to hierarchical machine translation systems (Chiang, 2007). Here, we might observe even larger reductions in phrase table size as there are many more entries.

References

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Yu Chen, Andreas Eisele, and Martin Kay. 2008. Improving statistical machine translation efficiency by triangulation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Yu Chen, Martin Kay, and Andreas Eisele. 2009. Intersecting multilingual data for faster and better statistical translations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 128–136, Boulder, Colorado, June. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory*. Wiley-Interscience, New York, NY, USA.

Nan Duan, Mu Li, and Ming Zhou. 2011. Improving phrase extraction via MBR phrase scoring and pruning. In *Proceedings of MT Summit XIII*, pages 189–197, Xiamen, China, September.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2007a. Estimating phrase pair relevance for machine translation pruning. In *Proceedings of MT Summit XI*, pages 159–165, Copenhagen, Denmark, September.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2007b. Translation model pruning via usage statistics for statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 21–24, Rochester, New York, April. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

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 *45th Annual Meeting of the Assoc. for Computational Linguistics (ACL): Poster Session*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *6th Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 115–124, Washington DC, September/October.

Wolfgang Macherey, Franz Och, Ignacio Thayer, and Jakob Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 725–734, Honolulu, HI, October. Association for Computational Linguistics.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of* *the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 258–267, Portland, Oregon, USA, June. Association for Computational Linguistics.

German Sanchis-Trilles, Daniel Ortiz-Martinez, Jesus Gonzalez-Rubio, Jorge Gonzalez, and Francisco Casacuberta. 2011. Bilingual segmentation for phrasetable pruning in statistical machine translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 257–264, Leuven, Belgium, May.

Andreas Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.

David Talbot and Miles Osborne. 2007. Smoothed Bloom filter language models: Tera-scale LMs on the cheap. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 468–476, Prague, Czech Republic, June. Association for Computational Linguistics.

Nadi Tomeh, Nicola Cancedda, and Marc Dymetman. 2009. Complexity-based phrase-table filtering for statistical machine translation. In *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, August.

Nadi Tomeh, Marco Turchi, Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2011. How good are your phrases? Assessing phrase quality with single class classification. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 261–268, San Francisco, California, December.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *16th Int. Conf. on Computational Linguistics (COLING)*, pages 836–841, Copenhagen, Denmark, August.

Mei Yang and Jing Zheng. 2009. Toward smaller, faster, and better hierarchical phrase-based SMT. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 237–240, Suntec, Singapore, August. Association for Computational Linguistics.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL): Workshop on Statistical Machine Translation*, pages 55–63, New York City, NY, June.

Richard Zens and Hermann Ney. 2008. Improvements in dynamic programming beam search for phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 195–205, Honolulu, Hawaii, October.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *25th German Conf. on Artificial Intelligence (KI2002)*, volume 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 18–32, Aachen, Germany, September. Springer Verlag.

Richard Zens, Evgeny Matusov, and Hermann Ney. 2004. Improved word alignment using a symmetric lexicon model. In *20th Int. Conf. on Computational Linguistics (COLING)*, pages 36–42, Geneva, Switzerland, August.