Improving Relative-Entropy Pruning using Statistical Significance

Wang Ling[1,2] *Nadi Tomeh*[3]
Guang Xiang[1] *Alan Black*[1] *Isabel Trancoso*[2]
(1)Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
(2)L^{2}F Spoken Systems Lab, INESC-ID, Lisboa, Portugal
(3)LIMSI-CNRS and Université Paris-Sud Orsay, France
{lingwang,guangx,awb}@cs.cmu.edu, nadi.tomeh@limsi.fr,
isabel.trancoso@inesc-id.pt

Abstract

Relative Entropy-based pruning has been shown to be efficient for pruning language models for more than a decade ago. Recently, this method has been applied to Phrase-based Machine Translation, and results suggest that this method is comparable the state-of-art pruning method based on significance tests. In this work, we show that these 2 methods are effective in pruning different types of phrase pairs. On one hand, relative entropy pruning searches for phrase pairs that can be composed using smaller constituents with a small or no loss in probability. On the other hand, significance pruning removes phrase pairs that are likely to be spurious. Then, we show that these methods can be combined in order to produce better results, over both metrics when used individually.

Proceedings of COLING 2012: Posters, pages 713–722,
COLING 2012, Mumbai, December 2012.

1 Introduction

Statistical Machine Translation systems are generally built on large amounts of parallel data. Typically, the training sentences are first aligned at the word level, then all phrase pairs that are consistent with the word alignment are extracted, scored and stored in the phrase table. While such extraction criterion performs well in practice, it produces translation models that are unnecessarily large with many phrase pairs that are useless for translation. This is undesirable at decoding time, since it leads to more search errors due to the large search space. Furthermore, larger models are more expensive to store, which limits the portability of such models to smaller devices.

Pruning is one approach to address this problem, where models are made more compact by discarding entries from the model, based on additional selection criteria. The challenge in this task is to choose the entries that will least degenerate the quality of the task for which the model is used. For language models, an effective algorithm based on relative entropy is described in (Seymore and Rosenfeld, 1996; Stolcke, 1998; Moore and Quirk, 2009). In these approaches, a criteria based on the KL divergence is applied, so that higher order n-grams are only included in the model when they provide enough additional information to the model, given the lower order n-grams. Recently, this concept was applied for translation model pruning (Ling et al., 2012; Zens et al., 2012), and results indicate that this method yields a better phrase table size and translation quality ratio than previous methods, such as the well known method in (Johnson et al., 2007), which uses the Fisher's exact test to calculate how well a phrase pair is supported by data.

In this work, we attempt to improve the relative entropy model, by combining it with the significance based approach presented in (Johnson et al., 2007).The main motivation is that, as suggested in (Ling et al., 2012), relative entropy and significance based methods are complementary. On one hand, relative entropy aims at pruning phrase pairs that can be reproduced using smaller constituents with a small or no loss in terms of the models predictions. On the other hand, significance pruning aims at removing phrase pairs that are spurious, and are originated from incorrect alignments at sentence or word level. This indicates that both methods can be combined to obtain better results. We propose a log-linear interpolation of the two metrics to achieve a better trade off between the number of phrase pairs and the translation quality.

This paper is structured as follows: Section 2 includes a brief summary of relative entropy and significance pruning approaches in sub-sections 2.1 and 2.2. Sub-section 2.3 analyses both algorithms and preceeds our combination approach in sub-section 2.4. The results obtained with the EUROPARL corpus (Koehn, 2005) are shown in Section 3. Finally, we conclude and present directions for future research in Section 4.

2 Combining Relative Entropy and Significance Pruning

In principle, any method of evaluation of phrase pairs can be used as the basis for pruning. This includes phrase counts and probabilities (Koehn et al., 2003), statistical significance tests (Johnson et al., 2007), and relative entropy scores (Ling et al., 2012; Zens et al., 2012) and many others (Deng et al., 2008; Venugopal et al., 2003; Tomeh et al., 2011), in addition to the features typically found in phrase tables (Och et al., 2004; Chiang et al., 2009). Each method reflects some characteristics of phrase pairs that are not sought by the other, and hence trying to combine them is a tempting idea. (Deng et al., 2008) incorporate several features into a log-linear model parametrized with y_k that are tuned, along with the extraction threshold, to maximize a translation quality, which makes the procedure extremely expensive. A similar model is used in (Venugopal et al., 2003) without any parameter tuning. (Zettlemoyer and Moore, 2007) use an already tuned model (using MERT) in a competitive linking algorithm to keep the best one-to-one phrase matching in each training sentence. In our work we favor efficiency and we focus on relative entropy and significance pruning, which can be efficiently computed, without the need to external information. They also deliver good practical performance.

2.1 Relative Entropy Pruning

Relative entropy pruning for translation models (Ling et al., 2012; Zens et al., 2012) has a solid foundation on information theory. The goal in these methods is to find a pruned model $P_p(t|s)$ that yields predictions that are as close as possible as the original model $P(t|s)$. More formally, we want to minimize the relative entropy or KL divergence between these models, expressed as follows:

$$D(P_p||P) = -\sum_{s,t} P(s,t) log \frac{P_p(t|s)}{P(t|s)} \tag{1}$$

In another words, for each phrase pair with source s and target t, we calculate the log difference between their probabilities $log\frac{P_p(t|s)}{P(t|s)}$. This value is then weighted by the empirical distribution $P(s,t)$, so that phrase pairs that are more likely to be observed in the data are less likely to be pruned. The empirical distribution is given as:

$$P(s,t) = \frac{C(s,t)}{N} \tag{2}$$

Where $C(s,t)$ denotes, the number of sentence pairs where s and t are observed, and N denotes the number of sentence pairs.

Computing $P_p(t|s)$ is the most computationally expensive operation in this model, since it involves finding all possible derivations of a phrase pair using smaller units, which involves a forced decoding step (Schwartz, 2008).

While minimizing $D(P_p||P)$ would lead to optimal results, such optimization is computationally infeasible. Thus, an approximation is the find the local values for each phrase pair:

$$\text{RelEnt(s,t)} = -P(s,t) log \frac{P_p(t|s)}{P(t|s)} \tag{3}$$

This score can be viewed as the relative entropy between $P_p(t|s)$ and $P(t|s)$, if only the phrase pair with source s and target t is pruned. The problem with this approximation is that, we might assume a given phrase pair A can be pruned, because it can be composed by phrase pairs B and C, only to discover later that B is also pruned.

2.2 Significance Pruning

Significance pruning of phrase tables (Johnson et al., 2007; Tomeh et al., 2009) relies on a statistical test that assesses the strength of the association between the source and target phrases in a phrase pair. Such association can be represented using a two-by-two contingency table:

$C(s,t)$	$C(s)-C(s,t)$
$C(t)-C(s,t)$	$N-C(s)-C(t)+C(s,t)$

where N is the size of the training parallel corpus, $C(s)$ is the count of the source phrase, $C(t)$ is the count of the target phrase, and $C(s,t)$ is the count of the co-occurences of s and t. The probability of this particular table is given by the the hypergeometric distribution:

$$p_h(C(s,t)) = \frac{\binom{C(s)}{C(s,t)}\binom{N-C(s)}{C(t)-C(s,t)}}{\binom{N}{C(t)}}.$$

The p-value correponds to the probability that s and t co-occur at least $C(s,t)$ times only due to chance. It is computed by Fisher's exact test by summing the probabilities of all contingency tables that are at least as extreme:

$$\text{p-value}(C(s,t)) = \sum_{k=C(s,t)}^{\infty} p_h(k).$$

We define the association score to be $-\log(\text{p-value})$ which varies between 0 and $infty$. The higher the association score, the less likely this phrases s and t co-occurred with the observed count $C(s,t)$ by chance.

2.3 Error Analysis

Table 1 shows examples of phrase table entries that are likely to be pruned for each method for a translation model using he EUROPARL dataset with 1.2M sentence pairs. The phrase pairs were chosen from the list of phrase pairs that would be pruned if we only pruned 1% of the table. We can see that both methods aim at pruning different types of phrase pairs.

In significance pruning, we observe that most of the filtered phrase pairs are spurious phrase pairs. These phrase pairs are generally originated from sentence level mis-alignments, which can occur in automatically aligned corpora. Another possible origin for spurious phrase pairs are Word-alignment errors. We can see that relative entropy pruning is not the best approach to address with these problems. For instance, if we calculate the divergence $log\frac{P_p(t|s)}{P(t|s)}$ for the phrase pair with source "it" and target "+", we will obtain $log(0)$, since it is cannot be composed using smaller units. Thus, it is unlikely that these phrase pairs will be pruned by relative pruning. Note, that while it is true that spurious phrase pairs will have a low empirical distribution probability, the same is true will longer and sparser phrase pairs that are actually correct, and in such cases the relative entropy model will prefer to prune the longer phrase pairs, since they can be composed using smaller constituents, which is not desired.

On the other hand, there is nothing intrinsically wrong in the phrase pairs that are pruned by relative entropy pruning. However, these phrase pairs are redundant and can be easily translated using smaller units. For instance, it is not surprising that the source phrase “0.005 %" can be translated “0.005 %", using the smaller units, “0.005" to “0.005" and “%" to “%", since it is unlikely that “0.005" or “%" to be translated another target phrase, or have a non-monotonous reordering. In significance pruning, for a moderately large corpora, it is unlikely that this phrase pair would be pruned early, since it is likely that the phrase pair is well supported by data.

Significance		Relative Entropy	
English	French	English	French
it	+	2 6 8 10 and	2 6 8 10 et
with	, entre	0,005 %	0.005 %
a	, un accord a été	! ! !	! ! !

Table 1: Selected examples of phrase pairs that have low scores according to Significance pruning (Left) and Relative Entropy pruning (Right). The examples are selected from the model built using the EUROPARL training dataset for French and English.

Thus, we can see that it is prominent that relative entropy and significance methods are complementary in terms of what types of phrase pairs that are pruned. We can see that all the phrase pairs pruned by significance pruning in the table would be unlikely to be pruned by relative entropy pruning, since these phrase pairs only have one word in the target side and so they cannot be decomposed into smaller units. On the other hand, it is also unlikely that the phrase pairs that are pruned using relative entropy, are pruned by significance pruning, since these phrases are well aligned and likely to be well supported by data.

2.4 Combination Method

In our work, we will attempt to achieve a better trade off between the number of phrase pairs that are pruned due to their redundancy and due to their spurious nature.

There are many different approaches that can be taken to combine these two scores. For instance, in Phrase-based machine translation multiple features are combined using a log-linear model. Thus, we could use a similar approach and set a weight α and combine the two scores as follows:

$$Score(s,t) = \alpha RelEnt(s,t) + (1-\alpha)Sig(s,t) \tag{4}$$

Where $RelEnt(s,t) = -P(\tilde{s},\tilde{t})log\frac{P_p(t|s)}{P(t|s)}$ is the relative entropy score and $Sig(s,t) = -log$ p-value$(C(s,t))$ is the significance score of the phrase pair with source s and target t.

However, one problem with this approach is that classification boundary for these two features does not seem to be linear from our analysis, especially since these features seem to be orthogonal. For instance, suppose that we have a phrase pair with a very high score using relative entropy (for instance 300), meaning that the phrase pair is definitely not redundant. However, in terms of p-value, the phrase pair is scored with a with a extremely low value (such as 10), which means that it is very likely that the phrase pair is not well-formed. If we simply interpolate the scores, we would expect the score of the phrase pair to be 155, with $\alpha = 0.5$, which is an average score. This is not necessarily a good decision, because regardless of how unique a phrase pair is, if the phrase pair is spurious it should not be kept in the model. The opposite is also true, if a phrase pair is well-formed, but it can be built using smaller phrase pairs, it means that it can be removed, since it is not useful in the model.

In another words, good phrase pairs must be well-formed and not redundant. Thus, we propose to select the minimum of the two scores rather than their average. More formally, we score each phrase pair as:

$$Score(s,t) = min(\alpha RelEnt(s,t), (1-\alpha)Sig(s,t)) \quad (5)$$

We still apply the scaling factor α, so that we can specify which score has a higher weight.

Using this score, for the example above, the phrase pair would be scored with the significance score of 10.

3 Experimental Results

3.1 Data Sets

Experiments were performed using the publicly available EUROPARL (Koehn, 2005) corpora for the English-French language pair. From this corpus, 1.2M sentence pairs were selected for training, 2000 for tuning and another 2000 for testing.

3.2 Baseline System

The baseline translation system was trained using a conventional pipeline similar to the one described in (Koehn et al., 2003).

First, the word alignments were generated using IBM model 4.

Then, the translation model was generate using the phrase extraction algorithm (Paul et al., 2010)(Koehn et al., 2007). The maximum size of the phrase pairs is set to 7, both for the source and the target language. The model uses as features:

- Translation probability
- Reverse translation probability
- Lexical translation probability
- Reverse lexical translation probability
- Phrase penalty

The reordering model is built using the lexicalized reordering model described in (Axelrod et al., 2005), with MSD (mono, swap and discontinuous) reordering features for orientations.

All the translation and reordering features are considered during the calculation of the relative entropy. As in (Zens et al., 2012), we removed all singleton phrase pairs from the phrase table. This will lower the effectiveness of significance pruning, since a large amount of least significant phrase pairs will be removed a priori. The filtered translation model contains, approximately 50 million phrase pairs.

As language model, a 5-gram model with Kneser-ney smoothing was used.

The baseline model was tuned using MERT tuning (Och, 2003). We did not rerun tuning again after pruning to avoid adding noise to the results.

Finally, we present the results evaluated with BLEU-4 (Papineni et al., 2002).

After computing the negative log likelihood of both scores, we also rescale both score's values by mean, so that scores will have similar values. This step is performed so the interpolation weights, in the results appear more intuitive.

3.3 Results

We can see the results in table 2, where the first two rows, represent the BLEU scores for relative entropy pruning and significance pruning, respectively. Then, we have the scores obtained using the scorer in 4 of these 2 scores, with α weights at intervals of 0.1. Finally, we have the scores using the scorer 5, also with the weight α set at intervals of 0.1.

From the results, we observe that using relative entropy pruning, we obtain better translation quality in terms of BLEU than significance pruning until 20%, where significance pruning works considerably better. This is because, at 20%, relative entropy pruning starts having to discard phrase pairs that have no smaller constituents, relying only on the empirical distribution. Thus, we would like to perform better by considering both scores.

However, we can see that using linear interpolation does not improve the results. This is because, as stated before, the two scores evaluate different aspects of phrase pairs. Thus, performing a weighted average of these two scores will simply degenerate the precision of the pruning decision. For instance, if one phrase pair has a 0 value according to relative entropy, implying that it is redundant, while the significance score is 300, because the phrase pair is well aligned, a uniform linear interpolation would give this phrase pair the score of 150. This is not the effect we desire, since if a phrase pair is classified is classified as redundant, it can be discarded regardless of how well-formed it is. The same applies to phrase pairs that are not-redundant but not significant. As we can see from the results, we can obtain results that range between the scores for using significance pruning and relative entropy pruning separately, but not improve over both of them.

On the other hand, we can see that using an weighted minimum of the 2 scores achieves much better results. We can see that results are equally good at higher phrase table sizes as the relative entropy. This indicates that at higher phrase table sizes, the pruning choices are governed by relative entropy pruning. At lower phrase tables sizes, we can see that we can achieve better results than each of the methods separately, where the 2 scores are combined to make better pruning decisions. Specifically, 20% of the phrase table size, the combined method for the best α (0.5) achives 27.16 BLEU points which is 0.3(1%) points over the significance pruning method and 1.51(6%) points over relative entropy pruning.

Conclusion

In this work, we evaluated two state of the art methods for translation model pruning, one based on significance tests and one based on relative entropy. While the former is effective at removing phrase pairs that are result of misalignments, the latter aims at removing phrase pairs that are redundant, since they can be formed using other phrase pairs. We showed that 2 the methods are complementary and a better pruning methodology can be obtained by combining them. We showed empirically that using linear interpolation is not the best approach to combine these scores, and better results can be obtained by taking the minimum from both scores at each data point.

The code used for calculating relative entropy and combining scores presented in this paper is currently integrated with MOSES[1].

Acknowledgments

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.

[1] available at https://github.com/moses-smt/mosesdecoder

Experiment	100%	80%	60%	40%	20%
Relative Entropy	27.50	**27.50**	**27.51**	27.37	25.65
Significance	27.50	27.39	27.24	27.20	26.86
Avg($\alpha = 0.9$)	27.50	27.48	27.48	27.35	26.21
Avg($\alpha = 0.8$)	27.50	27.48	27.46	27.36	26.21
Avg($\alpha = 0.7$)	27.50	27.49	27.46	27.34	26.21
Avg($\alpha = 0.6$)	27.50	27.49	27.43	27.32	26.21
Avg($\alpha = 0.5$)	27.50	27.48	27.43	27.33	26.21
Avg($\alpha = 0.4$)	27.50	27.47	27.44	27.31	26.21
Avg($\alpha = 0.3$)	27.50	27.48	27.36	27.31	26.21
Avg($\alpha = 0.2$)	27.50	27.47	27.38	27.31	26.21
Avg($\alpha = 0.1$)	27.50	27.46	27.37	27.31	26.21
Min($\alpha = 0.9$)	27.50	**27.50**	**27.51**	27.37	27.06
Min($\alpha = 0.8$)	27.50	**27.50**	**27.51**	**27.42**	27.15
Min($\alpha = 0.7$)	27.50	**27.50**	**27.51**	27.39	27.12
Min($\alpha = 0.6$)	27.50	**27.50**	**27.51**	27.38	27.11
Min($\alpha = 0.5$)	27.50	**27.50**	**27.51**	27.35	**27.16**
Min($\alpha = 0.4$)	27.50	**27.50**	**27.51**	27.36	27.14
Min($\alpha = 0.3$)	27.50	**27.50**	27.49	27.39	27.11
Min($\alpha = 0.2$)	27.50	**27.50**	27.49	27.41	27.15
Min($\alpha = 0.1$)	27.50	**27.50**	27.48	27.37	27.11

Table 2: Results for the EN-FR EUROPARL CORPORA. Each Column represents the size of the phrase table and each row represents a different pruning score. Each cell represents the BLEU score using a 2000 sentence pair test set.

References

Axelrod, A., Mayne, R. B., Callison-burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 218–226. Association for Computational Linguistics.

Deng, Y., Xu, J., and Gao, Y. (2008). Phrase table training for precision and recall: What makes a good phrase and a good phrase pair? In *Proceedings of ACL-08: HLT*, pages 81–88, Columbus, Ohio. Association for Computational Linguistics.

Johnson, J. H., Martin, J., Foster, G., and Kuhn, R. (2007). Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL 07*, pages 967–975.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Zens, R., Aachen, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C., and Bojar, O. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Ling, W., Graça, J., Trancoso, I., and Black, A. (2012). Entropy-based pruning for phrase-based machine translation. In *EMNLP-CoNLL*, pages 962–971.

Moore, R. C. and Quirk, C. (2009). Less is more: significance-based n-gram selection for smaller, better language models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 746–755, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D. A., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting on ACL*, pages 311–318.

Paul, M., Federico, M., and StÃ¼ker, S. (2010). Overview of the iwslt 2010 evaluation campaign. In *IWSLT '10: International Workshop on Spoken Language Translation*, pages 3–27.

Schwartz, L. (2008). Multi-source translation methods. In *Proceedings of AMTA*, pages 279–288.

Seymore, K. and Rosenfeld, R. (1996). Scalable backoff language models. In *Proceedings of ICSLP*, pages 232–235.

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

Tomeh, N., Cancedda, N., and Dymetman, M. (2009). Complexity-based phrase-table filtering for statistical machine translation. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 144–151, Ottawa, Ontario, Canada.

Tomeh, N., Turchi, M., Wisniewski, G., Allauzen, A., and Yvon, F. (2011). How good are your phrases? assessing phrase quality with single class classification. In Hwang, M.-Y. and Stueker, S., editors, *Proceedings of the eigth International Workshop on Spoken Language Translation (IWSLT)*, pages 261–268, San Francisco, CA.

Venugopal, A., Vogel, S., and Waibel, A. (2003). Effective phrase translation extraction from alignment models. In Hinrichs, E. and Roth, D., editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 319–326.

Zens, R., Stanton, D., and Xu, P. (2012). A systematic comparison of phrase table pruning techniques. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 972–983, Jeju Island, Korea.

Zettlemoyer, L. S. and Moore, R. C. (2007). Selective phrase pair extraction for improved 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 on XX*, NAACL '07, pages 209–212, Morristown, NJ, USA. Association for Computational Linguistics.