Translation Model Pruning via Usage Statistics for Statistical Machine Translation

Matthias Eck, Stephan Vogel, and Alex Waibel
InterACT Research
Carnegie Mellon University, Pittsburgh, USA
matteck@cs.cmu.edu, vogel+@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

We describe a new pruning approach to remove phrase pairs from translation models of statistical machine translation systems. The approach applies the original translation system to a large amount of text and calculates usage statistics for the phrase pairs. Using these statistics the relevance of each phrase pair can be estimated. The approach is tested against a strong baseline based on previous work and shows significant improvements.

1 Introduction

A relatively new device for translation systems are small portable devices like cell phones, PDAs and handheld game consoles. The idea here is to have a lightweight and convenient translation device e.g. for tourists that can be easily carried. Other applications include medical, relief, and military scenarios.

Preferably such a device will offer speech-to-speech translation for both (or multiple) translation directions. These devices have been researched and are starting to become commercially available (e.g. Isotani et al., 2003). The main challenges here are the severe restrictions regarding both memory and computing power on such a small portable device.

1.1 Statistical Machine Translation

Generally statistical machine translation systems have recently outperformed other translation approaches so it seems natural to also apply them in these scenarios.

A main component of every statistical machine translation system is the translation model. The translation model assigns translation probabilities to phrase[1] pairs of source and target phrases extracted from a parallel bilingual text. These phrase pairs are applied during the decoding process and their target sides are combined to form the final translation. A variety of algorithms to extract phrase pairs has been proposed. (e.g. Och and Ney, 2000 and Vogel, 2005).

Our proposed approach now tries to remove phrase pairs, which have little influence on the final translation performance, from a translation system (*pruning* of the translation model[2]). The goal is to reduce the number of phrase pairs and in turn the memory requirement of the whole translation system, while not impacting the translation performance too heavily.

The approach does not depend on the actual algorithm used to extract the phrase pairs and can be applied to every imaginable method that assigns probabilities to phrase pairs. We assume that the phrase pairs were pre-extracted before decoding. (in contrast to the proposed approaches to "online phrase extraction" (Zhang and Vogel, 2005; Callison-Burch et al., 2005)).

The task now is to remove enough pre-extracted phrase pairs in order to accommodate the possibly strict memory limitations of a portable device while restricting performance degradation as much as possible.

We will not specifically address the computing power limitations of the portable devices in this paper.

[1] A "phrase" here can also refer to a single word.
[2] Small language models are also desirable and the approaches could be applied as well but this was not investigated yet.

Proceedings of NAACL HLT 2007, Companion Volume, pages 21–24,
Rochester, NY, April 2007.

2 Previous work

Previous work mainly introduced two natural ideas to prune phrase pairs. Both are for example directly available in the Pharaoh decoder (Koehn, 2004).

Probability threshold
A very simple way to prune phrase pairs from a translation model is to use a probability threshold and remove all pairs for which the translation probability is below the threshold. The reasoning for this is that it is very unlikely that a translation with a very low probability will be chosen (over another translation candidate with a higher probability).

Translation variety threshold
Another way to prune phrase pairs is to impose a limit on the number of translation candidates for a certain phrase. That means the pruned translation model can only have equal or fewer possible translations for a given source phrase than the threshold. This is accomplished by sorting the phrase pairs for each source phrase according to their probability and eliminating low probability ones until the threshold is reached.

3 Pruning via Usage Statistics

The approach presented here uses a different idea inspired by the *Optimal Brain Damage* algorithm for neural networks (Le Cun et al., 1990).

The Optimal Brain Damage algorithm for neural networks computes a *saliency* for each network element. The saliency is the relevance for the performance of the network. In each pruning step the element with the smallest saliency is removed, and the network is re-trained and all saliencies are recalculated etc.

We can analogously view each phrase pair in the translation system as such a network element. The question is of course how to calculate the relevance for the performance for each phrase pair.

A simple approximation was already done in the previous work using a probability or variety threshold. Here the relevance is estimated using the phrase pair probability or the phrase pair rank as relevance indicators.

But these are not the only factors that influence the final selection of a phrase pair and most of these factors are not established during the training and phrase extraction process. Especially the following two additional factors play a major role in the importance of a phrase pair.

Frequency of the source phrase
We can clearly say that a phrase pair with a very common source phrase will be much more important than a phrase pair where the source phrase occurs only very rarely.

Actual use of the phrase-pair
But even phrase-pairs with very common source phrases might not be used for the final translation hypothesis. It is for example possible that it is part of a longer phrase pair that gets a higher probability so that the shorter phrase pair is not used.

Generally there are a lot of different factors influencing the estimated importance of a phrase pair and it seems hard to consider every influence separately. Hence the proposed idea does not use a combination of features to estimate the phrase pair importance. Instead the idea is to just apply the translation system to a large amount of text and see how often a phrase pair is actually used (i.e. influences the translation performance). If the translated text is large enough this will give a good statistics of the relevance of this respective phrase pair. This leads to the following algorithm:

Algorithm
Translate a large amount of (in-domain) data with the translation system (tuned on a development set) and collect the following two statistics for each phrase pair in the translation model.

- *c(*phrase pair*)* = Count how often a phrase pair was *considered* during decoding (i.e. was added to the translation lattice)
- *u(*phrase pair*)* = Count how often a phrase pair was *used* in the final translation (i.e. in the chosen path through the lattice).

The overall score for a phrase pair with simple smoothing (+1) is calculated as:

$$score(\text{phrase pair}) = \left[\log(c(\text{phrase pair})+1)\right] * \left[u(\text{phrase pair})+1\right]$$

We use the logarithm function to limit the influence of the *c* value. The *u* value is more important as this measures how often a phrase was actually used in a translation hypothesis. This scoring func-

tion was empirically found after experimenting with a variety of possible scoring terms.

The phrase pairs can then be sorted according to this score and the top *n* phrase pairs can be selected for a smaller phrase translation model.

4 Data and Experiments

4.1 Experimental Setup & Baseline

Translation system
The translation system that was used for the experiments is a state-of-the-art statistical machine translation system (Eck et al. 2006). The system uses a phrase extraction method described in Vogel (2005) and a 6-gram language model.

Training and testing data
The training data for all experiments consisted of the BTEC corpus (Takezawa et al., 2002) with 162,318 lines of parallel Japanese-English text. All translations were done from Japanese to English. The language model was trained on the English part of the training data.

The test set from the evaluation campaign of IWSLT 2004 (Akiba et al., 2004) was used as testing data. This data consists of 500 lines of tourism data. 16 reference translations to English were available.

Extracted phrases
Phrase pairs for n-grams up to length 10 were extracted (with low frequency thresholds for higher n-grams). This gave 4,684,044 phrase pairs (273,459 distinct source phrases). The baseline score using all phrase pairs was 59.11 (BLEU, Papineni et al., 2002) with a 95% confidence interval of [57.13, 61.09].

Baseline pruning
The approaches presented in previous work served as a baseline. The probability threshold was tested for 8 values (0 (no pruning), 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1) while the variety threshold tested for 14 values (1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 50, 100, 200, 500 (no pruning in this case)) and all combinations thereof. The final translation scores for different settings are very fluctuating. For that reason we defined the baseline score for each possible size as the best score that was reached with equal or less phrase pairs than the given size in any of the tested combinations.

4.2 Results for Pruning via Usage Statistics

For the proposed approach "Pruning via Usage Statistics", the translation system was applied to the 162,318 lines of Japanese training data.

As explained in section 3 it was now counted for each phrase pair how often it occurred in a translation lattice and how often it was used for the final translation. The phrase pairs were then sorted according to their relevance estimation and the top *n* phrase pairs were chosen for different values of *n*. The pruned phrase table was then used to translate the IWSLT 2004 test set. Table 1 shows the results comparing the baseline scores with the results using the described pruning. Figure 1 illustrates the scores. The plateaus in the baseline graph are due to the baseline definition as stated above.

	BLEU scores		
# of Phrase Pairs (*n*)	Baseline	Pruning	Relative score improvement
100,000	-	0.4735	-
200,000	0.3162	0.5008	58.38%
300,000	0.4235	0.5154	21.70%
400,000	0.4743	0.5241	10.50%
500,000	0.4743	0.5269	11.09%
600,000	0.4890	0.5359	9.59%
800,000	0.5194	0.5394	3.85%
1,000,000	0.5355	0.5442	1.62%
1,500,000	0.5413	0.5523	2.03%
2,000,000	0.5630	0.5749	2.11%
3,000,000	0.5778	0.5798	0.35%
4,000,000	0.5855	0.5865	0.17%
4,684,044	0.5911	0.5911	0.00%

Table 1: BLEU scores at different levels of pruning (Baseline: Best score with equal or less phrase pairs)

For more than 1 million phrase pairs the differences are not very pronounced. However the translation score for the proposed pruning algorithm is still not significantly lower than the 59.11 score at 2 million phrase pairs while the baseline drops slightly faster. For less than 1 million phrase pairs the differences become much more pronounced with relative improvements of up to 58% at 200,000 phrase pairs. It is interesting to note that the improved pruning removes infrequent source

phrases and to a lesser extent source vocabulary even for larger numbers of phrase pairs.

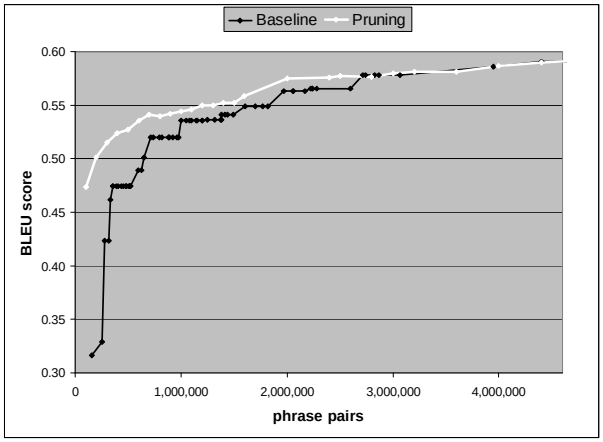

Figure 1: Pruning and baseline comparison

5 Conclusions and Future Work

The proposed pruning algorithm is able to outperform a strong baseline based on previously introduced threshold pruning ideas. Over 50% of phrase pairs can be pruned without a significant loss of performance. Even for very low memory situations the improved pruning remains a viable option while the baseline pruning performance drops heavily.

One idea to improve this new pruning approach is to exchange the *used* count with the count of the phrase occurring in the best path of the lattice according to a scoring metric. This would require having a reference translation available to be able to tell which path is the actual best one (metric-best path). It would be interesting to compare the performance if the statistics is done using the metric-best path on a smaller amount of data to the performance if the statistics is done using the model-best path on a larger amount (as there is no reference translation necessary).

The Optimal Brain Damage algorithm recalculates the *saliency* after removing each network element. It could also be beneficial to sequentially prune the phrase pairs and always re-calculate the statistics after removing a certain number of phrase pairs.

6 Acknowledgements

This work was partly supported by the US DARPA under the programs GALE and TRANSTAC.

7 References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii}. 2004. *Overview of the IWSLT04 Evaluation Campaign*. Proceedings of IWSLT 2004, Kyoto, Japan.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. *Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases.* Proceedings of ACL 2005, Ann Arbor, MI, USA.

Yann Le Cun, John S. Denker, and Sara A. Solla. 1990. *Optimal brain damage*. In Advances in Neural Information Processing Systems 2, pages 598-605. Morgan Kaufmann, 1990.

Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, Bing Zhao, Almut Silja Hildebrand, Stephan Vogel, and Alex Waibel. 2006. *The UKA/CMU Statistical Machine Translation System for IWSLT 2006.* Proceedings of IWSLT 2006, Kyoto, Japan.

Ryosuke Isotani, Kyoshi Yamabana, Shinichi Ando, Ken Hanazawa, Shin-ya Ishikawa and Ken.ichi Iso. 2003. *Speech-to-speech translation software on PDAs for travel conversation*. NEC research & development, Tokyo, Japan.

Philipp Koehn. 2004. *A Beam Search Decoder for Statistical Machine Translation Models.* Proceedings of AMTA 2004, Baltimore, MD, USA.

Franz Josef Och and Hermann Ney, 2000. *Improved statistical alignment models*, Proceedings of ACL 2000, Hongkong, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* Proceedings of ACL 2002, Philadelphia, PA, USA.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. *Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World.* Proceedings of LREC 2002, Las Palmas, Spain.

Stephan Vogel. 2005. *PESA: Phrase Pair Extraction as Sentence Splitting.* Proceedings of MTSummit X, Phuket, Thailand.

Ying Zhang and Stephan Vogel. 2005. *An Efficient Phrase-to-Phrase Alignment Model for Arbitrarily Long Phrases and Large Corpora*. Proceedings of EAMT 2005, Budapest, Hungary.