Translation Model Size Reduction for Hierarchical Phrase-based Statistical Machine Translation

Seung-Wook Lee[†] **Dongdong Zhang**[‡] **Mu Li**[‡] **Ming Zhou**[‡] **Hae-Chang Rim**[†]
[†] Dept. of Computer & Radio Comms. Engineering, Korea University, Seoul, South Korea
{swlee,rim}@nlp.korea.ac.kr
[‡] Microsoft Research Asia, Beijing, China
{dozhang,muli,mingzhou}@microsoft.com

Abstract

In this paper, we propose a novel method of reducing the size of translation model for hierarchical phrase-based machine translation systems. Previous approaches try to prune infrequent entries or unreliable entries based on statistics, but cause a problem of reducing the translation coverage. On the contrary, the proposed method try to prune only ineffective entries based on the estimation of the information redundancy encoded in phrase pairs and hierarchical rules, and thus preserve the search space of SMT decoders as much as possible. Experimental results on Chinese-to-English machine translation tasks show that our method is able to reduce almost the half size of the translation model with very tiny degradation of translation performance.

1 Introduction

Statistical Machine Translation (SMT) has gained considerable attention during last decades. From a bilingual corpus, all translation knowledge can be acquired automatically in SMT framework. Phrase-based model (Koehn et al., 2003) and hierarchical phrase-based model (Chiang, 2005; Chiang, 2007) show state-of-the-art performance in various language pairs. This achievement is mainly benefit from huge size of translational knowledge extracted from sufficient parallel corpus. However, the errors of automatic word alignment and non-parallelized bilingual sentence pairs sometimes have caused the unreliable and unnecessary translation rule acquisition. According to Bloodgood and Callison-Burch (2010) and our own preliminary experiments, the size of phrase table and hierarchical rule table consistently increases linearly with the growth of training size, while the translation performance tends to gain minor improvement after a certain point. Consequently, the model size reduction is necessary and meaningful for SMT systems if it can be performed without significant performance degradation. The smaller the model size is, the faster the SMT decoding speed is, because there are fewer hypotheses to be investigated during decoding. Especially, in a limited environment, such as mobile device, and for a time-urgent task, such as speech-to-speech translation, the compact size of translation rules is required. In this case, the model reduction would be the one of the main techniques we have to consider.

Previous methods of reducing the size of SMT model try to identify infrequent entries (Zollmann et al., 2008; Huang and Xiang, 2010). Several statistical significance testing methods are also examined to detect unreliable noisy entries (Tomeh et al., 2009; Johnson et al., 2007; Yang and Zheng, 2009). These methods could harm the translation performance due to their side effect of algorithms; similar multiple entries can be pruned at the same time deteriorating potential coverage of translation. The proposed method, on the other hand, tries to measure the redundancy of phrase pairs and hierarchical rules. In this work, redundancy of an entry is defined as its translational ineffectiveness, and estimated by comparing scores of entries and scores of their substituents. Suppose that the source phrase $s_1 s_2$ is always translated into $t_1 t_2$ with phrase entry $<s_1 s_2 \rightarrow t_1 t_2>$ where s_i and t_i are correspond-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 291–295,
Jeju, Republic of Korea, 8-14 July 2012.

ing translations. Similarly, source phrases s_1 and s_2 are always translated into t_1 and t_2, with phrase entries, $<s_1 \rightarrow t_1>$ and $<s_2 \rightarrow t_2>$, respectively. In this case, it is intuitive that $<s_1 s_2 \rightarrow t_1 t_2>$ could be unnecessary and redundant since its substituent always produces the same result. This paper presents statistical analysis of this redundancy measurement. The redundancy-based reduction can be performed to prune the phrase table, the hierarchical rule table, and both. Since the similar translation knowledge is accumulated at both of tables during the training stage, our reduction method performs effectively and safely. Unlike previous studies solely focus on either phrase table or hierarchical rule table, this work is the first attempt to reduce phrases and hierarchical rules simultaneously.

2 Proposed Model

Given an original translation model, TM, our goal is to find the optimally reduced translation model, TM^*, which minimizes the degradation of translation performance. To measure the performance degradation, we introduce a new metric named *consistency*:

$$C(TM, TM^*) = BLEU(D(s; TM), D(s; TM^*)) \quad (1)$$

where the function D produces the target sentence of the source sentence s, given the translation model TM. *Consistency* measures the similarity between the two groups of decoded target sentences produced by two different translation models. There are number of similarity metrics such as Dices coefficient (Kondrak et al., 2003), and Jaccard similarity coefficient. Instead, we use BLEU scores (Papineni et al., 2002) since it is one of the primary metrics for machine translation evaluation. Note that our *consistency* does not require the reference set while the original BLEU does. This means that only (abundant) source-side monolingual corpus is needed to predict performance degradation. Now, our goal can be rewritten with this metric; among all the possible reduced models, we want to find the set which can maximize the *consistency*:

$$TM^* = \underset{TM' \subset TM}{argmax}\ C(TM, TM') \quad (2)$$

In minimum error rate training (MERT) stages, a development set, which consists of bilingual sentences, is used to find out the best weights of features (Och, 2003). One characteristic of our method is that it isolates feature weights of the translation model from SMT log-linear model, trying to minimize the impact of search path during decoding. The reduction procedure consists of three stages: translation scoring, redundancy estimation, and redundancy-based reduction.

Our reduction method starts with measuring the translation scores of the individual phrase and the hierarchical rule. Similar to the decoder, the scoring scheme is based on the log-linear framework:

$$PS(p) = \sum_i \lambda_i h_i(p) \quad (3)$$

where h is a feature function and λ is its weight. As the conventional hierarchical phrase-based SMT model, our features are composed of $P(e|f)$, $P(f|e)$, $P_{lex}(e|f)$, $P_{lex}(f|e)$, and the number of phrases, where e and f denote a source phrase and a target phrase, respectively. P_{lex} is the lexicalized probability. In a similar manner, the translation scores of hierarchical rules are calculated as follows:

$$HS(r) = \sum_i \lambda_i h_i(r) \quad (4)$$

The features are as same as those that are used for phrase scoring, except the last feature. Instead of the phrase number penalty, the hierarchical rule number penalty is used. The weight for each feature is shared from the results of MERT. With this scoring scheme, our model is able to measure how important the individual entry is during decoding.

Once translation scores for all entries are estimated, our method retrieves substituent candidates with their combination scores. The combination score is calculated by accumulating translation scores of every member as follows:

$$CS(p_{1...n}) = \sum_{i=1}^{n} PS(p_i) \quad (5)$$

This scoring scheme follows the same manner what the conventional decoder does, finding the best phrase combination during translation. By comparing the original translation score with combination

scores of its substituents, the redundancy scores are estimated, as follows:

$$Red(p) = \min_{p_{1...n} \in Sub(p)} PS(p) - CS(p_{1...n}) \quad (6)$$

where Sub is the function that retrieves all possible substituents (the combinations of sub-phrases, and/or sub-rules that exactly produce the same target phrase, given the source phrase p). If the combination score of the best substituent is same as the translation score of p, the redundancy score becomes zero. In this case, the decoder always produces the same translation results without p. When the redundancy score is negative, the best substituent is more likely to be chosen instead of p. This implies that there is no risk to prune p; the search space is not changed, and the search path is not changed as well.

Our method can be varied according to the designation of Sub function. If both of the phrase table and the hierarchical rule table are allowed, cross reduction can be possible; the phrase table is reduced based on the hierarchical rule table and vice versa. With extensions of combination scoring and redundancy scoring schemes like following equations, our model is able to perform cross reduction.

$$CS(p_{1...n}, h_{1...m}) = \sum_{i=1}^{n} PS(p_i) + \sum_{i=1}^{m} HS(h_i) \quad (7)$$

$$Red(p) = \min_{<p_{1...n}, h_{1...m}> \in Sub(p)} PS(p) - CS(p_{1...n}, h_{1...m}) \quad (8)$$

The proposed method has some restrictions for reduction. First of all, it does not try to prune the phrase that has no substituents, such as unigram phrases; the phrase whose source part is composed of a single word. This restriction guarantees that the translational coverage of the reduced model is as high as those of the original translation model. In addition, our model does not prune the phrases and the hierarchical rules that have reordering within it to prevent information loss of reordering. For instance, if we prune phrase, $<s_1 s_2 s_3 \rightarrow t_3 t_1 t_2>$, phrases, $<s_1 s_2 \rightarrow t_1 t_2>$ and $<s_3 \rightarrow t_3>$ are not able to produce the same target words without appropriate reordering.

Once the redundancy scores for all entries have been estimated, the next step is to select the best N entries to prune to satisfy a desired model size. We can simply prune the first N from the list of entries sorted by increasing order of redundancy score. However, this method may not result in the optimal reduction, since each redundancy scores are estimated based on the assumption of the existence of all the other entries. In other words, there are dependency relationships among entries. We examine two methods to deal with this problem. The first is to ignore dependency, which is the more efficient manner. The other is to prune independent entries first. After all independent entries are pruned, the dependent entries are started to be pruned. We present the effectiveness of each method in the next section.

Since our goal is to reduce the size of all translation models, the reduction is needed to be performed for both the phrase table and the hierarchical rule table simultaneously, namely joint reduction. Similar to phrase reduction and hierarchical rule reduction, it selects the best N entries of the mixture of phrase and hierarchical rules. This method results in safer pruning; once a phrase is determined to be pruned, the hierarchical rules, which are related to this phrase, are likely to be kept, and vice versa.

3 Experiment

We investigate the effectiveness of our reduction method by conducting Chinese-to-English translation task. The training data, as same as Cui et al. (2010), consists of about 500K parallel sentence pairs which is a mixture of several datasets published by LDC. NIST 2003 set is used as a development set. NIST 2004, 2005, 2006, and 2008 sets are used for evaluation purpose. For word alignment, we use GIZA++[1], an implementation of IBM models (Brown et al., 1993). We have implemented a hierarchical phrase-based SMT model similar to Chiang (2005). The trigram target language model is trained from the Xinhua portion of English Gigaword corpus (Graff and Cieri, 2003). Sampled 10,000 sentences from Chinese Gigaword corpus (Graff, 2007) was used for source-side development dataset to measure consistency. Our main metric for translation performance evaluation is case-

[1] http://www.statmt.org/moses/giza/GIZA++.html

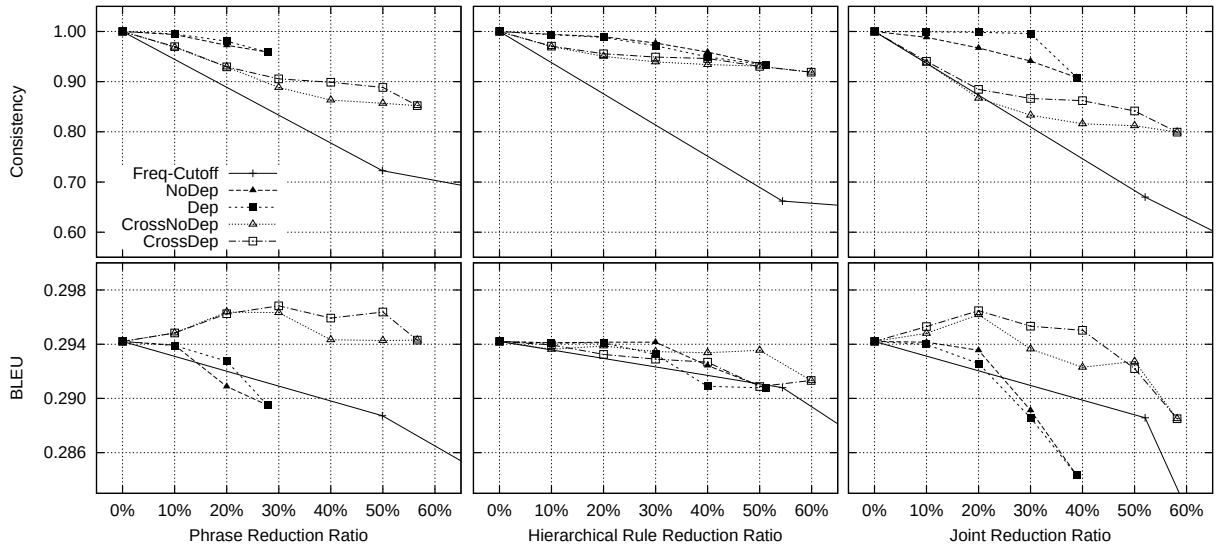

Figure 1: Performance comparison. BLEU scores and consistency scores are averaged over four evaluation sets.

insensitive BLEU-4 scores (Papineni et al., 2002).

As a baseline system, we chose the frequency-based cutoff method, which is one of the most widely used filtering methods. As shown in Figure 1, almost half of the phrases and hierarchical rules are pruned when cutoff=2, while the BLEU score is also deteriorated significantly. We introduced two methods for selecting the N pruning entries considering dependency relationships. The non-dependency method does not consider dependency relationships, while the dependency method prunes independent entries first. Each method can be combined with cross reduction. The performance is measured in three different reduction tasks: phrase reduction, hierarchical rule reduction, and joint reduction. As the reduction ratio becomes higher, the model size, i.e., the number of entries, is reduced while BLEU scores and coverage are decreased. The results show that the translation performance is highly co-related with the *consistency*. The co-relation scores measured between them on the phrase reduction and the hierarchical rule reduction tasks are 0.99 and 0.95, respectively, which indicates very strong positive relationship.

For the phrase reduction task, the dependency method outperforms the non-dependency method in terms of BLEU score. When the cross reduction technique was used for the phrase reduction task, BLEU score is not deteriorated even when more than half of phrase entries are pruned. This result implies that there is much redundant information stored in the hierarchical rule table. On the other hand, for the hierarchical rule reduction task, the non-dependency method shows the better performance. The dependency method sometimes performs worse than the baseline method. We expect that this is caused by the unreliable estimation of dependency among hierarchical rules since the most of them are automatically generated from the phrases. The excessive dependency of these rules would cause overestimation of hierarchical rule redundancy score.

4 Conclusion

We present a novel method of reducing the size of translation model for SMT. The contributions of the proposed method are as follows: 1) our method is the first attempt to reduce the phrase table and the hierarchical rule table simultaneously. 2) our method is a safe reduction method since it considers the redundancy, which is the practical ineffectiveness of individual entry. 3) our method shows that almost the half size of the translation model can be reduced without significant performance degradation. It may be appropriate for the applications running on limited environment, e.g., mobile devices.

Acknowledgement

The first author performed this research during an internship at Microsoft Research Asia. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA(National IT Industry Promotion Agency). (NIPA-2010-C1810-1002-0025)

References

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19:263–311, June.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43th Annual Meeting on Association for Computational Linguistics*, pages 263–270.

David Chiang. 2007. Hierarchical Phrase-based Translation. *Computational Linguistics*, 33:201–228, June.

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou, and Tiejun Zhao. 2010. Hybrid Decoding: Decoding with Partial Hypotheses Combination Over Multiple SMT Systems. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 214–222, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Graff and Christopher Cieri. 2003. English Gigaword. In *Linguistic Data Consortium, Philadelphia*.

David Graff. 2007. Chinese Gigaword Third Edition. In *Linguistic Data Consortium, Philadelphia*.

Fei Huang and Bing Xiang. 2010. Feature-Rich Discriminative Phrase Rescoring for SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 492–500, Stroudsburg, PA, USA. 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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can Improve Statistical Translation Models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 46–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Nadi Tomeh, Nicola Cancedda, and Marc Dymetman. 2009. Complexity-based Phrase-Table Filtering for Statistical Machine Translation.

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*, ACLShort '09, pages 237–240, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Zollmann, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A Systematic Comparison of Phrase-based, Hierarchical and Syntax-Augmented Statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1145–1152, troudsburg, PA, USA. Association for Computational Linguistics.