Better Filtration and Augmentation for Hierarchical Phrase-Based Translation Rules

Zhiyang Wang † **Yajuan Lü** † **Qun Liu** † **Young-Sook Hwang** ‡

†Key Lab. of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
wangzhiyang@ict.ac.cn

‡HILab Convergence Technology Center
C&I Business
SKTelecom
11, Euljiro2-ga, Jung-gu, Seoul 100-999, Korea
yshwang@sktelecom.com

Abstract

This paper presents a novel filtration criterion to restrict the rule extraction for the hierarchical phrase-based translation model, where a bilingual but relaxed well-formed dependency restriction is used to filter out bad rules. Furthermore, a new feature which describes the regularity that the source/target dependency edge triggers the target/source word is also proposed. Experimental results show that, the new criteria weeds out about 40% rules while with translation performance improvement, and the new feature brings another improvement to the baseline system, especially on larger corpus.

1 Introduction

Hierarchical phrase-based (HPB) model (Chiang, 2005) is the state-of-the-art statistical machine translation (SMT) model. By looking for phrases that contain other phrases and replacing the subphrases with nonterminal symbols, it gets hierarchical rules. Hierarchical rules are more powerful than conventional phrases since they have better generalization capability and could capture long distance reordering. However, when the training corpus becomes larger, the number of rules will grow exponentially, which inevitably results in slow and memory-consuming decoding.

In this paper, we address the problem of reducing the hierarchical translation rule table resorting to the dependency information of bilingual languages. We only keep rules that both sides are *relaxed-well-formed* (RWF) dependency structure (see the definition in Section 3), and discard others which do not satisfy this constraint. In this way, about 40% bad rules are weeded out from the original rule table. However, the performance is even better than the traditional HPB translation system.

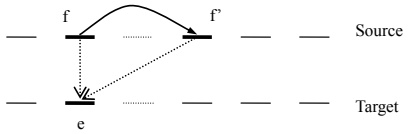

Figure 1: Solid wire reveals the dependency relation pointing from the child to the parent. Target word e is triggered by the source word f and it's head word f', $p(e|f \rightarrow f')$.

Based on the *relaxed-well-formed* dependency structure, we also introduce a new linguistic feature to enhance translation performance. In the traditional phrase-based SMT model, there are always lexical translation probabilities based on IBM model 1 (Brown et al., 1993), i.e. $p(e|f)$, namely, the target word e is triggered by the source word f. Intuitively, however, the generation of e is not only involved with f, sometimes may also be triggered by other context words in the source side. Here we assume that the dependency edge $(f \rightarrow f')$ of word f generates target word e (we call it head word trigger in Section 4). Therefore, two words in one language trigger one word in another, which provides a more sophisticated and better choice for the target word, i.e. Figure 1. Similarly, the dependency feature works well in Chinese-to-English translation task, especially on large corpus.

2 Related Work

In the past, a significant number of techniques have been presented to reduce the hierarchical rule table. He et al. (2009) just used the key phrases of source side to filter the rule table without taking advantage of any linguistic information. Iglesias et al. (2009) put rules into syntactic classes based on the number of non-terminals and patterns, and applied various filtration strategies to improve the rule table quality. Shen et al. (2008) discarded

Proceedings of the ACL 2010 Conference Short Papers, pages 142–146,
Uppsala, Sweden, 11-16 July 2010.

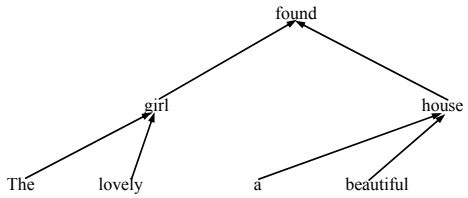

Figure 2: An example of dependency tree. The corresponding plain sentence is *The lovely girl found a beautiful house*.

most entries of the rule table by using the constraint that rules of the target-side are well-formed (WF) dependency structure, but this filtering led to degradation in translation performance. They obtained improvements by adding an additional dependency language model. The basic difference of our method from (Shen et al., 2008) is that we keep rules that both sides should be *relaxed-well-formed* dependency structure, not just the target side. Besides, our system complexity is not increased because no additional language model is introduced.

The feature of head word trigger which we apply to the log-linear model is motivated by the trigger-based approach (Hasan and Ney, 2009). Hasan and Ney (2009) introduced a second word to trigger the target word without considering any linguistic information. Furthermore, since the second word can come from any part of the sentence, there may be a prohibitively large number of parameters involved. Besides, He et al. (2008) built a maximum entropy model which combines rich context information for selecting translation rules during decoding. However, as the size of the corpus increases, the maximum entropy model will become larger. Similarly, In (Shen et al., 2009), context language model is proposed for better rule selection. Taking the dependency edge as condition, our approach is very different from previous approaches of exploring context information.

3 Relaxed-well-formed Dependency Structure

Dependency models have recently gained considerable interest in SMT (Ding and Palmer, 2005; Quirk et al., 2005; Shen et al., 2008). Dependency tree can represent richer structural information. It reveals long-distance relation between words and directly models the semantic structure of a sentence without any constituent labels. Figure 2 shows an example of a dependency tree. In this example, the word *found* is the root of the tree.

Shen et al. (2008) propose the well-formed dependency structure to filter the hierarchical rule table. A well-formed dependency structure could be either a single-rooted dependency tree or a set of sibling trees. Although most rules are discarded with the constraint that the target side should be well-formed, this filtration leads to degradation in translation performance.

As an extension of the work of (Shen et al., 2008), we introduce the so-called *relaxed-well-formed* dependency structure to filter the hierarchical rule table. Given a sentence $S = w_1 w_2 ... w_n$. Let $d_1 d_2 ... d_n$ represent the position of parent word for each word. For example, $d_3 = 4$ means that w_3 depends on w_4. If w_i is a root, we define $d_i = -1$.

Definition A dependency structure $w_i ... w_j$ is a *relaxed-well-formed* structure, where there is $h \notin [i, j]$, all the words $w_i ... w_j$ are directly or indirectly depended on w_h or -1 (here we define $h = -1$). If and only if it satisfies the following conditions

- $d_h \notin [i, j]$
- $\forall k \in [i, j],\ d_k \in [i, j]$ *or* $d_k = h$

From the definition above, we can see that the *relaxed-well-formed* structure obviously covers the well-formed one. In this structure, we don't constrain that all the children of the sub-root should be complete. Let's review the dependency tree in Figure 2 as an example. Except for the well-formed structure, we could also extract *girl found a beautiful house*. Therefore, if the modifier *The lovely* changes to *The cute*, this rule also works.

4 Head Word Trigger

(Koehn et al., 2003) introduced the concept of lexical weighting to check how well words of the phrase translate to each other. Source word f aligns with target word e, according to the IBM model 1, the lexical translation probability is $p(e|f)$. However, in the sense of dependency relationship, we believe that the generation of the target word e, is not only triggered by the aligned source word f, but also associated with f 's head word f'. Therefore, the lexical translation probability becomes $p(e|f \rightarrow f')$, which of course allows for a more fine-grained lexical choice of

the target word. More specifically, the probability could be estimated by the maximum likelihood (MLE) approach,

$$p(e|f \rightarrow f') = \frac{count(e, f \rightarrow f')}{\sum_{e'} count(e', f \rightarrow f')} \quad (1)$$

Given a phrase pair $\overline{f}$, $\overline{e}$ and word alignment a, and the dependent relation of the source sentence d_1^J (J is the length of the source sentence, I is the length of the target sentence). Therefore, given the lexical translation probability distribution $p(e|f \rightarrow f')$, we compute the feature score of a phrase pair ($\overline{f}$, $\overline{e}$) as

$$p(\overline{e}|\overline{f}, d_1^J, a) = \Pi_{i=1}^{|\overline{e}|} \frac{1}{|\{j|(j,i) \in a\}|} \sum_{\forall (j,i) \in a} p(e_i|f_j \rightarrow f_{d_j}) \quad (2)$$

Now we get $p(\overline{e}|\overline{f}, d_1^J, a)$, we could obtain $p(\overline{f}|\overline{e}, d_1^I, a)$ (d_1^I represents dependent relation of the target side) in the similar way. This new feature can be easily integrated into the log-linear model as lexical weighting does.

5 Experiments

In this section, we describe the experimental setting used in this work, and verify the effect of the *relaxed-well-formed* structure filtering and the new feature, head word trigger.

5.1 Experimental Setup

Experiments are carried out on the NIST[1] Chinese-English translation task with two different size of training corpora.

- **FBIS**: We use the FBIS corpus as the first training corpus, which contains 239K sentence pairs with 6.9M Chinese words and 8.9M English words.
- **GQ**: This is manually selected from the LDC[2] corpora. GQ contains 1.5M sentence pairs with 41M Chinese words and 48M English words. In fact, FBIS is the subset of GQ.

[1]www.nist.gov/speech/tests/mt

[2]It consists of six LDC corpora: LDC2002E18, LDC2003E07, LDC2003E14, Hansards part of LDC2004T07, LDC2004T08, LDC2005T06.

For language model, we use the SRI Language Modeling Toolkit (Stolcke, 2002) to train a 4-gram model on the first 1/3 of the Xinhua portion of GIGAWORD corpus. And we use the NIST 2002 MT evaluation test set as our development set, and NIST 2004, 2005 test sets as our blind test sets. We evaluate the translation quality using *case-insensitive* BLEU metric (Papineni et al., 2002) without dropping OOV words, and the feature weights are tuned by minimum error rate training (Och, 2003).

In order to get the dependency relation of the training corpus, we re-implement a beam-search style monolingual dependency parser according to (Nivre and Scholz, 2004). Then we use the same method suggested in (Chiang, 2005) to extract SCFG grammar rules within dependency constraint on both sides except that unaligned words are allowed at the edge of phrases. Parameters of head word trigger are estimated as described in Section 4. As a default, the maximum initial phrase length is set to 10 and the maximum rule length of the source side is set to 5. Besides, we also re-implement the decoder of Hiero (Chiang, 2007) as our baseline. In fact, we just exploit the dependency structure during the rule extraction phase. Therefore, we don't need to change the main decoding algorithm of the SMT system.

5.2 Results on FBIS Corpus

A series of experiments was done on the FBIS corpus. We first parse the bilingual languages with monolingual dependency parser respectively, and then only retain the rules that both sides are in line with the constraint of dependency structure. In Table 1, the *relaxed-well-formed* structure filtered out 35% of the rule table and the well-formed discarded 74%. *RWF* extracts additional 39% compared to *WF*, which can be seen as some kind of evidence that the rules we additional get seem common in the sense of linguistics. Compared to (Shen et al., 2008), we just use the dependency structure to constrain rules, not to maintain the tree structures to guide decoding.

Table 2 shows the translation result on FBIS. We can see that the *RWF* structure constraint can improve translation quality substantially both at development set and different test sets. On the Test04 task, it gains +0.86% BLEU, and +0.84% on Test05. Besides, we also used Shen et al. (2008)'s *WF* structure to filter both sides. Although it discard about 74% of the rule table, the

System	Rule table size
HPB	30,152,090
RWF	19,610,255
WF	7,742,031

Table 1: Rule table size with different constraint on FBIS. Here *HPB* refers to the baseline hierarchal phrase-based system, *RWF* means *relaxed-well-formed* constraint and *WF* represents the well-formed structure.

System	Dev02	Test04	Test05
HPB	0.3285	0.3284	0.2965
WF	0.3125	0.3218	0.2887
RWF	0.3326	**0.3370****	**0.3050**
RWF+Tri	0.3281	/	0.2965

Table 2: Results of FBIS corpus. Here *Tri* means the feature of head word trigger on both sides. And we don't test the new feature on Test04 because of the bad performance on development set. * or ** = significantly better than baseline ($p < 0.05$ or 0.01, respectively).

over-all BLEU is decreased by 0.66%-0.78% on the test sets.

As for the feature of head word trigger, it seems not work on the FBIS corpus. On Test05, it gets the same score with the baseline, but lower than *RWF* filtering. This may be caused by the data sparseness problem, which results in inaccurate parameter estimation of the new feature.

5.3 Result on GQ Corpus

In this part, we increased the size of the training corpus to check whether the feature of head word trigger works on large corpus.

We get 152M rule entries from the GQ corpus according to (Chiang, 2007)'s extraction method. If we use the *RWF* structure to constrain both sides, the number of rules is 87M, about 43% of rule entries are discarded. From Table 3, the new feature works well on two different test sets. The gain is +2.21% BLEU on Test04, and +1.33% on Test05. Compared to the result of the baseline, only using the *RWF* structure to filter performs the same as the baseline on Test05, and +0.99% gains on Test04.

System	Dev02	Test04	Test05
HPB	0.3473	0.3386	0.3206
RWF	0.3539	**0.3485****	**0.3228**
RWF+Tri	0.3540	**0.3607****	**0.3339***

Table 3: Results of GQ corpus. * or ** = significantly better than baseline ($p < 0.05$ or 0.01, respectively).

6 Conclusions

This paper proposes a simple strategy to filter the hierarchal rule table, and introduces a new feature to enhance the translation performance. We employ the *relaxed-well-formed* dependency structure to constrain both sides of the rule, and about 40% of rules are discarded with improvement of the translation performance. In order to make full use of the dependency information, we assume that the target word e is triggered by dependency edge of the corresponding source word f. And this feature works well on large parallel training corpus.

How to estimate the probability of head word trigger is very important. Here we only get the parameters in a generative way. In the future, we we are plan to exploit some discriminative approach to train parameters of this feature, such as EM algorithm (Hasan et al., 2008) or maximum entropy (He et al., 2008).

Besides, the quality of the parser is another effect for this method. As the next step, we will try to exploit bilingual knowledge to improve the monolingual parser, i.e. (Huang et al., 2009).

Acknowledgments

This work was partly supported by National Natural Science Foundation of China Contract 60873167. It was also funded by SK Telecom, Korea under the contract 4360002953. We show our special thanks to Wenbin Jiang and Shu Cai for their valuable suggestions. We also thank the anonymous reviewers for their insightful comments.

References

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

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL*

'05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228.

Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 541–548.

Saša Hasan and Hermann Ney. 2009. Comparison of extended lexicon models in search and rescoring for smt. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 17–20.

Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Jesús Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 372–381.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 321–328.

Zhongjun He, Yao Meng, Yajuan Lü, Hao Yu, and Qun Liu. 2009. Reducing smt rule table with monolingual key phrase. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 121–124.

Liang Huang, Wenbin Jiang, and Qun Liu. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1222–1231.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009. Rule filtering by pattern for efficient hierarchical translation. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 380–388.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Joakim Nivre and Mario Scholz. 2004. Deterministic dependency parsing of english text. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 64–70.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 72–80.

Andreas Stolcke. 2002. Srilman extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.