Reducing SMT Rule Table with Monolingual Key Phrase

Zhongjun He[†] **Yao Meng**[†] **Yajuan Lü**[‡] **Hao Yu**[†] **Qun Liu**[‡]
[†] Fujitsu R&D Center CO., LTD, Beijing, China
{hezhongjun, mengyao, yu}@cn.fujitsu.com
[‡] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{lvyajuan, liuqun}@ict.ac.cn

Abstract

This paper presents an effective approach to discard most entries of the rule table for statistical machine translation. The rule table is filtered by monolingual *key phrases*, which are extracted from source text using a technique based on term extraction. Experiments show that 78% of the rule table is reduced without worsening translation performance. In most cases, our approach results in measurable improvements in BLEU score.

1 Introduction

In statistical machine translation (SMT) community, the state-of-the-art method is to use rules that contain hierarchical structures to model translation, such as the hierarchical phrase-based model (Chiang, 2005). Rules are more powerful than conventional phrase pairs because they contain structural information for capturing long distance reorderings. However, hierarchical translation systems often suffer from a large rule table (the collection of rules), which makes decoding slow and memory-consuming.

In the training procedure of SMT systems, numerous rules are extracted from the bilingual corpus. During decoding, however, many of them are rarely used. One of the reasons is that these rules have low quality. The rule quality are usually evaluated by the conditional translation probabilities, which focus on the correspondence between the source and target phrases, while ignore the quality of phrases in a monolingual corpus.

In this paper, we address the problem of reducing the rule table with the information of monolingual corpus. We use $C\text{-}value$, a measurement of automatic term recognition, to score source phrases. A source phrase is regarded as a *key phrase* if its score greater than a threshold. Note that a source phrase is either a *flat* phrase consists of words, or a *hierarchical* phrase consists of both words and variables. For rule table reduction, the rule whose source-side is not key phrase is discarded.

Our approach is different from the previous research. Johnson et al. (2007) reduced the phrase table based on the significance testing of phrase pair co-occurrence in bilingual corpus. The basic difference is that they used statistical information of bilingual corpus while we use that of monolingual corpus. Shen et al. (2008) proposed a string-to-dependency model, which restricted the target-side of a rule by dependency structures. Their approach greatly reduced the rule table, however, caused a slight decrease of translation quality. They obtained improvements by incorporating an additional dependency language model. Different from their research, we restrict rules on the source-side. Furthermore, the system complexity is not increased because no additional model is introduced.

The hierarchical phrase-based model (Chiang, 2005) is used to build a translation system. Experiments show that our approach discards 78% of the rule table without worsening the translation quality.

2 Monolingual Phrase Scoring

2.1 Frequency

The basic metrics for phrase scoring is the frequency that a phrase appears in a monolingual corpus. The more frequent a source phrase appears in a corpus, the greater possibility the rule that contains the source phrase may be used.

However, one limitation of this metrics is that if we filter the rule table by the source phrase with lower frequency, most long phrase pairs will be discarded. Because the longer the phrase is, the less possibility it appears. However, long phrases

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 121–124,
Suntec, Singapore, 4 August 2009.

are very helpful for reducing ambiguity since they contains more information than short phrases.

Another limitation is that the frequency metrics focuses on a phrase appearing by itself while ignores it appears as a substring of longer phrases. It is therefore inadequate for hierarchical phrases.

We use an example for illustration. Considering the following three rules (the subscripts indicate word alignments):

R_1 :
接受$_1$ 布什$_2$ 总统$_3$ 的$_4$ 邀请$_5$
accept$_1$ President$_3$ Bush$_2$'s$_4$ invitation$_5$

R_2 :
接受$_1$ 布什$_2$ X$_3$ 的$_4$ 邀请$_5$
accept$_1$ X$_3$ Bush$_2$'s$_4$ invitation$_5$

R_3 :
接受$_1$ X$_2$ 的$_3$ 邀请$_4$
accept$_1$ X$_2$'s$_3$ invitation$_4$

We use f_1, f_2 and f_3 to represent their source-sides, respectively. The hierarchical phrases f_2 and f_3 are sub-strings of f_1. However, R_3 is suggested to be more useful than R_2. The reason is that f_3 may appears in various phrases, such as "接受 法国 的 邀请, accept France 's invitation". While f_2 almost always appears in f_1, indicating that the variable X may not be replaced with other words expect "President". It indicates that R_2 is not likely to be useful, although f_2 may appears frequently in a corpus.

2.2 C-value

$C\text{-}value$, a measurement of automatic term recognition, is proposed by Frantzi and Ananiadou (1996) to extract nested collocations, collocations that substrings of other longer ones.

We use $C\text{-}value$ for two reasons: on one hand, it uses rich factors besides phrase frequency, e.g. the phrase length, the frequency that a sub-phrase appears in longer phrases. Thus it is appropriate for extracting hierarchical phrases. On the other hand, the computation of $C\text{-}value$ is efficient.

Analogous to (Frantzi and Ananiadou, 1996), we use 4 factors (L, F, S, N) to determine if a phrase p is a key phrase:

1. $L(p)$, the length of p;
2. $F(p)$, the frequency that p appears in a corpus;
3. $S(p)$, the frequency that p appears as a sub-string in other longer phrases;
4. $N(p)$, the number of phrases that contain p as a substring.

Algorithm 1 Key Phrase Extraction

Input: Monolingual Text
Output: Key Phrase Table *KP*

1: Extract candidate phrases
2: **for** all phrases p in length descending order **do**
3: **if** $N(p) = 0$ **then**
4: $C\text{-}value = (L(p) - 1) \times F(p)$
5: **else**
6: $C\text{-}value = (L(p) - 1) \times (F(p) - \frac{S(p)}{N(p)})$
7: **end if**
8: **if** $C\text{-}value \geq \varepsilon$ **then**
9: add p to *KP*
10: **end if**
11: **for** all sub-strings q of p **do**
12: $S(q) = S(q) + F(p) - S(p)$
13: $N(q) = N(q) + 1$
14: **end for**
15: **end for**

Given a monolingual corpus, key phrases can be extracted efficiently according to Algorithm 1.

Firstly (line 1), all possible phrases are extracted as candidates of key phrases. This step is analogous to the rule extraction as described in (Chiang, 2005). The basic difference is that there are no word alignment constraints for monolingual phrase extraction, which therefore results in a substantial number of candidate phrases. We use the following restrictions to limit the phrase number:

1. The length of a candidate phrase is limited to pl;
2. The length of the initial phrase used to create hierarchical phrases is limited to ipl;
3. The number of variables in hierarchical phrases is limited to nv, and there should be at least 1 word between variables;
4. The frequency of a candidate phrase appears in a corpus should be greater than $freq$.

In our experiments, we set $pl = 5$, $ipl = 10$, $nv = 2$, $freq = 3$. Note that the first 3 settings are used in (Chiang, 2005) for rule extraction.

Secondly (line 3 to 7), for each candidate phrase, $C\text{-}value$ is computed according to the phrase appears by itself (line 4) or as a substring of other long phrases (line 6). The $C\text{-}value$ is in direct proportion to the phrase length (L) and occurrences (F, S), while in inverse proportion to the number of phrases that contain the phrase as a substring (N). This overcomes the limitations of frequency measurement. A phrase is regarded as a key phrase if its $C\text{-}value$ is greater than a threshold ε.

Finally (line 11 to 14), $S(q)$ and $N(q)$ are updated for each substring q.

We use the example in Section 2.1 for illustration. The quadruple for f_1 is $(5, 2, 0, 0)$, indicating that the phrase length is 5 and appears 2 times by itself in the corpus. Therefore $C\text{-}value(f_1) = 8$. The quadruple for f_2 is $(4, 2, 2, 1)$, indicating that the phrase length is 4 and appears 2 times in the corpus. However, the occurrences are as a substring of the phrase f_1. Therefore, $C\text{-}value(f_2) = 0$. While the quadruple for f_3 is $(3, 11, 11, 9)$, which indicates that the phrase length is 3 and appears 11 times as a substring in 9 phrases, thus $C\text{-}value(f_3) = 19.6$. Given the threshold $\varepsilon = 5$, f_1 and f_3 are viewed as key phrases. Thus R_2 will be discarded because its source-side is not a key phrase.

3 Experiments

Our experiments were carried out on two language pairs:

- **Chinese-English**: For this task, the corpora are from the NIST evaluation. The parallel corpus [1] consists of 1M sentence pairs . We trained two trigram language models: one on the Xinhua portion of the Gigaword corpus, and the other on the target-side of the parallel corpus. The test sets were NIST MT06 GALE set (06G) and NIST set (06N) and NIST MT08 test set.

- **German-English**: For this task, the corpora are from the WMT [2] evaluation. The parallel corpus contains 1.3M sentence pairs. The target-side was used to train a trigram language model. The test sets were WMT06 and WMT07.

For both the tasks, the word alignment were trained by GIZA++ in two translation directions and refined by "grow-diag-final" method (Koehn et al., 2003). The source-side of the parallel corpus is used to extract key phrases.

3.1 Results

We reimplemented the state-of-the-art hierarchical MT system, Hiero (Chiang, 2005), as the baseline system. The results of the experiments are shown in Table 1 and Table 2.

Table 1 shows the $C\text{-}value$ threshold effect on the size of the rule table, as well as the BLEU scores. Originally, 103M and 195M rules are respectively extracted for Chinese-English and German-English. For both the two tasks, about 78% reduction of the rule table (for Chinese-English $\varepsilon = 200$ and for German-English $\varepsilon = 100$) does not worsen translation performance. We achieved improvements in BLEU on most of the test corpora, except a slight decrease (0.06 point) on WMT07.

We also compared the effects of *frequency* and $C\text{-}value$ metrics for the rule table reduction on Chinese-English test sets. The rule table is reduced to the same size (22% of original table) using the two metrics, separately. However, as shown in Table 2, the *frequency* method decreases the BLEU scores, while the $C\text{-}value$ achieves improvements. It indicates that $C\text{-}value$ is more appropriate than *frequency* to evaluate the importance of phrases, because it considers more factors.

With the rule table filtered by key phrases on the source side, the number of source phrases reduces. Therefore during decoding, a source sentence is suggested to be decomposed into a number of "key phrases", which are more reliable than the discarded phrases. Thus the translation quality does not become worse.

3.2 Adding C-value as a Feature

Conventional phrase-based approaches performed phrase segmentation for a source sentence with a uniform distribution. However, they do not consider the weights of source phrases. Although any strings can be phrases, it is believed that some strings are more likely phrases than others. We use $C\text{-}value$ to describe the weight of a phrase in a monolingual corpus and add it as a feature to the translation model:

[1] LDC2002E18 (4,000 sentences), LDC2002T01, LDC2003E07, LDC2003E14, LDC2004T07, LDC2005T10, LDC2004T08_HK_Hansards (500,000 sentences)

[2] http://www.statmt.org/wmt07/shared-task.html

C-value	Chinese-English				Germany-English		
Threshold ε	Rule Table (%)	06G	06N	08	Rule Table (%)	06	07
0	100%	12.43	28.58	21.57	100%	27.30	27.95
5	61%	12.22	28.40	21.33	54%	27.39	**28.05**
20	44%	12.24	28.29	21.21	37%	27.47	27.94
100	28%	12.36	28.56	21.67	22%	**27.54**	27.89
200	22%	**12.66**	**28.69**	**22.12**	17%	27.26	27.80
300	20%	12.41	27.76	21.52	15%	27.41	27.69
400	18%	11.88	26.98	20.70	13%	27.36	27.76
500	16%	11.65	26.40	20.32	12%	27.25	27.76

Table 1: C-value threshold effect on the rule table size and BLEU scores.

System	Rule Table (%)	06G	06N	08
Baseline	100%	12.43	28.58	21.57
Frequency	22%	12.24	27.77	21.20
C-value	22%	12.66	28.69	22.12^{*}
+CV-Feature	22%	12.89^{*}	29.22^{*+}	22.56^{*+}

Table 2: BLEU scores on the test sets of the Chinese-English task. * means significantly better than baseline at $p < 0.01$. $^{+}$ means significantly better than C-value at $p < 0.05$.

$$h(F_1^J) = \sum_{k=1}^{K} log(C\text{-}value(\widetilde{f}_k)) \qquad (1)$$

where, $\widetilde{f}_k$ is the source-side of a rule.

The results are shown in the row *+CV-Feature* in Table 2. Measurable improvements are obtained on all test corpora of the Chinese-English task by adding the $C\text{-}value$ feature. The improvements over the baseline are statistically significant at $p < 0.01$ by using the significant test method described in (Koehn, 2004).

4 Conclusion

In this paper, we successfully discarded most entries of the rule table with monolingual key phrases. Experiments show that about 78% of the rule table is reduced and the translation quality does not become worse. We achieve measurable improvements by incorporating $C\text{-}value$ into the translation model.

The use of key phrases is one of the simplest method for the rule table reduction. In the future, we will use sophisticated metrics to score phrases and reduce the rule table size with the information of both the source and target sides.

References

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *COLING1996*, pages 41–46.

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.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

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, Columbus, Ohio, June.