Partial Matching Strategy for Phrase-based Statistical Machine Translation

Zhongjun He[1,2] and **Qun Liu**[1] and **Shouxun Lin**[1]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100190, China
[2]Graduate University of Chinese Academy of Sciences
Beijing, 100049, China
{zjhe,liuqun,sxlin}@ict.ac.cn

Abstract

This paper presents a partial matching strategy for phrase-based statistical machine translation (PBSMT). Source phrases which do not appear in the training corpus can be translated by word substitution according to partially matched phrases. The advantage of this method is that it can alleviate the data sparseness problem if the amount of bilingual corpus is limited. We incorporate our approach into the state-of-the-art PBSMT system Moses and achieve statistically significant improvements on both small and large corpora.

1 Introduction

Currently, most of the phrase-based statistical machine translation (PBSMT) models (Marcu and Wong, 2002; Koehn et al., 2003) adopt full matching strategy for phrase translation, which means that a phrase pair $(\widetilde{f}, \widetilde{e})$ can be used for translating a source phrase $\bar{f}$, only if $\widetilde{f} = \bar{f}$. Due to lack of generalization ability, the full matching strategy has some limitations. On one hand, the data sparseness problem is serious, especially when the amount of the bilingual data is limited. On the other hand, for a certain source text, the phrase table is redundant since most of the bilingual phrases cannot be fully matched.

In this paper, we address the problem of translation of *unseen phrases*, the source phrases that are not observed in the training corpus. The alignment template model (Och and Ney, 2004) enhanced phrasal generalizations by using words classes rather than the words themselves. But the phrases are overly generalized. The hierarchical phrase-based model (Chiang, 2005) used hierarchical phrase pairs to strengthen the generalization ability of phrases and allow long distance reorderings. However, the huge grammar table greatly increases computational complexity. Callison-Burch et al. (2006) used paraphrases of the trainig corpus for translating unseen phrases. But they only found and used the semantically similar phrases. Another method is to use multi-parallel corpora (Cohn and Lapata, 2007; Utiyama and Isahara, 2007) to improve phrase coverage and translation quality.

This paper presents a partial matching strategy for translating unseen phrases. When encountering unseen phrases in a source sentence, we search partially matched phrase pairs from the phrase table. Then we keep the translations of the matched part and translate the unmatched part by word substitution. The advantage of our approach is that we alleviate the data sparseness problem without increasing the amount of bilingual corpus. Moreover, the partially matched phrases are not necessarily synonymous. We incorporate the partial matching method into the state-of-the-art PBSMT system, Moses. Experiments show that, our approach achieves statistically significant improvements not only on small corpus, but also on large corpus.

2 Partial Matching for PBSMT

2.1 Partial Matching

We use *matching similarity* to measure how well the source phrases match each other. Given two source phrases $\widetilde{f}_1^J$ and $\widetilde{f'}_1^J$, the matching similarity is computed as:

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 161–164,
Columbus, Ohio, USA, June 2008.

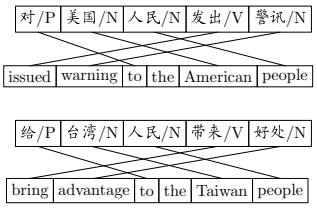

Figure 1: An example of partially matched phrases with the same POS sequence and word alignment.

$$SIM(\widetilde{f}_1^J, \widetilde{f'}_1^J) = \frac{\sum_{j=1}^{J} \delta(f_j, f'_j)}{J} \quad (1)$$

where,

$$\delta(f, f') = \begin{cases} 1 & \text{if } f = f' \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Therefore, partial matching takes full matching ($SIM(\widetilde{f}, \bar{f}) = 1.0$) as a special case. Note that in order to improve search efficiency, we only consider the partially matched phrases with the same length.

In our experiments, we use a matching threshold α to tune the precision of partial matching. Low threshold indicates high coverage of unseen phrases, but will suffer from much noise. In order to alleviate this problem, we search partially matched phrases under the constraint that they must have the same parts-of-speech (POS) sequence. See Figure 1 for illustration. Although the matching similarity of the two phrases is only 0.2, as they have the same POS sequence, the word alignments are the same. Therefore, the lower source phrase can be translated according to the upper phrase pair with correct word reordering. Furthermore, this constraint can sharply decrease the computational complexity since there is no need to search the whole phrase table.

2.2 Translating Unseen Phrases

We translate an unseen phrase f_1^J according to the partially matched phrase pair ($f'^J_1, e'^I_1, \widetilde{a}$) as follows:

1. Compare each word between f_1^J and f'^J_1 to get the position set of the different words: $P = \{j | f_j \neq f'_j, j = 1, 2, \ldots, J\}$;
2. Remove f'_j from f'^J_1 and e'_{a_j} from e'^I_1, where $j \in P$;
3. Find the translation e for $f_j (j \in P)$ from the phrase table and put it into the position a_j in e'^I_1 according to the word alignment $\widetilde{a}$.

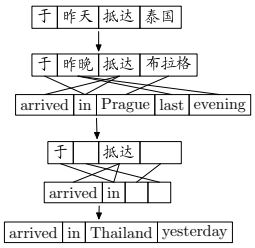

Figure 2: An example of phrase translation.

Figure 2 shows an example. In fact, we create a translation template dynamically in step 2:

$$\langle \text{于 X}_1 \text{ 抵达 X}_2, \text{arrived in X}_2 \text{ X}_1 \rangle \quad (3)$$

Here, on the source side, each of the non-terminal X corresponds to a single source word. In addition, the removed sub-phrase pairs should be consistent with the word alignment matrix.

Following conventional PBSMT models, we use 4 features to measure phrase translation quality: the translation weights $p(\widetilde{f}|\widetilde{e})$ and $p(\widetilde{e}|\widetilde{f})$, the lexical weights $p_w(\widetilde{f}|\widetilde{e})$ and $p_w(\widetilde{e}|\widetilde{f})$. The new constructed phrase pairs keep the translation weights of their "parent" phrase pair. The lexical weights are computed by word substitution. Suppose $S\{(f', e')\}$ is the pair set in ($\widetilde{f'}$,$\widetilde{e'}$,$\widetilde{a}$) which replaced by $S\{(f, e)\}$ to create the new phrase pair ($\widetilde{f}$,$\widetilde{e}$,$\widetilde{a}$), the lexical weight is computed as:

$$p_w(\widetilde{f}|\widetilde{e}, \widetilde{a}) = \frac{p_w(\widetilde{f'}|\widetilde{e'}, \widetilde{a}) \times \prod_{(f,e) \in S\{(f,e)\}} p_w(f|e)}{\prod_{(f',e') \in S\{(f',e')\}} p_w(f'|e')} \quad (4)$$

Therefore, the newly constructed phrase pairs can be used for decoding as they have already existed in the phrase table.

2.3 Incorporating Partial Matching into the PBSMT Model

In this paper, we incorporate the partial matching strategy into the state-of-the-art PBSMT system, Moses[1]. Given a source sentence, Moses firstly uses the full matching strategy to search all possible *translation options* from the phrase table, and then uses a beam-search algorithm for decoding.

[1]http://www.statmt.org/moses/

Therefore, we do incorporation by performing partial matching for phrase translation before decoding. The advantage is that the main search algorithm need not be changed.

For a source phrase $\widetilde{f}$, we search partially matched phrase pair ($\widetilde{f}'$, $\widetilde{e}'$, $\widetilde{a}$) from the phrase table. If $SIM(\widetilde{f}, \widetilde{f}')$=1.0, which means $\widetilde{f}$ is observed in the training corpus, thus $\widetilde{e}'$ can be directly stored as a translation option. However, if $\alpha \leq SIM(\widetilde{f}, \widetilde{f}') < 1.0$, we construct translations for $\widetilde{f}$ according to Section 2.2. Then the newly constructed translations are stored as translation options.

Moses uses translation weights and lexical weights to measure the quality of a phrase translation pair. For partial matching, besides these features, we add matching similarity $SIM(\widetilde{f}, \widetilde{f}')$ as a new feature. For a source phrase, we select top N translations for decoding. In Moses, N is set by the pruning parameter *ttable-limit*.

3 Experiments

We carry out experiments on Chinese-to-English translation on two tasks: **Small-scale task**, the training corpus consists of 30k sentence pairs (840K + 950K words); **Large-scale task**, the training corpus consists of 2.54M sentence pairs (68M + 74M words). The 2002 NIST MT evaluation test data is used as the development set and the 2005 NIST MT test data is the test set. The baseline system we used for comparison is the state-of-the-art PBSMT system, Moses.

We use the ICTCLAS toolkit[2] to perform Chinese word segmentation and POS tagging. The training script of Moses is used to train the bilingual corpus. We set the maximum length of the source phrase to 7, and record word alignment information in the phrase table. For the language model, we use the SRI Language Modeling Toolkit (Stolcke, 2002) to train a 4-gram model on the Xinhua portion of the Gigaword corpus.

To run the decoder, we set ttable-limit=20, distortion-limit=6, stack=100. The translation quality is evaluated by BLEU-4 (case-sensitive). We perform minimum-error-rate training (Och, 2003) to tune the feature weights of the translation model to maximize the BLEU score on development set.

[2]http://www.nlp.org.cn/project/project.php?proj_id=6

α	1.0	0.7	0.5	0.3	0.1
BLEU	24.44	24.43	24.86	**25.31**	25.13

Table 1: Effect of matching threshold on BLEU score.

3.1 Small-scale Task

Table 1 shows the effect of matching threshold on translation quality. The baseline uses full matching (α=1.0) for phrase translation and achieves a BLEU score of 24.44. With the decrease of the matching threshold, the BLEU scores increase. when α=0.3, the system obtains the highest BLEU score of 25.31, which achieves an absolute improvement of 0.87 over the baseline. However, if the threshold continue decreasing, the BLEU score decreases. The reason is that low threshold increases noise for partial matching.

The effect of matching threshold on the coverage of n-gram phrases is shown in Figure 3. When using full matching (α=1.0), long phrases (length$\geq$3) face a serious data sparseness problem. With the decrease of the threshold, the coverage increases.

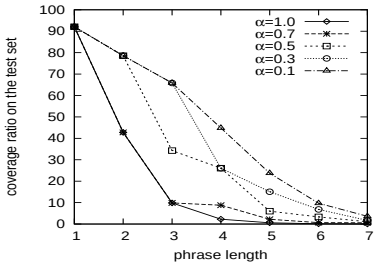

Figure 3: Effect of matching threshold on the coverage of n-gram phrases.

Table 2 shows the phrase number of 1-best output under α=1.0 and α=0.3. When α=1.0, the long phrases (length$\geq$3) only account for 2.9% of the total phrases. When α=0.3, the number increases to 10.7%. Moreover, the total phrase of α=0.3 is less than that of α=1.0, since source text is segmented into more long phrases under partial matching, and most of the long phrases are translated from partially matched phrases (the row *0.3*$\leq SIM <$*1.0*).

3.2 Large-scale Task

For this task, the BLEU score of the baseline is 30.45. However, for partial matching method with

Phrase Length		1	2	3	4	5	6	7	total
α=1.0		19485	4416	615	87	12	2	1	24618
α=0.3	SIM=1.0	14750	2977	387	48	10	1	0	21195
	0.3$\leq SIM <$1.0	0	1196	1398	306	93	17	12	

Table 2: Phrase number of 1-best output. α=1.0 means full matching. For α=0.3, SIM=1.0 means full matching, $0.3 \leq SIM < 1.0$ means partial matching.

α=0.5[3], the BLEU score is 30.96, achieving an absolute improvement of 0.51. Using Zhang's significant tester (Zhang et al., 2004), both the improvements on the two tasks are statistically significant at $p < 0.05$.

The improvement on large-scale task is less than that on small-scale task since larger corpus relieves data sparseness. However, the partial matching approach can also improve translation quality by using long phrases. For example, the segmentation and translation for the Chinese sentence "但是 经济 产出 的 长期 趋势 将" are as follows:

Full matching:
长期 | 经济 产出 | 但是 | 的 | 趋势 | 将
long term | *economic output* | , but | *the* | *trend* | will

Partial matching:
但是 | 经济 产出 的 长期 趋势 | 将
but | *the long-term trend of economic output* | will

Here the source phrase "经济 产出 的 长期 趋势" cannot be fully matched. Thus the decoder breaks it into 4 short phrases, but performs an incorrect reordering. Using partial matching, the long phrase is translated correctly since it can partially matched the phrase pair "*经济 发展 的 必然 趋势, the inevitable trend of economic development*".

3.3 Conclusion

This paper presents a partial matching strategy for phrase-based statistical machine translation. Phrases which are not observed in the training corpus can be translated according to partially matched phrases by word substitution. Our method can relieve data sparseness problem without increasing the amount of the corpus. Experiments show that our approach achieves statistically significant improvements over the state-of-the-art PBSMT system Moses.

In future, we will study sophisticated partial matching methods, since current constraints are excessively strict. Moreover, we will study the effect of word alignment on partial matching, which may affect word substitution and reordering.

[3]Due to time limit, we do not tune the threshold for large-scale task.

Acknowledgments

We would like to thank Yajuan Lv and Yang Liu for their valuable suggestions. This work was supported by the National Natural Science Foundation of China (NO. 60573188 and 60736014), and the High Technology Research and Development Program of China (NO. 2006AA010108).

References

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proc. of NAACL06*, pages 17–24.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of ACL05*, pages 263–270.

T. Cohn and M. Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proc. of ACL07*, pages 728–735.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL03*, pages 127–133.

D. Marcu and W. Wong. 2002. A phrasebased joint probabilitymodel for statistical machine translation. In *Proc. of EMNLP02*, pages 133–139.

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL03*, pages 160–167.

A. Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. of ICSLP02*, pages 901–904.

M. Utiyama and H. Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. of NAACL-HLT07*, pages 484–491.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proc. of LREC04*, pages 2051–2054.