A Chunk-Based Reordering Model for Phrase-Based SMT Systems

Yidong CHEN, Xiaodong SHI, Changle ZHOU, Qingyang HONG

Department of Cognitive Science, School of Information Science and Technology, Xiamen University, Xiamen, Fujian, P. R. China

{ydchen, mandel, dozero, qyhong}@xmu.edu.cn

Abstract:

This paper proposed a novel reordering model based on the reordering of source language chunks. This model is used as a preprocessing step of phrase-based translation models and could be well integrated with them. At the same time, as a chunk-based model, syntax information could be concerned in the process of reordering while the entire parsing of the source sentence is not required. Two experiments were carried out and the results showed that the proposed model could improve the performance of a phrase-based statistical machine translation (SMT) system greatly.

Keywords:

Statistical Machine Translation (SMT); Phrase-Based Translation Models (PBTM); Reordering Models; Chunk-Based Models

1. Introduction

In the late 1980s and early 1990s, statistical techniques were first applied for machine translation in the work of IBM [1, 2], which led to a dramatic improvement of the quality of current machine translation systems. Among the statistical machine translation (SMT) models, phrase-based models [3, 4, 5] have achieved great success and become the dominating models. Typically, the alignment template based translation model [5] obtained the best performance in the U.S. National Institute of Standards and Technology (NIST)/TIDES MT evaluations from 2001 to 2006, and was considered as the state-of-art SMT model.

In spite of the success they have achieved, phrase-based SMT models are beset by a number of difficult theoretical and practical problems, one of which is global reordering problem [6]. Many recent studies on SMT work hard to improve phrase-based SMT model by integrating global reordering and yielded some promising results (see Related Work in section 5 for detail).

In this paper, we proposed a novel model based on the reordering of source language chunks. It is used as a preprocessing step of phrase-based translation models and could be well integrated with them. At the same time, as a chunk-based model, syntax information could be concerned in the process of reordering while the entire parsing of the source sentence is not required.

The rest of this paper is organized as follows. Section 2 gives the basic idea and framework of the reordering model. Section 3 and Section 4 describe the reordering algorithm used in the training and decoding time, respectively. Section 5 addresses the related work and the differences among them and our work. Section 6 reports the experimental results. Section 7 gives the conclusion.

2. Framework

One of the most crucial problems of a phrase-based system is that it is lack of the ability of global reordering. The reason behind this problem is the non-grammatical boundary of the phrases used in phrase-based systems. To overcome this problem, some researchers tried to incorporate syntactic information by integrating parsing [7, 8, 9, 10], which have gotten promising results.

However, performing entirely parsing is slow and sometimes would bring up many errors. These problems of entire parsing could be harmful for the whole translation system. So, in our improvement, we decide to use shallow parsing, which is much simple than entire parsing and thus could be more efficient and accuracy.

Moreover, to make full use of the advantages of phrase-based systems, we adopt a similar approach similar to [7, 8], i.e. integrating a reordering model in the preprocessing step of phrase-based systems.

In Figure 1 and Figure 2, we address the training steps and the decoding steps of our model respectively.

Among the steps described above, two ones are important, i.e. the reordering model used in the training time and the reordering model used in the decoding time. We will discuss these two steps and the corresponding data training steps in detail in Section 3 and Section 4, respectively.

978-1-4244-2780-2/08/$25.00 ©2008 IEEE

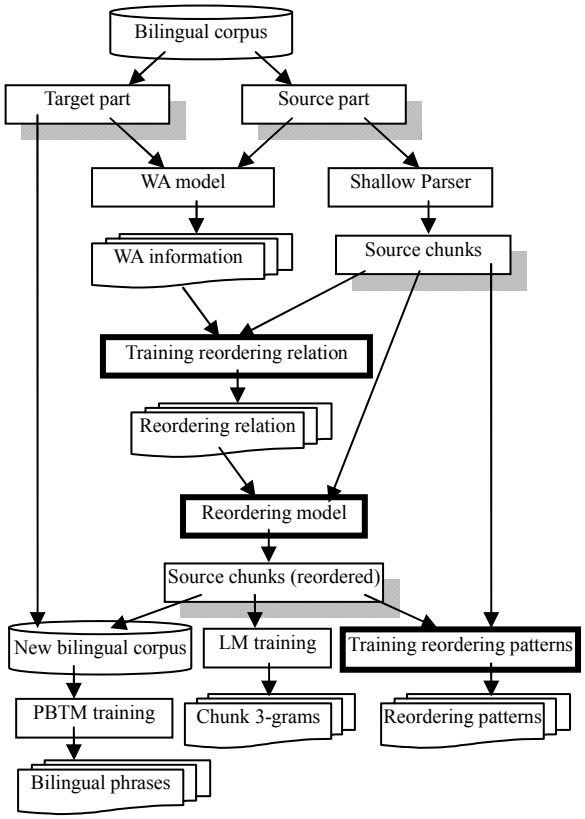

Figure 1. Training Steps.

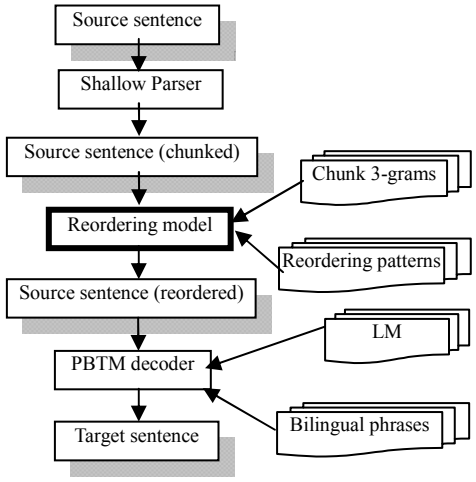

Figure 2. Decoding Steps.

3. The Reordering Model in Training Time

For a sentence with n chunks, the problem of finding the best reordering result is equivalent to a TSP problem, and thus is NP-hard [11]. In this section, we will show that by giving some limitations this problem could be solved efficiently in polynomial time.

3.1. Chunk Reordering vs. Integer Sorting

The chunk reordering problem could be considered as a problem of finding a permutation of the chunks that is the best one according to the target language order, and thus is similar to the problem of sorting, whose aim is to find a permutation of a given integer sequence so that the integers are in ascending or descending order.

Formally, given an integer sequence S_0, the aim of the sorting is to find the $\hat{S}$ using Formula 1.

$$\hat{S} = \arg\max_{S} |\{(a_i, a_j) \mid a_i \in S \wedge a_j \in S \wedge i < j \wedge < a_i, a_j > \in R\}| \quad (1)$$

where, S is a permutation of S_0 and R is the binary relation $\leq$ or $\geq$.

Similarly, we may define the aim of the chunk reordering problem as to find the best permutation $\hat{O}$ of given chunk sequence O_0 by using Formula 2, which is shown as follows.

$$\hat{O} = \arg\max_{O} \sum_{\substack{c_i \in O \wedge c_j \in O \wedge \\ i<j \wedge <a_i, a_j> \in R}} \text{Prob}(< c_i, c_j >) \quad (2)$$

where, O is a permutation of O_0 and R is a reordering relation. A reordering relation is an extension of a binary relation. Each element $< c, c' >$ belonging to the reordering relation is attached a probability $\text{Prob}(< c, c' >)$ which shows how likely the chunk c should be put before the chunk c' as far as the target language order is concerned.

3.2. Training the Reordering Relation

Given a word-aligned bilingual corpus, with the source sentences chunked, the training of the reordering relation could be realized in a straightforward way, which is described as follows.

First, gather all the chunk pairs $< c, c' >$ using the word alignment matrix of each sentence pair, and then count the frequencies of their emergences, $N(< c, c' >)$. A chunk pair $< c, c' >$ will be recorded if and only if it satisfies the following condition:

$$|\{< i, j > | i \in TSet(c) \wedge j \in TSet(c') \wedge i < j\}| > 0.5$$

where $TSet(c)$ is a set of the word indexes of the target words that are aligned to source words in chunk c according to the word alignment matrix.

Then, merge the chunk pairs and calculate the probability according to the relative frequencies using Formula 3 as follows.

$$\text{Prob}(< c, c' >) = \frac{N(< c, c' >)}{N(< c, c' >) + N(< c', c >)} \quad (3)$$

3.3. A Selection-Sort-Like Reordering Algorithm

As discussed in 3.1, the chunk reordering problem is very similar to a sorting problem. Why not try using sorting algorithms to solve the reordering problem? Actually, the chunk reordering problem could be solved using an algorithm similar to the selection sort algorithm, if some limitation (see below) is obeyed.

First of all, it is better to look more closely on the selection sort. The selection sort for a given array of integers performs sorting by repeatedly putting the smallest element in the unprocessed portion of the array to the beginning of it until the whole array is sorted. Where, the aim of the each iteration of the selection sort algorithm is to find the $\hat{i}$ using Formula 4:

$$\hat{i} = \arg\max_{i \in s}(|\{j \mid j \in s \wedge j \neq i \wedge < i, j > \in R\}|) \quad (4)$$

Here, s is the unprocessed portion of the integer array and R is the binary relation ≤ or ≥.

Likewise, given a sequence of chunks, our reordering algorithm will perform reordering by repeatedly putting the chunk, which satisfies Formula 5, in the unprocessed portion of the sequence to the beginning of it until the whole sequence is reordered.

$$\hat{c} = \arg\max_{c \in s} \sum_{c' \in s \wedge c' \neq c \wedge < c, c' > \in R} \text{Prob}(< c, c' >) \quad (5)$$

In Formula 5, s is the unprocessed portion of the chunk sequence and R is the reordering relation.

The algorithm described above has a worst-case complexity of $O(n^3)$, where n is the number of chunks in the chunk sequence.

4. The Reordering Model in Decoding Time

Although the algorithm described in 3.3 is efficient, it has an obvious limitation. When using this algorithm, no context information could be incorporated, since the algorithm regards the chunks to be reordered independent to each other. To incorporate context information, we use a totally reordering model in the decoding time.

4.1. Basic Idea

Given a chunked source sentence $\bar{\mathbf{c}} = \bar{c}_1^L$ and its reordered version $\bar{\mathbf{d}} = \bar{d}_1^L$, where $\bar{\mathbf{d}}$ is a permutation of $\bar{\mathbf{c}}$. The reordering problem could be defined as a problem to find the $\hat{\bar{d}}_1^L$ which makes the probability $\Pr(\bar{d}_1^L \mid \bar{c}_1^L)$ maximum.

Note that the problem described above could be considered as a translation problem. Thus we may use phrase-based translation models to settle it. The advantage of using such models is that the context information could be taken into account during the course of reordering.

4.2. Training

Two kinds of data should be trained

- Chunk tag n-gram.
- Reordering patterns.

Here we define a reordering pattern as a 3-tuple, $(CTS, Perm, Prob)$, where CTS is a sequence of consecutive chunk tags, $Perm$ is an integer sequence stands for a potential permutation of the chunks in CTS, and $Prob$ is the probability of the reordering pattern, which could be estimated using relative frequencies.

Given a chunked Chinese sentence and its reordered version, as shown in Figure 3 (a) and Figure 3 (b), respectively, the reordering patterns listed in Table 1 will be extracted.

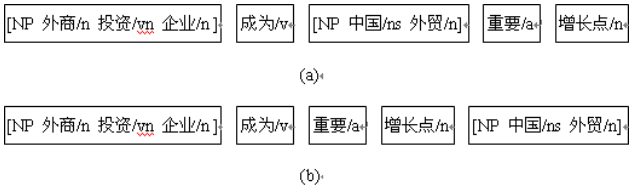

Figure 3. A chunked Chinese sentence and its reordered version.

Table 1. The extracted reordering patterns.

CTS	*Perm*	*Prob*
NP v	1 2	…
NP a n	2 3 1	…
a n	1 2	…

Please note that the *Prob* parts for each reordering patterns will not be calculated until the entire training corpus has been processed.

4.3. A DP-Based Reordering Algorithm

In order to clearly describe the reordering process, we define the quantity $Q(l, \bar{d})$ as the maximum probability of a chunk sequence that ends with the chunk $\bar{d}$ and covers positions 1 to l of the chunked source sentence. $Q(L+1, \$)$ is the probability of the optimal reordering result. The \$ symbol is the sentence boundary marker. We obtain the following dynamic programming recursion:

$$Q(0, \$) = 1 \quad (6)$$

$$Q(l, \bar{d}) = \max_{\substack{0 \le l' < l, \\ \bar{d}', \tilde{\bar{d}}}} Q(l', \bar{d}') \cdot p(\bar{c}_{l'+1}^{l} \mid \tilde{\bar{d}}) \cdot p(\tilde{\bar{d}} \mid \bar{d}') \quad (7)$$

$$Q(L+1, \$) = \max_{\bar{d}'} Q(L, \bar{d}') \cdot p(\$ \mid \bar{d}') \quad (8)$$

where, $\tilde{\bar{d}}$ in Formula 7 is a possible reordering result

for chunk sequence $\bar{c}_{l'+1}^{l}$, according to the reordering patterns.

Given the DP recursions shown in Formula 6, 7 and 8, a DP-based reordering algorithm could be constructed easily.

5. Related Works

Approaches proposed by Xia and McCord [7] and Collins et al. [8] are similar to ours. All the three authors select the strategy that improves a phrase-based using reordering in the preprocessing step. Our approach differs from their work in that ours do not need entire parsing of the source sentence, and thus is more efficient.

Model presented by Chiang [12], Zens and Ney [13] and Xiong et al. [14] all involved enhancing phrase-based systems by incorporating global reordering. However, in their model, no syntactic information is considered. Reordering in our model, on the contrary, is based on chunk, which is a syntactic unit. Though it has not been proved formally, more and more researchers agreed that, syntactic information could be helpful when dealing many phenomena, including word reordering, during translation.

The studies of Schafer and Yarowsky [15] and Watanabe et al. [16] were also dependent on shallow parsing of the source sentences. However, in their models the chunks are not only used as reordering units but also used as the translation units. Our model only uses chunks as reordering units and will finally use phrase-based systems to translate the reordered source sentence. By this means, our model could make full use of the advantages of phrase-based systems.

6. Experiments

Two experiments were carried out. The first one tested the performance of the reordering model, and the second one considers the influence of the chunk level.

6.1. The Performance of the Reordering Model

In this experiment, we evaluated the performance of our chunk-based reordering model. We used a monotone phrase-based SMT system called Caravan [17] as the baseline system. BLEU score [18] was used to evaluate the translation performance of the translation systems.

The statistics for the data used in our experiments are shown in Table 2. We used the writing part of the test set from *2005 China's National 863 MT Evaluation* as our test data. The Chinese-English bilingual corpus was used to extract bilingual phrases and reordering patterns. The treebank was used to train the shallow parser. The English corpus, which is the English part of the bilingual corpus, was used to train English language model.

Table 2. The statistics for the data used.

	Amounts
Bilingual corpus	833,394 sentence pairs
TreeBank	18,782 sentence
English corpus	833,394 sentence
Test set	489 sentences

We used GIZA++ package [19] to perform word alignment. CRF++ Toolkit [20] was used to train the chunker. Language model was trained using SRI Language Modeling Toolkit [21] with modified Kneser-Ney smoothing [22]. Only trigram language model was trained on the training corpus.

The BLEU scores of the systems are listed in Table 3, as followed.

Table 3. The BLEU scores of the systems.

Systems	BLEU-4 case sensitive
Caravan	0.1612
Caravan + Chunk-based reordering	0.1923

It should be learned from the results above that Caravan worked much better after it was integrated with the chunk reordering component. This implies that our chunk reordering model may bring up great improvement.

The speeds of the systems were also evaluated and the results are listed in Table 4. The results show that, our chunk reordering model will not slow down the decoding significantly, and thus indicate that the model is efficient.

Table 4. The speeds of the systems.

Systems	Sents/min
Caravan	333
Caravan + Chunk-based reordering	296

6.2. The Influence of the Chunk Level

In the proposed chunk-based reordering model, only one-level chunk was used. We perform this experiment in order to test the capacity of the model.

In this experiment, we first calculated the reordering depths for 4103 sentences in Chinese Penn TreeBank that have English translations, and then computed their distribution. Here, a reordering depth is defined using Formula 9, 10 and 11:

$$RD(T) = \max_{p \in Path(T)} RDP(p) \tag{9}$$

$$RDP(p) = \sum_{n \in p} RC(n) \tag{10}$$

$$RC(n) = \begin{cases} 1 & ;\textit{if any two children of} \\ & \quad n \textit{ shoud be reordered} \\ 0 & ;\textit{otherwise} \end{cases} \tag{11}$$

where, $Path(T)$ gives all the paths of T.

Please note that the sub-trees that satisfy Formula

11 were treated as leaf nodes.

$$|Span(n)| <= ML_{bp} \wedge |Span(Parent(n))| > ML_{bp} \quad (12)$$

where, $|\cdot|$ calculates the cardinal of a given set, $Span(n)$ gives a set of the leaf nodes in the sub-tree rooted at n, $Parent(n)$ returns the parent node of node n, ML_{bp} is the maximum length of bilingual phrases.

The statistical results are listed in Table 5. The results show that the ratio of sentences whose reordering depth is 0 or 1 reaches two third. This indicates that the proposed reordering model that relied on one-level chunk may cover most reordering cases.

Table 5. The statistical results.

Reordering depth	0	1	2	3	4	5	6	7	8	9	10	11	12
Sentence counts	1594	1134	697	386	169	80	24	11	4	2	1	0	1

7. Conclusions

In this paper we presented a novel reordering model based on source language chunks. The model has three advantages. First, it could be easily integrated with traditional phrase-based translation models. Second, it could use syntax information while performing the reordering. Third, it is efficient since it only relies on shallow parsing. The experimental results shown that the proposed model could improve the performance of a phrase-based SMT system significantly.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 60573189), the National 863 High Technology Research and Development Program of China (Grant No. 2006AA01Z139, 2006AA010107 and 2006AA010108), the Natural Science Foundation of Fujian Province (Grant No. 2006J0043) and the Fund of Key Research Project of Fujian Province (Grant No. 2006H0038).

References

[1] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S Rossin. "A Statistical Approach to Machine Translation". *Computational Linguistics*, Vol. 16, No. 2, pp. 79-85, 1990.

[2] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mecer. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics*, Vol. 19, No. 2, pp. 263-311, 1993.

[3] Richard Zens, Franz J. Och, and Hermann Ney. "Phrase-Based Statistical Machine Translation". In *Proceedings of the 25th Annual German Conference on AI: Advances in Artificial Intelligence*, Aachen, Germany, pp. 18-22, 2002.

[4] Philipp Koehn, Franz J. Och, and Danie Marcu. "Statistical Phrase-Based Translation". In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, Canada, pp. 127-133, 2003.

[5] Franz J. Och, and Hermann Ney "The Alignment Template Approach to Statistical Machine Translation". *Computational Linguistics*, Vol. 30, No. 4, pp. 417-499, 2004.

[6] Chris Quirk, and Arul Menezes. "Do We Need Phrases? Challenging the Conventional Wisdom in Statistical Machine Translation". In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, New York, USA, pp. 9-16, 2006.

[7] Fei Xia, and Michael McCord. "Improving a Statistical MT System with Automatically Learned Rewrite Patterns". In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 508-514, 2004.

[8] Michael Collins, Philipp Koehn, and Ivona Kučerová "Clause Restructuring for Statistical Machine Translation". In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, USA, pp. 531-540, 2005.

[9] Chris Quirk, Arul Menezes, and Colin Cherry. "Dependency Treelet Translation: Syntactically Informed Phrasal SMT". In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, USA, pp. 271-279, 2005.

[10] Yang Liu, Qun Liu, and Shouxun Lin. "Tree-to-String Alignment Template for Statistical Machine Translation". In *Proceedings of the 21st International Conference on Computational Linguistics (COLING 2006) and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, Sydney, Australia, pp. 609-616, 2006.

[11] Kevin Knight. "Decoding Complexity in Word-Replacement Translation Models". *Computational Linguistics*, Vol. 25, No. 2, pp. 607-615, 1999.

[12] David Chiang. "A Hierarchical Phrase-Based Model for Statistical Machine Translation". In *Proceeding of the 43rd Annual Meeting of the*

Association for Computational Linguistics (ACL 2005), Ann Arbor, USA, pp. 263-270, 2005.

[13] Richard Zens, and Hermann Ney. "Discriminative Reordering Models for Statistical Machine Transla-tion". In *Proceedings of the Workshop on Statistical Machine Translation*, New York, USA. pp. 55-63, 2006.

[14] Deyi Xiong, Qun Liu, and Shouxun Lin. "Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation". In *Proceedings of the 21st International Conference on Computational Linguistics (COLING 2006) and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, Sydney, Australia, pp. 521-528, 2006.

[15] Charles Schafer, and David Yarowsky. "Statistical Machine Translation Using Coercive Two-Level Syntactic Transduction". In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Philadelphia, USA, pp. 20-28, 2003.

[16] Taro Watanabe, Eiichiro Sumita, and Hiroshi G. Okuno. "Chunk-Based Statistical Translaiton". In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, pp. 303-310, 2003.

[17] Yidong Chen, Xiaodong Shi, and Changle Zhou. "The XMU Phrase-Based Statistical Machine Translation System for IWSLT 2006". In *Proceedings of International Workshop on Spoken Language Translation*, Kyoto, Japan, pp. 153-157, 2006.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. "BLEU: A Method for Automatic Evaluation of Machine Translation". In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, USA, pp. 311-318, 2002.

[19] Franz J. Och and Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*, Vol. 29, No. 1, pp. 19-51, 2003.

[20] http://crfpp.sourceforge.net/

[21] Andreas Stolcke. "Srilm - An Extensible Language Modeling Toolkit". In *Proceedings of the International Conference on Spoken language Processing*, volume 2, pp. 901-904, 2002.

[22] Stanley F. Chen, and Joshua Goodman. "An Empirical Study of Smoothing Techniques for Language Modeling". *Technical Report TR-10-98*, Harvard University Center for Research in Computing Technology, 1998.