Improved Chunk-level Reordering for Statistical Machine Translation

Yuqi Zhang, Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Abstract

Inspired by previous chunk-level reordering approaches to statistical machine translation, this paper presents two methods to improve the reordering at the chunk level. By introducing a new lattice weighting factor and by reordering the training source data, an improvement is reported on TER and BLEU. Compared to the previous chunk-level reordering approach, the BLEU score improves 1.4% absolutely. The translation results are reported on IWSLT Chinese-English task.

1. Introduction

In machine translation, reordering is one of the major problems, since different languages have different word order requirements. In current phrase-based Statistical Machine Translation (SMT) systems, distance-based reordering constraints are widely used, such as IBM constraints [1], local constraints [2] and distortion limit [3]. With these models phrase-based SMT is powerful in word reordering within short distance. However, long-distance reordering is still problematic.

In order to solve the long-distance reordering problem, it has been realized that syntactic information should be used. Some approaches have applied at the word-level, such as morphology [4], POS tags [5] and word classes [6]. They are particularly useful for the language with rich morphology for reducing the data sparseness. Another kinds of syntax reordering methods require parse trees, such as the work in [7], [8], [9], [10]. The parse tree is more powerful to capture the sentence structures. However, it is expensive to create tree structures and building a good quality parser is also a hard task.

What we are interested in here is to use an intermediate syntax between POS tag and parse tree: chunks, as the basic unit for reordering. It is not only because chunks are with more syntax than POS tags, but also they are closer to the definition of a "phrase" in phrase-based SMT and easy to use. We have not found much work to do reordering at the chunk level. Schafer [11] has developed a word-chunk two levels syntactic transduction which uses chunks on both language sides. It is a whole translation system. Here, we only apply chunks on source language and are more interested in using chunk knowledge in the phrase-based translation framework.

In this paper, we will improve the approach described in [12] by adding a weight model using the rules probability and repeating training on the reordered sentence pairs. In Section 3, the baseline systems are introduced. Section 4 is the main part of the paper, where the new methods to improve the baseline model are presented. Section 5 describes the experiments and the analysis. Finally, Section 6 is the conclusion.

2. Related work

In the previous chunk level reordering work, [12] has represented the reorderings generated with some rules in a weighted lattice. The lattice is weighted with language model trained on reordered source data. The information from the reordering rules is not used.

The previous work to input a graph to SMT system was done by [13]. Another work with weighted graph is done by [14]. In their N-gram-based SMT system, reordering is handled by a statistical machine reordering (SMR) system, which translate an original source language to a reordered source language. The output of the SMR system is a weighted graph. Their reordering is done at word class level.

Another work is to use multiple reordered inputs instead of single input to the SMT system. [9] represents reordered sentences in a N-best list.

3. Baseline system

3.1. The baseline phrase-based SMT system

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\begin{aligned} \hat{e}_1^{\hat{I}} &= \underset{I, e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} & (1) \\ &= \underset{I, e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} & (2) \end{aligned}$$

This decomposition into two knowledge sources is known as the source-channel approach to statistical machine translation [15]. It allows an independent modeling of the target language model $Pr(e_1^I)$ and the translation model $Pr(f_1^J|e_1^I)$. The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence. The argmax operation denotes the search problem, i.e., the generation of the output sentence in the target language.

An alternative to the classical source-channel approach is the direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$. Using a log-linear model [16], we obtain:

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{{e'}_1^{I'}} \exp\left(\sum_{m=1}^M \lambda_m h_m({e'}_1^{I'}, f_1^J)\right)} \quad (3)$$

The denominator represents a normalization factor that depends only on the source sentence f_1^J. Therefore, we can omit it during the search process. As a decision rule, we obtain:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (4)$$

This approach is a generalization of the source-channel approach. It has the advantage that additional models $h(\cdot)$ can be easily integrated into the overall system. The model scaling factors λ_1^M are trained according to the maximum entropy principle, e.g., using the GIS algorithm. Alternatively, one can train them with respect to the final translation quality measured by an error criterion [17].

The log linear model is a natural framework to integrate many models. During the search of the baseline system we are using the following models:

- phrase translation models (including phrase count features)
- word-based translation models
- word and phrase penalty
- target language model (6-gram)
- jump reordering model (assigning costs based on the jump width)

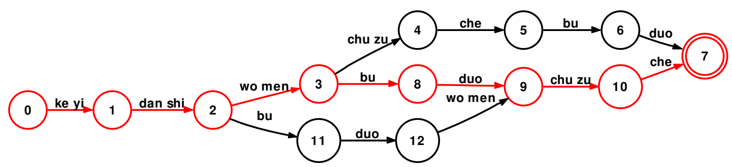
All the experiments in the paper are evaluated without rescoring. More details about the baseline system can be found in [18].

Figure 1: An example of source reordering.

source	ke yi	dan shi	wo men	chu zu	che	bu	duo
POS	v	c	r	v	n	d	m
chunks	v	c	r	NP		VP	
English gloss	yes	but	we	taxi		not many	

used reordering rules
NP VP → VP NP
r NP VP → r VP NP
r NP VP → VP r NP

Reordering Lattice:



3.2. Chunking reordering system

The baseline reordering system we use was described in [12]. The reordering is done in preprocessing stage on the source language side. A source sentence is firstly parsed into chunks. These chunks will be reordered by some rules which are automatically extracted from chunk-to-word alignment. All the reorderings are compacted in a lattice. One arc refers to a word. We have shown an example in Figure 1. In the first table of the example, a source sentence is POS tagged and chunked. Five chunks are generated from seven words. The English gloss is also shown at the last row for each chunks. The three rules for reordering the chunks are listed in the second table. Then the corresponding lattice with the three rules is generated.

Note that when building the lattice, the monotone word sequence without any reordering is guaranteed to be included.

The chunk parser is the maximum entropy tool YASMET [1]. The F-measure is 63.3 for chunk tagging. Since the chunking requires POS tags, "Inst. of Computing Tech., Chinese Lexical Analysis System. (ICTCLAS)" [19] is used. It does word segmentation and Part-Of-Speech tagging in one pass.

The lattice is weighted with a trigram reordered source language model. Each path of the lattice is a permutation $f_{\pi_1}^{\pi_J} = f_{\pi_1}, ..., f_{\pi_J}$ for a given source sentence f_1^J. π_j is the permutation position of word f_j. The weight model used in the decoder is:

$$h_{\text{slm}}(f_{\pi_1}^{\pi_J}, f_1^J) = \log p(f_{\pi_1}^{\pi_J} | f_1^J) \quad (5)$$

$$= \sum_{j=1}^{J} \log p(f_{\pi_j} | f_{\pi_{j-1}}, f_{\pi_{j-2}}) \quad (6)$$

4. Improved chunk reordering system

Two methods will be reported to improve the chunk reordering:

1. new model to weigh the lattice.
2. add additional reordered training data.

4.1. Lattice weighting

Besides the Equation (5), an additional weight model is introduced to evaluate each permutation. The reordering model h_{reorder} is computed using the probabilities of the reordering rules.

After chunk parsing, the original source sentence f_1^J consists of a sequence of chunks: $f_1^J = c_1^N$. π_n is the permutation position of the chunk c_n.

$$h_{\text{reorder}}(\pi_1^N, c_1^N) = \log(p(\pi_1^N | c_1^N)) \quad (7)$$

For a reordered sentence, the π_1^N is generated with a sequence of reordering rules r_1^K. These rules segment source chunks c_1^N into k parts $\tilde{c}_1 ... \tilde{c}_k$. $\tilde{c}$ is

[1]http://www-i6.informatik.rwth-aachen.de/web/Software /index.html

a sequence of chunk c. Similar to phrase-based translation model, we introduce a "hidden" variable B for the segmentations. One permutation can be produced by different rule set with different segmentations. Then, for a given segmentation B, the probability of a permutation is computed by the multiplication of rules probability. For a rule r_k:($\tilde{\pi}_k$, $\tilde{c}_k$), its left hand side is the chunk sequence $\tilde{c}_k$ and its right hand side is the $\tilde{c}_k$'s permutation: $\tilde{\pi}_k$. So, $p(\pi_1^N|c_1^N)$ can be represented as:

$$\begin{aligned} p(\pi_1^N|c_1^N) &= \sum_B p(\pi_1^N, B|c_1^N) && (8)\\ &= \sum_B p(B|c_1^N)\cdot p(\pi_1^N|c_1^N, B) && (9)\\ &= \sum_B \alpha(c_1^N)\cdot p(\pi_1^N|c_1^N, B) && (10)\end{aligned}$$

$$\begin{aligned} p(\pi_1^N|c_1^N, B) &= p(\tilde{\pi}_1^K|\tilde{c}_1^K) && (11)\\ &= \prod_{k=1}^{K} p(\tilde{\pi}_k|\tilde{c}_k) && (12)\end{aligned}$$

When we assume all segmentations have the same probability $\alpha(c_1^N)$, the reordering probability is only relevant to the probabilities of reordering rules, where $p(\tilde{\pi}_k|\tilde{c}_k)$ is defined in Equation (13). It is calculated via relative frequencies. $N(\tilde{\pi}_k|\tilde{c}_k)$ is the count of the rule r_k in the rules training data and $N(\tilde{c}_k)$ is the count of the rules with the same left hand side of r_k.

$$p(\tilde{\pi}_k|\tilde{c}_k) = \frac{N(\tilde{\pi}_k, \tilde{c}_k)}{N(\tilde{c}_k)} \quad (13)$$

Both models $h_{\text{slm}}(f_{\pi_1}^{\pi_J}, f_1^J)$ and $h_{\text{reorder}}(\pi_1^N, c_1^N)$ are integrated into the Equation (4).

4.2. Reordering training data

So far, only the test data is reordered. The training source data is still keeping the original word order, which is inconsistent with the test data. We follow the phrase extraction method described in [13] to filter out all portions of the test source sentence and their translations from the phrase pairs of the training data. Some long phrases could be broken because of the inconsistency of word order between test and training data. It will affect the lexical choice during decoding.

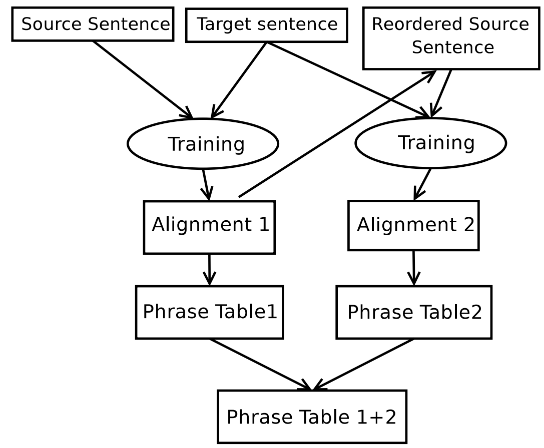


Figure 2: Illustration of the combination of reordered and non-reordered training data.

In order to solve this problem, the phrase table is expanded by extracting phrases from an additional alignment. Besides the alignment training on original data, a second GIZA++ [2] training is run on the reordered training data. The two phrase tables are combined by summing the counts of the same phrase pairs. The process is illustrated in Figure 2. Different from the test data, the training data is reordered not with the rules, but by the alignment. "reordered f" in Figure 2 is generated by reordering the chunks according to the "Alignment 1" to make the source chunks to have similar word order of the target side.

5. Experiments

5.1. Corpus statistics

We perform translation experiments on the Basic Traveling Expression Corpus (BTEC) for the Chinese-English task. It is a speech translation task in the domain of tourism-related information. All data come from the package for the IWSLT

[2] http://www.fjoch.com/GIZA++.html

Table 1: Statistics of training and test corpora for the IWSLT tasks.

		Chinese	English
Train	Sentences	43 k	
	Words	380 k	420 k
	Vocabulary	11 760	9 933
Dev dev2	Sentences	500	
	Words	3 578	3 908
	OOVs	73	–
Test dev3	Sentences	506	
	Words	3 837	3 970
	OOVs	70	–

2007 evaluation. The development corpus is dev2 (IWSLT04 eval data) and the test corpus is dev3 (IWSLT05 eval data). Both dev4 (IWSLT06 dev data) and dev5 (IWSLT06 eval data) and their references are added into training data as bilingual corpora. The corpus statistics are shown in Table 1.

The scaling factors are optimized for the BLEU score. The translation is evaluated case-insensitive and with punctuation marks.

5.2. Evaluation criteria

WER (word error rate). The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference sentence.
PER (position-independent word error rate). The PER compares the words in the hypothesis and references ignoring the word order.
TER (translation error rate). The TER [20] is computed as the number of edits needed to change a system output so that it exactly matches a given reference. The edits include insertions, deletions, substitutions and shifts.
BLEU. This score measures the precision of unigrams, bigrams, trigrams and fourgrams with respect to a reference translation with a penalty for too short sentences [21]. The BLEU score measures accuracy.

5.3. Results

In Table 2, the translation results for the IWSLT05 eval data are reported. The experiments are run comparing to the baseline which is the source reordering weighed only by the source language model. The results of new methods are shown step by step.

- "+ruleProb" uses the probabilities of the reordering rules to weight the reordering lattice. At this step, the BLEU score improves 0.7%.

- "+reordered train data" is the result of enlarging the training data by adding reordered source sentences. After this step, the BLEU is 1.3% better than the baseline.

In order to know clearly the situation of the chunk reordering, the comparisons between the source reordering, monotone translation and the RWTH's best system are shown in Table 3. The "RWTH's best system" is described in Section 3.1, where the max-jump width is 7. We could observe that source reordering is much faster (The "Time" is for the whole test set.). But the BLEU score is worse. That could be explained by the inconsistency between chunks and phrases. Source reordering approach only reorder chunks, while not do reordering inside chunks because the local word reordering is included in phrase pairs. However, since the boundary of chunks and phrases could be cross each other, the local word reordering would be hurt.

The intention of the syntactic approach is to reorder some words over large distances. It is especially often happened in question sentences, in which question words like "where" and "when" are at the end of a sentence, unlike in English at the beginning of a sentence. In Table 4, some translation examples are listed. Besides the source and reference, the chunked source sentence and the alignments between the source and reference are

Table 2: Translation performance for the Chinese-English IWSLT task

test	WER[%]	PER[%]	TER[%]	BLEU[%]
baseline: source reorder	33.5	27.2	32.0	59.0
+ ruleProb	33.1	27.0	32.0	59.7
+ reordered train data	32.7	27.8	31.5	60.3

Table 3: Comparison with the RWTH best system

	BLEU[%]	TIME
monotone	56.0	14 sec.
RWTH-best-system	62.4	62 min.
source reorder improved	60.3	4 min.

Table 4: Translation Examples

source	我想要一个面向海滩的房间.
chunks	我_r 想_v 要_v 一个_m [VP 面向_v 海滩_n] 的_u 房间_n ._w
reference	I'd like a room facing the beach.
source reorder improved	i would like a room facing the beach .
RWTH-best-system	i would like a beach facing the room .
source	你拿到这些书了吗?
chunks	你_r [VRD 拿_v 到_v] 这些_r [NP 书_n 了_y] 吗_y ?_w
reference	Do you have these books available?
source reorder improved	do you have these books ?
RWTH-best-system	you have to book ?
source	有很多鱼的地方在哪?
chunks	有_v [NP 很多_m 鱼_n] 的_u 地方_n 在_p [NP 哪_r] ?_w
reference	What place has a lot of fish?
source reorder improved	where can i find a lot of fish ?
RWTH-best-system	there are many fish where ?
source	它将于什么时候结束?
chunks	它_r 将_d 于_p [NP 什么_r 时候_n] 结束_v ?_w
reference	At what time does it end?
source reorder improved	what time will it be over ?
RWTH-best-system	when will it be over ?

aslo given. We compare improved source reordering approach ("+reordered train data" in Table 2) to the RWTH's best system output. The chunk-reordering approach works better in this case of reordering question words.

6. Conclusion and future work

In this paper, chunk-based source reordering method has been improved by two methods, namely lattice weighting with the rules probability and reordered training data. Translation results were reported for IWSLT Chinese-English translation task. The total BLEU score improves 1.4%. In the next step, we would try to fix the gap between phrases and chunks. More analysis on the reordering rules are also necessary.

7. Acknowledgements

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Textübersetzung" (Ne572/5). The findings reported and the views expressed in this research are those of the author and do not necessarily reflect the position of the DARPA and DFG.

8. References

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–72, March 1996.

[2] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney, "Novel reordering approaches in phrase-based statistical machine translation," in *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, Ann Arbor, Michigan, June 2005, pp. 167–174.

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, Edmonton, Canada, May/June 2003, pp. 127–133.

[4] S. Nießen and H. Ney, "Morpho-syntactic analysis for reordering in statistical machine translation," in *Proc. MT Summit VIII*, Santiago de Compostela, Galicia, Spain, Sept. 2001, pp. 247–252.

[5] M. Popović and H. Ney, "POS-based word reorderings for statistical machine translation," in *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, 2006.

[6] M. R. Costa-jussà and J. A. R. Fonollosa, "Statistical machine reordering," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, Sydney, Australia, July 2006, pp. 70–76.

[7] M. Collins, P. Koehn, and I. Kucerova, "Clause restructuring for statistical machine translation," in *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, June 2005, pp. 531–540.

[8] C. Wang, M. Collins, and P. Koehn, "Chinese syntactic reordering for statistical machine translation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 737–745.

[9] C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan, "A probabilistic approach to syntax-based reordering for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 720–727.

[10] D. Zhang, M. Li, C.-H. Li, and M. Zhou, "Phrase reordering model integrating syntactic knowledge for SMT," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 533–540.

[11] C. Schafer and D. Yarowsky, "Statistical machine translation using coercive two-level syntac-

tic transduction," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 9–16.

[12] Y. Zhang, R. Zens, and H. Ney, "Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation," in *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*. Rochester, New York: Association for Computational Linguistics, April 2007, pp. 1–8.

[13] R. Zens, F. J. Och, and H. Ney, "Phrase-based statistical machine translation," in *25th German Conf. on Artificial Intelligence (KI2002)*, ser. Lecture Notes in Artificial Intelligence (LNAI), M. Jarke, J. Koehler, and G. Lakemeyer, Eds., vol. 2479. Aachen, Germany: Springer Verlag, September 2002, pp. 18–32.

[14] M. R. Costa-jussià, J. M. Crego, P. Lambert, M. Khalilov, J. A. R. Fonollosa, J. B. Mariño, and R. E. Banchs, "Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 167–170.

[15] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, June 1990.

[16] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pp. 295–302.

[17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[18] A. Mauser, D. Vilar, G. Leusch, Y. Zhang, and H. Ney, "The rwth statistical machine translation system for the iwslt 2007 evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation*, Trento, Italy, 2007.

[19] H.-P. Zhang, Q. Liu, X.-Q. Cheng, H. Zhang, and H.-K. Yu, "Chinese lexical analysis using hierarchical hidden markov model," in *Proc. of the second SIGHAN workshop on Chinese language processing*, Morristown, NJ, USA, 2003, pp. 63–70.

[20] M. Snover, L. M. Bonnie Dorr, Richard Schwartz, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of the 7th Conference of the Association for Machine Translation in the Americas*, 2006, pp. 223–231.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pp. 311–318.