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 novel methods to improve the reordering at the chunk level. By introducing a new lattice weighting factor, reordering the training source data and applying reordering during decoding, a significant improvement is reported on WER, PER, TER and BLEU. Compare to the previous chunk-level reordering approach, the BLEU score gains 2% improvement absolutely. The translation result is 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]. These syntax could be viewed as the word classification, since both morphology analysis and POSs are tags of a set of words. So they are particularly useful for the language with much morphologies for reducing the word sparseness. Another kind of syntax used for reordering is parse tree, 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 build 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 more close to the definition of "phrase" in phrase-based SMT and easy to use. We have not found much work to do reordering at the chunk level, only [11]. In that work, source sentences have been parsed into chunks. The chunk reordering is done by some rules learned automatically. All the reordered sentences are presented in a lattice.

In this paper, we will improve the approach by adding weights to the lattice with the rules probability. And not only the reordering on the source language side, we also report improvements by reordering both on the source language side and during decoding. Additionally, reordered training corpora can also improve the system.

2. Related work

In the previous chunk level reordering work, [11] 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 direct from the reordering rules is not used.

The earlist work to input a graph to SMT system was done by [12]. Another work with weighted graph is done by [13]. 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 [14]. 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 [15], we obtain:

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

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 [16].

The log linear model is a natural framework to integrate many models. During 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)

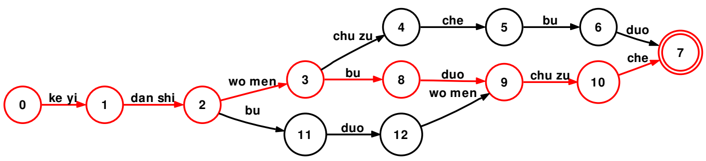
This is a standard phrase-based monotone SMT system, except the reorderings captured inside phrases, no particular reordering model is applied.

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 $\rightarrow$ VP NP
r NP VP $\rightarrow$ r VP NP
r NP VP $\rightarrow$ VP r NP

Reordering Lattice:



3.2. Chunking reordering system

The baseline reordering system we use was described in [11]. 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 POSs, "Inst. of Computing Tech., Chinese Lexical Analysis System.(ICTCLAS)" [17] is used. It can do word segmentation and Art-Of-Speech tagging integratively.

The lattice is weighted with a trigram reordered source language model. Each path of the lattice is a permutation $S'(S' = f_{\pi_1^J})$ for a given source sentence $S(S = f_1^J)$. The weight model used in the docoder is:

$$
\begin{aligned}
h(S'|S) &= \log(p(S')) && (5)\\
&= \log(p(f_{\pi_1^J}|f_1^J)) && (6)\\
&= \sum_j \log(p(f_{\pi_j}|f_{\pi_{j-1}}, f_{\pi_{j-2}})) && (7)
\end{aligned}
$$

4. Improve chunk reordering system

Three methods will be reported to improve the chunk reordering:

1. new model to weight the lattice.
2. add additional reordered training data.
3. beside the source reordering, a reordering model in decoding.

4.1. Lattice weighting

We expand the Equation(5) with reordering probability $p_reorder(S'|S)$, which is computed by the probabilities of the used reordering rules.

$$W(S'|S) = \lambda_1 \cdot \log(p(S'|S)) + \lambda_2 \cdot \log(p(S')) \quad (8)$$

λ_1 and λ_2 are the model scaling factors, which can be integrated into the Equation (4).

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

The reordering model $p(S'|S)$ could be viewed as a translation from original source sentence S to reordered sentence S'. S consists of a sequence of chunks: $S = c_1^N$. S' is a permutation of the chunks: $S' = \pi_1^N$. For a given S, a S' is generated with a sequence of reordering rules r_1^K. These rules segment source chunks c_1^N into k parts $\tilde{c_1}$... $\tilde{c_k}$. Similar to phrase-based translation model, we also 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 $\tilde{c_k}$ and its right hand side is the $\tilde{c_k}$'s permutation: $\tilde{\pi_k}$. So, $p(S'|S)$ can be represented as:

$$\begin{aligned} p(S'|S) &= \sum_B p(\pi_1^N, B|c_1^N) && (9)\\ &= \sum_B p(B|c_1^N) \cdot p(\pi_1^N|c_1^N, B) && (10)\\ &= \sum_B \alpha(c_1^N) \cdot p(\pi_1^N|c_1^N, B) && (11) \end{aligned}$$

$$\begin{aligned} p(\pi_1^N|c_1^N, B) &= p(\tilde{\pi}_1^K|\tilde{c}_1^K) && (12)\\ &= \prod_k p(\tilde{\pi}_k|\tilde{c}_k) && (13) \end{aligned}$$

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(14). It is calculated via relative frequencies.

$$p(\tilde{\pi}_k|\tilde{c}_k) = \frac{N(\tilde{\pi}_k, \tilde{c}_k)}{N(\tilde{c}_k)} \qquad (14)$$

4.2. Reordering training data

So far, the source reordering model is applied only on the test data. The training process and phrase extraction use the methods described in [12]. The training source data is not reordered at all, here we refer it to "monotone" source sentence. The test data is reordered. Then, for phrase extraction, there is mismatch between the monotone training source sentence and the reordered test source sentence. Some long phrases will be broken because of the new word order of the test data. It will affect the lexical choice during decoding.

In order to solve this problem, another GIZA++ [2] training is run on the reordered source training data. Here, we call the training on monotone source sentences as "monotone training"; the training on reordered source sentences as "reordered training". The reordered training result will work as an additional resource for the phrase extraction. The reordered source training data is generated by reorder the chunks according to the word alignment to make the source words to have similar word order of the target language.

4.3. Reordering model in decoding

What will happen if we not only reorder the source words, but also run some reordering model during search process? It is well known that distance-based distortion model is powerful to reorder phrases in short distance. It may still helpful for the whole system.

In this work, we apply IBM constraint [19] reordering in he beam search algorithm. Translation is done from the beginning of the sentence to the end. The target sentence is generated phrase by phrase. The constraints allow one of the first k still untranslated phrases to be selected. In the following experiments, a maximum of two phrases are allowed to be skipped at a time.

5. Experiments

5.1. Corpus Statistics

We perform translation experiments on the Basic Traveling Expression Corpus (BTEC) [20] 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	43k	
	Words	379k	420k
	Vocabulary	11944	9933
Dev dev2	Sentences	500	
	Words	3 557	3 908
	OOVs	75	–
Test dev3	Sentences	506	
	Words	3 813	3 970
	OOVs	72	–

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 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, we show the translation results for the BTEC task. First, we observe that the results get improvement overall evaluation criteria. Both on dev and test, the BLEU score which is optimized on improves about 2% absolutely.

The monotone translation is the standard phrase-based translation system as it is described in Section 3.1. Our baseline system is the chunk-level source reordering which is reported in [11]. More details are presented in Section 3.2. From the baseline, the scores of adding each technique are shown step by step in Table 2.

- “+ruleProb” uses the probabilities of the reordering rules to weight the reordering lattice.
- “+reordered train data” is the result of enlarging the training data by adding reordered source sentences. We have gotten better PER, since more reordered chunks could be found in the training data. This is seen on the test data, but not obviously on the development data.
- “+IBM(w=2)” is the results of applying IBM reordering with the constrain window size 2. The improvement with this method means that the source reordering rules have not translated well all the possible reorderings. On the other hand, it is good to know a system can use both source reordering and decoding reordering at a time. We could develop the two models respectively.

In Table 3, some translation examples are listed. The translation hypothesis with all the three techniques is shown in row “source reorder improved”.

6. Conclusion and Future Work

We have presented some improvements on the chunk-based source reordering method, namely

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

		WER[%]	PER[%]	TER[%]	BLEU[%]
dev	monotone	45.1	35.3	42.4	42.0
	source reorder baseline	46.0	36.0	43.3	48.4
	+ ruleProb	44.9	34.9	42.7	48.6
	+ reordered train data	44.4	35.0	42.3	49.5
	+ IBM (w=2)	43.9	35.2	42.0	50.5
test	monotone	39.9	31.6	38.0	47.5
	source reorder baseline	39.9	31.5	37.8	52.5
	+ ruleProb	40.1	31.7	38.1	53.3
	+ reordered train data	39.4	31.2	37.6	54.2
	+ IBM (w=2)	38.4	29.8	36.1	54.8

Table 3: Translation Examples

source	到 成田 机场 还要 多 长 时间 ?
reference	how much longer will it take to get to narita?
source reorder baseline	to narita how many more hours ?
source reorder improved	how long does it take to narita ?
source	你 今天 玩 得 开心 吗 ?
reference	did you have a good time today?
source reorder baseline	you had a good time for today ?
source reorder improved	do you have a good time today ?
source	有 对 孩子 们 的 折扣 吗 ?
reference	do you have discounts for children?
source reorder baseline	do you have for children discount ?
source reorder improved	do you have a discount for children ?
source	没有 处方 我 能 买 这个 吗 ?
reference	can I buy this without a prescription?
source reorder baseline	i do not have a prescription can buy this ?
source reorder improved	can i buy this without a prescription ?

lattice weighting with the rules probability, reordered training data and additional IBM reordering during decoding. Translation results were reported for IWSLT Chinese-English translation task. These three methods improve the baseline results step by step. It achieves overall 2% BLEU score on both dev and test data and results improvement on all evaluation criteria.

Since the reordering ability of the rules are still not clear, in the next step, we plan to do more analysis on the reordering rules, including manual analysis. We also plan to integrate chunks into phrase tables and model distortion reorderings at chunk level.

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)

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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] C. Tillmann and H. Ney, "Word reordering and a dynamic programming beam search algorithm for statistical machine translation," *Computational Linguistics*, vol. 29, no. 1, pp. 97–133, March 2003.

[19] A. L. Berger, P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. R. Gillett, A. S. Kehler, and R. L. Mercer, "Language translation apparatus and method of using context-based translation models, United States Patent 5510981," April 1996.

[20] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proc. of the Third Int. Conf. on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, May 2002, pp. 147–152.

[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.