A Topic-Based Reordering Model for Statistical Machine Translation

Xing Wang, Deyi Xiong*, Min Zhang, Yu Hong, and Jianmin Yao

Provincial Key Laboratory for Computer Information Processing Technology
Soochow University, Suzhou, China
xingwsuda@gmail.com,
{dyxiong,minzhang,yhong,jyao}@suda.edu.cn

Abstract. Reordering models are one of essential components of statistical machine translation. In this paper, we propose a topic-based reordering model to predict orders for neighboring blocks by capturing topic-sensitive reordering patterns. We automatically learn reordering examples from bilingual training data, which are associated with document-level and word-level topic information induced by LDA topic model. These learned reordering examples are used as evidences to train a topic-based reordering model that is built on a maximum entropy (MaxEnt) classifier. We conduct large-scale experiments to validate the effectiveness of the proposed topic-based reordering model on the NIST Chinese-to-English translation task. Experimental results show that our topic-based reordering model achieves significant performance improvement over the conventional reordering model using only lexical information.

Keywords: statistical machine translation, reordering model, topic information.

1 Introduction

In recent years, phrase-based SMT has been widely used. It segments a source sentence into a phrase sequence, then translates and reorders these phrases in the target. Phrase reordering in the target is critical issue for SMT [6]. A great variety of models have been proposed to address this issue. Lexical, syntactical and semantic information are explored to predicate phrase orientations. Unfortunately these models only focus on sentence-level information and neglect document-level information.

Document-level information has proved very useful in many NLP tasks. In SMT literature, some researchers utilize document level information to improve translation and language models [4,5,13,15]. The translation knowledge which appears in the preceding sentences' translations is used to guide translations of

* Corresponding author.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 414–421, 2014.

the current sentence as well as the succeeding sentences. With regard to reordering models, document-level information has also been studied but in an implicit fashion. [12] propose a topic-based similarity model for translation rule selection in hierarchical phrase-based SMT. They report that topic information is not only helpful for phrase and monotone rule selection but also for reordering rule selection. [2] also find that phrase reorderings are sensitive to domains where they occur and that reordering model adaptation can improve translation quality. These two studies suggest that document-level information can be used for phrase reordering. However, to the best of our knowledge, there is no attempt to directly incorporate document-level information into reordering models. In this paper, we investigate how document-level information, especially topic information can be explicitly integrated into reordering models.

We therefore propose a novel topic-based reordering model to improve phrase reordering for SMT by using document-level topic information. We employ two kinds of topic information in our topic-based reordering model: (1) document topic assignment, and (2) word topic assignment. We integrate our topic-based reordering model into a state-of-the-art phrase-based SMT system. We train the system with large-scale Chinese-English bilingual data. Experimental results on the NIST benchmark data show that our topic-based reordering model is able to achieve substantial performance improvement over the baseline. We also find that both document topic features and word topic features are able to improve phrase reordering and the combination of these two groups of features obtains the best performance.

The rest of the paper is organized as follows. In section 2, we present the topic-based reordering model with novel topic features and detail the training process of the model. The integration of the proposed model into SMT system is presented in the section 3. We describe our experiments, including the details on experimental setup and present the experimental results in section 4. In section 5, we provide a brief analysis on the results. Finally, we conclude the paper in section 6.

2 Topic-Based Reordering Model

In this section, we present the proposed topic-based reordering model, different features that are used in our topic-based reordering model as well as the training procedure of the model.

2.1 Reordering Model

We discuss our topic-based reordering model under the ITG constraints [11]. The following three Bracketing Transduction Grammar (BTG) rules are used to constrain translation and reordering.

$$A \to [A^1, A^2] \tag{1}$$

$$A \to \langle A^1, A^2 \rangle \tag{2}$$

$$A \to f/e \tag{3}$$

where A is a block which consists of a pair of source and target strings. A^1 and A^2 are two consecutive blocks. The rule (1) and rule (2) are reordering rules which are used to merge two blocks into a larger block in a straight or inverted order. Rule (3) is a lexical rule that translates a source phrase f into a target phrase e and correspondingly generates a block A. A maximum entropy (MaxEnt) reordering model is proposed by [14] to predict orders for neighboring blocks under the ITG constraints. Following them, we also use a maximum entropy classifier to predict ITG orientation $o \in \{monotonous, inverted\}$ for two neighboring blocks A^1 and A^2 as MaxEnt is able to handle arbitrary and overlapping features from different knowledge sources. The classifier can be formulated as follows:

$$p(o|C(A^1, A^2)) = \frac{\exp(\sum_i \theta_i f_i(o, C(A^1, A^2)))}{\sum_{o'} \exp(\sum_i \theta_i f_i(o', C(A^1, A^2)))} \tag{4}$$

where $C(A^1, A^2)$ indicates the attributes of A^1 and A^2, f_i are binary features, θ_i are weights of these features.

2.2 Features

We integrate three kinds of features into the topic-based reordering model:1) boundary word features; 2) document topic features (DT) and 3) word topic features (WT). All these features are in the following binary form:

$$f_i(o, C(A^1, A^2)) = \begin{cases} 1, & if\ (o, C(A^1, A^2))\ satisfies\ certain\ condition \\ 0, & else \end{cases} \tag{5}$$

Boundary Word Features: Following [14], we adopt boundary words as our basic features.

Document Topic Features: Given the document topic distribution inferred by LDA[1], we first choose the topic with the maximum probability to be the document topic. Then we use this topic as the document topic feature.

Word Topic Features: Topics of boundary words on the source side can be also used as features to capture topic-sensitive reordering patterns. Unfortunately, most of boundary words are function words or stop words (the percentage is 44.67% when we set the topic number 30). These words normally distribute evenly over all inferred topics and therefore their topics can not be used as indicators to distinguish topic-sensitive orientations. We believe that the integration of topics of boundary function words may jeopardize translation quality.

In order to address this issue, we use the topic assignments of content words that locate at the left/rightmost positions on the source phrases in question.

2.3 Training

In order to train the topic-based reordering model, topic distributions of source language documents from bilingual corpora are estimated by LDA tools. Training instances are generated and associated with their corresponding topic information. Given the features extracted from these training instances, we use a maximum entropy toolkit[1] to train the maximum entropy classifier (i.e., the topic-based reordering model). We perform 100 iterations of L-BFGS algorithm implemented in the training toolkit with both Gaussian prior and event cutoff set to 1 to avoid over-fitting.

3 Decoding

In this section, we integrate the proposed reordering model into SMT system. The log-linear model as described in [9] is adopted to combine various sub-models for obtaining the best translation, which is formulated as follows.

$$e_{best} = \operatorname*{argmax}_{e} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right\} \tag{6}$$

where h_m are sub-models or features of the whole log-linear model, λ_m are their weights. We apply the topic-based reordering model as one sub-model to the log-linear model to constrain the phrase reordering, and tune the wight of the sub-model on development set.

We infer the topic distributions of the given test documents before translation according to the topic model trained on the corpora with document boundaries.

4 Experiments

In this section, we present our experiments on Chinese-to-English translation tasks with large-scale training data. With the aim of evaluating the effectiveness of the novel topic-based reordering model, we carried out a series of experiments with different configurations.

4.1 Setup

We adopted a state-of-the-art BTG-based phrasal system with a CKY-style decoder [14] as our baseline system and integrated the topic-based reordering model into this system.

Our training data consists of 2.8M sentence pairs with 81M Chinese words and 86M English words (including punctuation tokens) from LDC data[2]. There are 10, 326 documents in the training data. We chose NIST MT Evaluation test set 2003 (MT03) as our development set, MT02, MT04 as our test sets. The number of documents/sentences in the development and test sets are listed in Table 1.

1 Available at `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

2 The corpora include LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08(Hong Kong Hansards/Law/News).

Table 1. The number of documents/sentences in the development set and test sets

	MT02	MT 03	MT04
The numbers of documents	100	100	200
The numbers of sentences	878	919	1788

Table 2. BLEU scores of the topic-based reordering model using document-level topic features on the development set with topic number K varying from 10 to 50

	DevSet(MT03)
Baseline	35.28
K = 10	35.48
K = 30	35.76
K = 50	35.39

We obtained the word alignments by running GIZA++ [9] on the training data in both directions and applying the "grow-diag-final-and" refinement [7]. We applied SRI Language Modeling Toolkit[3] [8] to train a 5-gram model with Kneser-Ney smoothing on the Xinhua portion of English Gigaword corpus. Case-insensitive BLEU-4 [10] was used as our evaluation metric. In order to alleviate the impact of the instability of MERT, we run it three times for all our experiments and present the average BLEU scores on the three runs following the suggestion by [4].

To train the topic-based reordering model, we used the open source software tool GibbsLDA++[4] for topic estimation and inference. GibbsLDA++ is an implementation of LDA using Gibbs Sampling technique for parameter estimation and inference. We removed Chinese stop words (1, 205 words in total) and rare words (54, 073 words in total) before topic estimation. Then we set the number of topic K, and used the default setting of the tool for training and inference. Document and word topic assignments for training sentences were obtained according to topic distributions. Finally, 52M training instances were generated to train the topic-based reordering model.

4.2 Impact of the Number of Topics

We carried out the first group of experiments to study the influence of the number of topics K on our topic-based reordering model. In these experiments, the topic-based reordering model was trained on the corpora with document boundaries. Only boundary words and document topic features are used in the trained reordering model. The results are shown in Table 2.

From Table 2, we can observe that we gain an improvement of 0.28 BLEU points when the number of topics K increases from 10 to 30. However a noticeable drop in the BLEU scores (-0.37 BLEU points) is observed, when K further

[3] Available at `http://www.speech.sri.com/projects/srilm/download.html`

[4] Available at `http://sourceforge.net/projects/gibbslda/`

Table 3. BLEU scores of topic-based reordering model with different features on the test sets (DT : document topic features; WT : word topic features)

	MT02	MT04	Avg
Baseline	33.90	34.86	34.38
Baseline + DT	34.11	35.52	34.82
Baseline + WT	34.09	35.05	34.57
Baseline + DT + WT	**34.40**	**35.96**	**35.18**

increases to 50. Therefore, the number of topics K is set to 30 in all experiments hereafter.

4.3 Effect of Different Topic Features

We ran our second group of experiments to investigate the effects of using different features in the topic-based reordering model. All reordering models were trained over the corpora with document boundaries. The reordering model of *Baseline* only uses boundary words as features.

The results of the topic-based reordering model compared with the conventional reordering model are presented in Table 3. As can be seen in the table, improvements are achieved by integrating topic features into the reordering model. Adding document topic features (+DT) and word topic features (+WT) leads to an average improvement of 0.44 BLEU points and 0.19 BLEU points respectively. Using all topic features(+DT+WT), our model achieves an improvement of up to 1.1 BLEU points on the MT04 test set, and an average improvement of 0.8 BLEU points over the *Baseline* on the two test sets. These improvements suggest that topic features are indeed useful for phrase reordering.

5 Analysis

In this section, we conduct a brief analysis of the results shown in the previous section, to take a deeper look into why the topic-based reordering model improves translation performance.

Table 4 shows an example which compares the translation generated by the *Baseline* to that generated by the *Baseline+DT+WT* system. By checking the translation process of the source sentence, we find that the baseline system incorrectly merges two phrases ("CEPA" and "to the legal profession in Hong Kong") and makes a wrong reordering prediction. Under the guidance of topic information, our system avoids making the same error.

Furthermore, we view the reordering problem from a global perspective. The Chinese word "比/bǐ" which means "make a comparison" occurs frequently. The phrases to the right of the word "比/bǐ"are always moved to the left of the corresponding target word after translation. In other words, inverted orientation is often used when the source word "比/bǐ" occurs in documents with topic on economic performance comparison. However, when the word occur in documents

Table 4. A Chinese to English translation example showing the difference between the baseline and the system using our topic-based reordering model

src	(港/gǎng 澳/ào 台/tái) CEPA **为/wěi** 香港/xiānggǎng 法律/fǎlǜ 界/jiè 打开/dǎkāi 新/xīn 局面/júmiàn
Baseline	(Hong Kong , Macao and Taiwan) **to** the legal profession in Hong Kong CEPA opens new situation
Baseline + DT + WT	(Hong Kong , Macao and Taiwan) CEPA opens new prospects **to** the legal profession in Hong Kong
ref	(Hong Kong , Macao and Taiwan) the CEPA opens up new prospects **for** Hong Kong legal fields

Table 5. Examples of different reordering phenomena in different topics, where phrases in brackets with the same index are translation of each other

Topic about economy	src	... (比/$b\check{\imath}$)^1 (5月份/5$yu\grave{e}f\grave{e}n$)^2 (下降/$xi\grave{a}ji\grave{a}ng$ 3.8%)^3...
	trg	... (down 3.8%)^3 (from)^1 (May)^2 ...
Topic about sport	src	... (五/$w\check{u}$)^4 (比/$b\check{\imath}$)^5 (一/$y\bar{\imath}$)^6 ...
	trg	... (five)^4 (to)^5 (one)^6 ...

with a topic on sport, the inversion phenomena disappear because the word sense changes into "game score".

Table 5 displays two examples where the word "比/bǐ"with different topics has different reorderings in bilingual data. This suggests that phrase reorderings are sensitive to topic and training instances with document-level topic features are capable of capturing topic-sensitive reordering patterns.

These examples and analysis empirically explain why reordering is topic-sensitive and our experimental results in Section 5 convincingly demonstrate that our proposed solution is effective in capturing topic information for phrase reordering.

6 Conclusions

In this paper, we have presented a topic-based reordering model to incorporate topic information into reordering model. To capture topic-sensitive reordering patterns, two kinds of topic information, namely the document-level topic and word-level topic, are used in our reordering model. Experimental results show that our model achieves substantial performance improvement over the baseline. This demonstrates the advantage of exploiting topic information for phrase reordering. Finally, we investigate why incorporating topic information into reordering model improves SMT performance.

In future, instead of only using the document topic with the maximum probability, we would like to utilize topic distribution to represent document topic features. Furthermore, we plan to explore more document-level information to predict orders for neighboring blocks. Finally, we are also interested in investigating new methods to infer topics for sentences with no document boundaries.

Acknowledgement. The authors are supported by National Natural Science Foundation of China (No. 61373095,61373097,61272259) and Natural Science Foundation of Jiangsu Province (No. BK20140355).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research, 601–608 (2002)
2. Chen, B., Foster, G., Kuhn, R.: Adaptation of reordering models for statistical machine translation. In: Proceedings of NAACL-HLT, pp. 938–946 (2013)
3. Clark, J.H., Dyer, C., Lavie, A., Smith, N.A.: Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In: Proceedings of ACL, pp. 176–181 (2011)
4. Gong, Z., Zhang, M., Zhou, G.: Cache-based document-level statistical machine translation. In: Proceedings of EMNLP, pp. 909–919 (2011)
5. Gong, Z., Zhang, M., Tan, C., Zhou, G.: N-gram-based tense models for statistical machine translation. In: Proceedings of EMNLP, pp. 276–285 (2012)
6. Knight, K.: Decoding complexity in word-replacement translation models. Computational Linguistics 25(4), 607–615 (1999)
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of NAACL-HLT, pp. 48–54 (2003)
8. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of INTERSPEECH (2002)
9. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
11. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics 23(3), 377–403 (1997)
12. Xiao, X., Xiong, D., Zhang, M., Liu, Q., Lin, S.: A topic similarity model for hierarchical phrase-based translation. In: Proceedings of ACL, pp. 750–758 (2012)
13. Xiong, D., Ben, G., Zhang, M., Lü, Y., Liu, Q.: Modeling lexical cohesion for document-level machine translation. In: Proceedings of IJCAI, pp. 2183–2189 (2013)
14. Xiong, D., Liu, Q., Lin, S.: Maximum entropy based phrase reordering model for statistical machine translation. In: Proceedings of ACL, pp. 521–528 (2006)
15. Xiong, D., Zhang, M.: A topic-based coherence model for statistical machine translation. In: Proceedings of AAAI, pp. 2183–2189 (2013)