A Topic Similarity Model for Hierarchical Phrase-based Translation

Xinyan Xiao[†] **Deyi Xiong**[‡] **Min Zhang**[‡*] **Qun Liu**[†] **Shouxun Lin**[†]

[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan, liuqun, sxlin}@ict.ac.cn

[‡]Human Language Technology
Institute for Infocomm Research
{dyxiong, mzhang*}@i2r.a-star.edu.sg

Abstract

Previous work using topic model for statistical machine translation (SMT) explore topic information at the word level. However, SMT has been advanced from word-based paradigm to phrase/rule-based paradigm. We therefore propose a topic similarity model to exploit topic information at the synchronous rule level for hierarchical phrase-based translation. We associate each synchronous rule with a topic distribution, and select desirable rules according to the similarity of their topic distributions with given documents. We show that our model significantly improves the translation performance over the baseline on NIST Chinese-to-English translation experiments. Our model also achieves a better performance and a faster speed than previous approaches that work at the word level.

1 Introduction

Topic model (Hofmann, 1999; Blei et al., 2003) is a popular technique for discovering the underlying topic structure of documents. To exploit topic information for statistical machine translation (SMT), researchers have proposed various topic-specific lexicon translation models (Zhao and Xing, 2006; Zhao and Xing, 2007; Tam et al., 2007) to improve translation quality.

Topic-specific lexicon translation models focus on word-level translations. Such models first estimate word translation probabilities conditioned on topics, and then adapt lexical weights of phrases by these probabilities. However, the state-of-the-art SMT systems translate sentences by using sequences of synchronous rules or phrases, instead of translating word by word. Since a synchronous rule is rarely factorized into individual words, we believe that it is more reasonable to incorporate the topic model directly at the rule level rather than the word level.

Consequently, we propose a **topic similarity** model for hierarchical phrase-based translation (Chiang, 2007), where each synchronous rule is associated with a topic distribution. In particular,

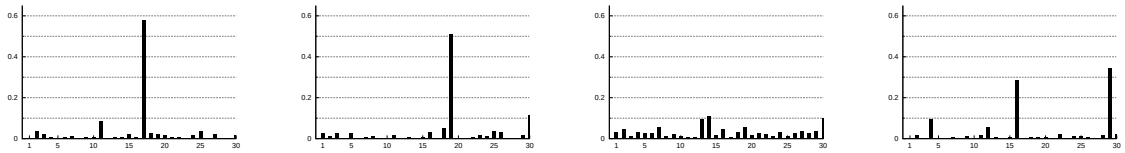
- Given a document to be translated, we calculate the topic similarity between a rule and the document based on their topic distributions. We augment the hierarchical phrase-based system by integrating the proposed topic similarity model as a new feature (Section 3.1).

- As we will discuss in Section 3.2, the similarity between a generic rule and a given source document computed by our topic similarity model is often very low. We don't want to penalize these generic rules. Therefore we further propose a topic sensitivity model which rewards generic rules so as to complement the topic similarity model.

- We estimate the topic distribution for a rule based on both the source and target side topic models (Section 4.1). In order to calculate similarities between target-side topic distributions of rules and source-side topic distributions of given documents during decoding, we project

*Corresponding author

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 750–758,
Jeju, Republic of Korea, 8-14 July 2012.



(a) 作战 能力 ⇒ operational capability (b) 给予 X_1 ⇒ grands X_1 (c) 给予 X_1 ⇒ give X_1 (d) X_1 举行 会谈 X_2 ⇒ held talks X_1 X_2

Figure 1: Four synchronous rules with topic distributions. Each sub-graph shows a rule with its topic distribution, where the X-axis means topic index and the Y-axis means the topic probability. Notably, the rule (b) and rule (c) shares the same source Chinese string, but they have different topic distributions due to the different English translations.

the target-side topic distributions of rules into the space of source-side topic model by one-to-many projection (Section 4.2).

Experiments on Chinese-English translation tasks (Section 6) show that, our method outperforms the baseline hierarchial phrase-based system by +0.9 BLEU points. This result is also +0.5 points higher and 3 times faster than the previous topic-specific lexicon translation method. We further show that both the source-side and target-side topic distributions improve translation quality and their improvements are complementary to each other.

2 Background: Topic Model

A topic model is used for discovering the topics that occur in a collection of documents. Both Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) are types of topic models. LDA is the most common topic model currently in use, therefore we exploit it for mining topics in this paper. Here, we first give a brief description of LDA.

LDA views each document as a mixture proportion of various topics, and generates each word by multinomial distribution conditioned on a topic. More specifically, as a generative process, LDA first samples a document-topic distribution for each document. Then, for each word in the document, it samples a topic index from the document-topic distribution and samples the word conditioned on the topic index according the topic-word distribution.

Generally speaking, LDA contains two types of parameters. The first one relates to the document-topic distribution, which records the topic distribution of each document. The second one is used for topic-word distribution, which represents each topic as a distribution over words. Based on these parameters (and some hyper-parameters), LDA can infer a topic assignment for each word in the documents. In the following sections, we will use these parameters and the topic assignments of words to estimate the parameters in our method.

3 Topic Similarity Model

Sentences should be translated in consistence with their topics (Zhao and Xing, 2006; Zhao and Xing, 2007; Tam et al., 2007). In the hierarchical phrase based system, a synchronous rule may be related to some topics and unrelated to others. In terms of probability, a rule often has an uneven probability distribution over topics. The probability over a topic is high if the rule is highly related to the topic, otherwise the probability will be low. Therefore, we use topic distribution to describe the relatedness of rules to topics.

Figure 1 shows four synchronous rules (Chiang, 2007) with topic distributions, some of which contain nonterminals. We can see that, although the source part of rule (b) and (c) are identical, their topic distributions are quite different. Rule (b) contains a highest probability on the topic about "China-U.S. relationship", which means rule (b) is much more related to this topic. In contrast, rule (c) contains an even distribution over various topics. Thus, given a document about "China-U.S. relationship", we hope to encourage the system to apply rule (b) but penalize the application of rule (c). We achieve this by calculating similarity between the topic distributions of a rule and a document to be translated.

More formally, we associate each rule with a **rule-topic distribution** $P(z|r)$, where r is a rule, and z is a topic. Suppose there are K topics, this distribution

can be represented by a K-dimension vector. The k-th component $P(z = k|r)$ means the probability of topic k given the rule r. The estimation of such distribution will be described in Section 4.

Analogously, we represent the topic information of a document d to be translated by a **document-topic distribution** $P(z|d)$, which is also a K-dimension vector. The k-th dimension $P(z = k|d)$ means the probability of topic k given document d. Different from rule-topic distribution, the document-topic distribution can be directly inferred by an off-the-shelf LDA tool.

Consequently, based on these two distributions, we select a rule for a document to be translated according to their **topic similarity** (Section 3.1), which measures the relatedness of the rule to the document. In order to encourage the application of generic rules which are often penalized by our similarity model, we also propose a topic sensitivity model (Section 3.2).

3.1 Topic Similarity

By comparing the similarity of their topic distributions, we are able to decide whether a rule is suitable for a given source document. The topic similarity computes the distance of two topic distributions. We calculate the topic similarity by Hellinger function:

$$\begin{aligned} &Similarity(P(z|d), P(z|r)) \\ &= \sum_{k=1}^{K} \left(\sqrt{P(z = k|d)} - \sqrt{P(z = k|r)}\right)^2 \end{aligned} \quad (1)$$

Hellinger function is used to calculate distribution distance and is popular in topic model (Blei and Lafferty, 2007).[1] By topic similarity, we aim to encourage or penalize the application of a rule for a given document according to their topic distributions, which then helps the SMT system make better translation decisions.

3.2 Topic Sensitivity

Domain adaptation (Wu et al., 2008; Bertoldi and Federico, 2009) often distinguishes general-domain data from in-domain data. Similarly, we divide the rules into topic-insensitive rules and topic-sensitive rules according to their topic distributions. Let's revisit Figure 1. We can easily find that the topic distribution of rule (c) distribute evenly. This indicates that it is insensitive to topics, and can be applied in any topics. We call such a rule a topic-insensitive rule. In contrast, the distributions of the rest rules peak on a few topics. Such rules are called topic-sensitive rules. Generally speaking, a topic-insensitive rule has a fairly flat distribution, while a topic-sensitive rule has a sharp distribution.

A document typically focuses on a few topics, and has a sharp topic distribution. In contrast, the distribution of topic-insensitive rule is fairly flat. Hence, a topic-insensitive rule is always less similar to documents and is punished by the similarity function.

However, topic-insensitive rules may be more preferable than topic-sensitive rules if neither of them are similar to given documents. For a document about the "military" topic, the rule (b) and (c) in Figure 1 are both dissimilar to the document, because rule (b) relates to the "China-U.S. relationship" topic and rule (c) is topic-insensitive. Nevertheless, since rule (c) occurs more frequently across various topics, it may be better to apply rule (c).

To address such issue of the topic similarity model, we further introduce a topic sensitivity model to describe the topic sensitivity of a rule using entropy as a metric:

$$\begin{aligned} &Sensitivity(P(z|r)) \\ &= -\sum_{k=1}^{K} P(z = k|r) \times \log\left(P(z = k|r)\right) \end{aligned} \quad (2)$$

According to the Eq. (2), a topic-insensitive rule has a large entropy, while a topic-sensitive rule has a smaller entropy. By incorporating the topic sensitivity model with the topic similarity model, we enable our SMT system to balance the selection of these two types of rules. Given rules with approximately equal values of Eq. (1), we prefer topic-insensitive rules.

4 Estimation

Unlike document-topic distribution that can be directly learned by LDA tools, we need to estimate the rule-topic distribution according to our requirement. In this paper, we try to exploit the topic information

[1] We also try other distance functions, including Euclidean distance, Kullback-Leibler divergence and cosine function. They produce similar results in our preliminary experiments.

of both source and target language. To achieve this goal, we use both source-side and target-side monolingual topic models, and learn the correspondence between the two topic models from word-aligned bilingual corpus.

Specifically, we use two types of rule-topic distributions: one is source-side rule-topic distribution and the other is target-side rule-topic distribution. These two rule-topic distributions are estimated by corresponding topic models in the same way (Section 4.1). Notably, only source language documents are available during decoding. In order to compute the similarity between the target-side topic distribution of a rule and the source-side topic distribution of a given document， we need to project the target-side topic distribution of a synchronous rule into the space of the source-side topic model (Section 4.2).

A more principle way is to learn a bilingual topic model from bilingual corpus (Mimno et al., 2009). However, we may face difficulty during decoding, where only source language documents are available. It requires a marginalization to infer the monolingual topic distribution using the bilingual topic model. The high complexity of marginalization prohibits such a summation in practice. Previous work on bilingual topic model avoid this problem by some monolingual assumptions. Zhao and Xing (2007) assume that the topic model is generated in a monolingual manner, while Tam et al., (2007) construct their bilingual topic model by enforcing a one-to-one correspondence between two monolingual topic models. We also estimate our rule-topic distribution by two monolingual topic models, but use a different way to project target-side topics onto source-side topics.

4.1 Monolingual Topic Distribution Estimation

We estimate rule-topic distribution from word-aligned bilingual training corpus with document boundaries explicitly given. The source and target side distributions are estimated in the same way. For simplicity, we only describe the estimation of source-side distribution in this section.

The process of rule-topic distribution estimation is analogous to the traditional estimation of rule translation probability (Chiang, 2007). In addition to the word-aligned corpus, the input for estimation also contains the source-side topic-document distribution of every documents inferred by LDA tool.

We first extract synchronous rules from training data in a traditional way. When a rule r is extracted from a document d with topic distribution $P(z|d)$, we collect an instance $(r, P(z|d), c)$, where c is the fraction count of an instance as described in Chiang, (2007). After extraction, we get a set of instances $\mathcal{I} = \{(r, P(z|d), c)\}$ with different document-topic distributions for each rule. Using these instances, we calculate the topic probability $P(z = k|r)$ as follows:

$$P(z=k|r) = \frac{\sum_{I\in\mathcal{I}} c \times P(z=k|d)}{\sum_{k'=1}^{K}\sum_{I\in\mathcal{I}} c \times P(z=k'|d)} \quad (3)$$

By using both source-side and target-side document-topic distribution, we obtain two rule-topic distributions for each rule in total.

4.2 Target-side Topic Distribution Projection

As described in the previous section, we also estimate the target-side rule-topic distribution. However, only source document-topic distributions are available during decoding. In order to calculate the similarity between the target-side rule-topic distribution of a rule and the source-side document-topic distribution of a source document, we need to project target-side topics into the source-side topic space. The projection contains two steps:

- In the first step, we learn the topic-to-topic correspondence probability $p(z_f|z_e)$ from target-side topic z_e to source-side topic z_f.
- In the second step, we project the target-side topic distribution of a rule into source-side topic space using the correspondence probability.

In the first step, we estimate the correspondence probability by the co-occurrence of the source-side and the target-side topic assignment of the word-aligned corpus. The topic assignments are output by LDA tool. Thus, we denotes each sentence pair by $(\mathbf{z_f}, \mathbf{z_e}, \mathbf{a})$, where $\mathbf{z_f}$ and $\mathbf{z_e}$ are the topic assignments of source-side and target-side sentences respectively, and $\mathbf{a}$ is a set of links $\{(i, j)\}$. A link (i, j) means a source-side position i aligns to a target-side position j. Thus, the co-occurrence of a source-side topic with index k_f and a target-side

e-topic	f-topic 1	f-topic 2	f-topic 3
enterprises	农业(agricultural)	企业(enterprise)	发展(develop)
rural	农村(rural)	市场(market)	经济(economic)
state	农民(peasant)	国有(state)	科技(technology)
agricultural	改革(reform)	公司(company)	我国(China)
market	财政(finance)	金融(finance)	技术(technique)
reform	社会(social)	银行(bank)	产业(industry)
production	保障(safety)	投资(investment)	结构(structure)
peasants	调整(adjust)	管理(manage)	创新(innovation)
owned	政策(policy)	改革(reform)	加快(accelerate)
enterprise	收入(income)	经营(operation)	改革(reform)
$p(z_f\|z_e)$	0.38	0.28	0.16

Table 1: Example of topic-to-topic correspondence. The last line shows the correspondence probability. Each column means a topic represented by its top-10 topical words. The first column is a target-side topic, while the rest three columns are source-side topics.

topic k_e is calculated by:

$$\sum_{(\mathbf{z_f},\mathbf{z_e},\mathbf{a})} \sum_{(i,j)\in\mathbf{a}} \delta(z_{f_i}, k_f) * \delta(z_{e_j}, k_e) \qquad (4)$$

where $\delta(x, y)$ is the Kronecker function, which is 1 if $x = y$ and 0 otherwise. We then compute the probability of $P(z = k_f|z = k_e)$ by normalizing the co-occurrence count. Overall, after the first step, we obtain an correspondence matrix $\mathbf{M}_{K_e\times K_f}$ from target-side topic to source-side topic, where the item $M_{i,j}$ represents the probability $P(z_f = i|z_e = j)$.

In the second step, given the correspondence matrix $\mathbf{M}_{K_e\times K_f}$, we project the target-side rule-topic distribution $P(z_e|r)$ to the source-side topic space by multiplication as follows:

$$T(P(z_e|r)) = P(z_e|r) \otimes \mathbf{M}_{K_e\times K_f} \qquad (5)$$

In this way, we get a second distribution for a rule in the source-side topic space, which we called projected target-side topic distribution $T(P(z_e|r))$.

Obviously, our projection method allows one target-side topic to align to multiple source-side topics. This is different from the one-to-one correspondence used by Tam et al., (2007). From the training result of the correspondence matrix $\mathbf{M}_{K_e\times K_f}$, we find that the topic correspondence between source and target language is not necessarily one-to-one. Typically, the probability $P(z = k_f|z = k_e)$ of a target-side topic mainly distributes on two or three source-side topics. Table 1 shows an example of a target-side topic with its three mainly aligned source-side topics.

5 Decoding

We incorporate our topic similarity model as a new feature into a traditional hiero system (Chiang, 2007) under discriminative framework (Och and Ney, 2002). Considering there are a source-side rule-topic distribution and a projected target-side rule-topic distribution, we add four features in total:

- $Similarity\,(P(z_f|d), P(z_f|r))$
- $Similarity(P(z_f|d), T(P(z_e|r)))$
- $Sensitivity(P(z_f|r))$
- $Sensitivity(T(P(z_e|r))$

To calculate the total score of a derivation on each feature listed above during decoding, we sum up the correspondent feature score of each applied rule.[2]

The source-side and projected target-side rule-topic distribution are calculated before decoding. During decoding, we first infer the topic distribution $P(z_f|d)$ for a given document on source language. When applying a rule, it is straightforward to calculate these topic features. Obviously, the computational cost of these features is rather small.

In the topic-specific lexicon translation model, given a source document, it first calculates the topic-specific translation probability by normalizing the entire lexicon translation table, and then adapts the lexical weights of rules correspondingly. This makes the decoding slower. Therefore, comparing with the previous topic-specific lexicon translation method, our method provides a more efficient way for incorporating topic model into SMT.

6 Experiments

We try to answer the following questions by experiments:

1. Is our topic similarity model able to improve translation quality in terms of BLEU? Furthermore, are source-side and target-side rule-topic distributions complementary to each other?

[2]Since glue rule and rules of unknown words are not extracted from training data, here, we just ignore the calculation of the four features for them.

System	MT06	MT08	Avg	Speed
Baseline	30.20	21.93	26.07	12.6
TopicLex	30.65	22.29	26.47	3.3
SimSrc	30.41	**22.69**	26.55	11.5
SimTgt	30.51	22.39	26.45	11.7
SimSrc+SimTgt	30.73	**22.69**	**26.71**	11.2
Sim+Sen	**30.95**	**22.92**	**26.94**	10.2

Table 2: Result of our topic similarity model in terms of BLEU and speed (words per second), comparing with the traditional hierarchical system ("Baseline") and the topic-specific lexicon translation method ("TopicLex"). "SimSrc" and "SimTgt" denote similarity by source-side and target-side rule-distribution respectively, while "Sim+Sen" activates the two similarity and two sensitivity features. "Avg" is the average BLEU score on the two test sets. Scores marked in bold mean significantly (Koehn, 2004) better than *Baseline* ($p < 0.01$).

2. Is it helpful to introduce the topic sensitivity model to distinguish topic-insensitive and topic-sensitive rules?

3. Is it necessary to project topics by one-to-many correspondence instead of one-to-one correspondence?

4. What is the effect of our method on various types of rules, such as phrase rules and rules with non-terminals?

6.1 Data

We present our experiments on the NIST Chinese-English translation tasks. The bilingual training data contains 239K sentence pairs with 6.9M Chinese words and 9.14M English words, which comes from the FBIS portion of LDC data. There are 10,947 documents in the FBIS corpus. The monolingual data for training English language model includes the Xinhua portion of the GIGAWORD corpus, which contains 238M English words. We used the NIST evaluation set of 2005 (MT05) as our development set, and sets of MT06/MT08 as test sets. The numbers of documents in MT05, MT06, MT08 are 100, 79, and 109 respectively.

We obtained symmetric word alignments of training data by first running GIZA++ (Och and Ney, 2003) in both directions and then applying refinement rule "grow-diag-final-and" (Koehn et al., 2003). The SCFG rules are extracted from this word-aligned training data. A 4-gram language model was trained on the monolingual data by the SRILM toolkit (Stolcke, 2002). Case-insensitive NIST BLEU (Papineni et al., 2002) was used to measure translation performance. We used minimum error rate training (Och, 2003) for optimizing the feature weights.

For the topic model, we used the open source LDA tool GibbsLDA++ for estimation and inference.[3] GibssLDA++ is an implementation of LDA using gibbs sampling for parameter estimation and inference. The source-side and target-side topic models are estimated from the Chinese part and English part of FBIS corpus respectively. We set the number of topic $K = 30$ for both source-side and target-side, and use the default setting of the tool for training and inference.[4] During decoding, we first infer the topic distribution of given documents before translation according to the topic model trained on Chinese part of FBIS corpus.

6.2 Effect of Topic Similarity Model

We compare our method with two baselines. In addition to the traditional hiero system, we also compare with the topic-specific lexicon translation method in Zhao and Xing (2007). The lexicon translation probability is adapted by:

$$p(f|e, D_F) \propto p(e|f, D_F)P(f|D_F) \quad (6)$$

$$= \sum_k p(e|f, z = k)p(f|z = k)p(z = k|D_F) \quad (7)$$

However, we simplify the estimation of $p(e|f, z = k)$ by directly using the word alignment corpus with

[3] http://gibbslda.sourceforge.net/

[4] We determine K by testing {15, 30, 50, 100, 200} in our preliminary experiments. We find that $K = 30$ produces a slightly better performance than other values.

Type	Count	Src%	Tgt%
Phrase-rule	3.9M	83.4	84.4
Monotone-rule	19.2M	85.3	86.1
Reordering-rule	5.7M	85.9	86.8
All-rule	28.8M	85.1	86.0

Table 3: Percentage of topic-sensitive rules of various types of rule according to source-side ("Src") and target-side ("Tgt") topic distributions. Phrase rules are fully lexicalized, while monotone and reordering rules contain nonterminals (Section 6.5).

topic assignment that is inferred by the GibbsLDA++. Despite the simplification of estimation, the improvement of our implementation is comparable with the improvement in Zhao et al.,(2007). Given a new document, we need to adapt the lexical translation weights of the rules based on topic model. The adapted lexicon translation model is added as a new feature under the discriminative framework.

Table 2 shows the result of our method comparing with the traditional system and the topic-lexicon specific translation method described as above. By using all the features (last line in the table), we improve the translation performance over the baseline system by 0.87 BLEU point on average. Our method also outperforms the topic-lexicon specific translation method by 0.47 points. This verifies that topic similarity model can improve the translation quality significantly.

In order to gain insights into why our model is helpful, we further investigate how many rules are topic-sensitive. As described in Section 3.2, we use entropy to measure the topic sensitivity. If the entropy of a rule is smaller than a certain threshold, then the rule is topic sensitive. Since documents often focus on some topics, we use the average entropy of document-topic distribution of all training documents as the threshold. We compare both source-side and target-side distribution shown in Table 3. We find that more than 80 percents of the rules are topic-sensitive, thus provides us a large space to improve the translation by exploiting topics.

We also compare these methods in terms of the decoding speed (words/second). The baseline translates 12.6 words per second, while the topic-specific lexicon translation method only translates 3.3 words in one second. The overhead of the topic-specific lexicon translation method mainly comes from the adaptation of lexical weights. It takes 72.8% of the time to do the adaptation, despite only lexical weights of the used rules are adapted. In contrast, our method has a speed of 10.2 words per second for each sentence on average, which is three times faster than the topic-specific lexicon translation method.

System	MT06	MT08	Avg
Baseline	30.20	21.93	26.07
One-to-One	30.27	22.12	26.20
One-to-Many	30.51	22.39	26.45

Table 4: Effects of one-to-one and one-to-many topic projection.

Meanwhile, we try to separate the effects of source-side topic distribution from the target-side topic distribution. From lines 4-6 of Table 2. We clearly find that the two rule-topic distributions improve the performance by 0.48 and 0.38 BLEU points over the baseline respectively. It seems that the source-side topic model is more helpful. Furthermore, when combine these two distributions, the improvement is increased to 0.64 points. This indicates that the effects of source-side and target-side distributions are complementary.

6.3 Effect of Topic Sensitivity Model

As described in Section 3.2, because the similarity features always punish topic-insensitive rules, we introduce topic sensitivity features as a complement. In the last line of Table 2, we obtain a further improvement of 0.23 points, when incorporating topic sensitivity features with topic similarity features. This suggests that it is necessary to distinguish topic-insensitive and topic-sensitive rules.

6.4 One-to-One Vs. One-to-Many Topic Projection

In Section 4.2, we find that source-side topic and target-side topics may not exactly match, hence we use one-to-many topic correspondence. Yet another method is to enforce one-to-one topic projection (Tam et al., 2007). We achieve one-to-one projection by aligning a target topic to the source topic with the largest correspondence probability as calculated in Section 4.2.

Table 4 compares the effects of these two method-

System	MT06	MT08	Avg
Baseline	30.20	21.93	26.07
Phrase-rule	30.53	22.29	26.41
Monotone-rule	30.72	22.62	26.67
Reordering-rule	30.31	22.40	26.36
All-rule	30.95	22.92	26.94

Table 5: Effect of our topic model on three types of rules. Phrase rules are fully lexicalized, while monotone and reordering rules contain nonterminals.

s. We find that the enforced one-to-one topic method obtains a slight improvement over the baseline system, while one-to-many projection achieves a larger improvement. This confirms our observation of the non-one-to-one mapping between source-side and target-side topics.

6.5 Effect on Various Types of Rules

To get a more detailed analysis of the result, we further compare the effect of our method on different types of rules. We divide the rules into three types: phrase rules, which only contain terminals and are the same as the phrase pairs in phrase-based system; monotone rules, which contain non-terminals and produce monotone translations; reordering rules, which also contain non-terminals but change the order of translations. We define the monotone and reordering rules according to Chiang et al., (2008).

Table 5 show the results. We can see that our method achieves improvements on all the three types of rules. Our topic similarity method on monotone rule achieves the most improvement which is 0.6 BLEU points, while the improvement on reordering rules is the smallest among the three types. This shows that topic information also helps the selections of rules with non-terminals.

7 Related Work

In addition to the topic-specific lexicon translation method mentioned in the previous sections, researchers also explore topic model for machine translation in other ways.

Foster and Kunh (2007) describe a mixture-model approach for SMT adaptation. They first split a training corpus into different domains. Then, they train separate models on each domain. Finally, they combine a specific domain translation model with a general domain translation model depending on various text distances. One way to calculate the distance is using topic model.

Gong et al. (2010) introduce topic model for filtering topic-mismatched phrase pairs. They first assign a specific topic for the document to be translated. Similarly, each phrase pair is also assigned with one specific topic. A phrase pair will be discarded if its topic mismatches the document topic.

Researchers also introduce topic model for cross-lingual language model adaptation (Tam et al., 2007; Ruiz and Federico, 2011). They use bilingual topic model to project latent topic distribution across languages. Based on the bilingual topic model, they apply the source-side topic weights into the target-side topic model, and adapt the n-gram language model of target side.

Our topic similarity model uses the document topic information. From this point, our work is related to context-dependent translation (Carpuat and Wu, 2007; He et al., 2008; Shen et al., 2009). Previous work typically use neighboring words and sentence level information, while our work extents the context into the document level.

8 Conclusion and Future Work

We have presented a topic similarity model which incorporates the rule-topic distributions on both the source and target side into traditional hierarchical phrase-based system. Our experimental results show that our model achieves a better performance with faster decoding speed than previous work on topic-specific lexicon translation. This verifies the advantage of exploiting topic model at the rule level over the word level. Further improvement is achieved by distinguishing topic-sensitive and topic-insensitive rules using the topic sensitivity model.

In the future, we are interesting to find ways to exploit topic model on bilingual data without document boundaries, thus to enlarge the size of training data. Furthermore, our training corpus mainly focus on news, it is also interesting to apply our method on corpus with more diverse topics. Finally, we hope to apply our method to other translation models, especially syntax-based models.

Acknowledgement

The authors were supported by High-Technology R&D Program (863) Project No 2011AA01A207 and 2012BAH39B03. This work was done during Xinyan Xiao's internship at I^2R. We would like to thank Yun Huang, Zhengxian Gong, Wenliang Chen, Jun lang, Xiangyu Duan, Jun Sun, Jinsong Su and the anonymous reviewers for their insightful comments.

References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proc of WMT 2009*.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *AAS*, 1(1):17–35.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

Marine Carpuat and Dekai Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of the MT Summit XI*.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. EMNLP 2008*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Zhengxian Gong, Yu Zhang, and Guodong Zhou. 2010. Statistical machine translation based on lda. In *Proc. IUCS 2010*, page 286 ‐ 290, Oct.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proc. EMNLP 2008*.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proc. of UAI 1999*, pages 289–296.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL 2003*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP 2004*.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proc. of EMNLP 2009*.

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL 2002*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL 2003*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL 2002*.

Nick Ruiz and Marcello Federico. 2011. Topic adaptation for lecture translation through bilingual latent semantic models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, July.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2009. Effective use of linguistic and contextual information for statistical machine translation. In *Proc. EMNLP 2009*.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. ICSLP 2002*.

Yik-Cheung Tam, Ian R. Lane, and Tanja Schultz. 2007. Bilingual lsa-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proc. Coling 2008*.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proc. ACL 2006*.

Bin Zhao and Eric P. Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Proc. NIPS 2007*.