Journal of Artificial Intelligence Research 50 (2014) 1-30 Submitted 12/13; published 05/14

Topic-Based Dissimilarity and Sensitivity Models for Translation Rule Selection

Min Zhang MINZHANG@SUDA.EDU.CN
Provincial Key Laboratory for Computer Information Processing Technology,
Soochow University, Suzhou, China

Xinyan Xiao XIAOXINYAN@ICT.AC.CN
IIP Key Lab, Institute of Computing Technology,
Chinese Academy of Sciences, China

Deyi Xiong DYXIONG@SUDA.EDU.CN
Provincial Key Laboratory for Computer Information Processing Technology,
Soochow University, Suzhou, China

Qun Liu LIUQUN@ICT.AC.CN
CNGL, School of Computing, Dublin City University, Ireland
IIP Key Lab, Institute of Computing Technology,
Chinese Academy of Sciences, China

Abstract

Translation rule selection is a task of selecting appropriate translation rules for an ambiguous source-language segment. As translation ambiguities are pervasive in statistical machine translation, we introduce two topic-based models for translation rule selection which incorporates global topic information into translation disambiguation. We associate each synchronous translation rule with source- and target-side topic distributions.With these topic distributions, we propose a topic dissimilarity model to select desirable (less dissimilar) rules by imposing penalties for rules with a large value of dissimilarity of their topic distributions to those of given documents. In order to encourage the use of non-topic specific translation rules, we also present a topic sensitivity model to balance translation rule selection between generic rules and topic-specific rules. Furthermore, we project target-side topic distributions onto the source-side topic model space so that we can benefit from topic information of both the source and target language. We integrate the proposed topic dissimilarity and sensitivity model into hierarchical phrase-based machine translation for synchronous translation rule selection. Experiments show that our topic-based translation rule selection model can substantially improve translation quality.

1. Introduction

Translation rules are bilingual segments[1] that establish translation equivalences between the source and target language. They are widely used in statistical machine translation (SMT) with various representations ranging from word pairs to bilingual phrases and synchronous rules in word-, phrase- and syntax-based SMT respectively. Normally, a large number of translation rules can be learnt from bilingual training data for a single source segment which occurs in different contexts. For example, Xiong, Zhang, and Li (2012) observe that each Chinese verb can be translated with more

1. Here a segment is defined as a string of terminals and/or nonterminals.

than 140 different translation rules on average. Therefore how to select an appropriate translation rule for an ambiguous source segment is a very crucial issue in SMT.

Traditionally the appropriateness of a translation rule is measured with multiple probabilities estimated from word-aligned data, such as bidirectional translation probabilities (Koehn, Och, & Marcu, 2003). As such probabilities fail to capture local and global contexts of highly ambiguous source segments, they are not sufficient to select correct translation rules for these segments. Therefore various approaches have been proposed to capture rich contexts at the sentence level to help select proper translation rules for phrase- (Carpuat & Wu, 2007a) or syntax-based SMT (Chan, Ng, & Chiang, 2007; He, Liu, & Lin, 2008; Liu, He, Liu, & Lin, 2008). These studies show that local features, such as surrounding words, syntactic information and so on, are helpful for translation rule selection.

Beyond these contextual features at the sentence level, we conjecture that translation rules are also related to high-level global information, such as the topic (Hofmann, 1999; Blei, Ng, & Jordan, 2003) information at the document level. In order to visualize the relatedness between translation rules and document topics, we show four hierarchical phrase-based translation rules with their topic distributions in Figure 1. From the figure, we can observe that

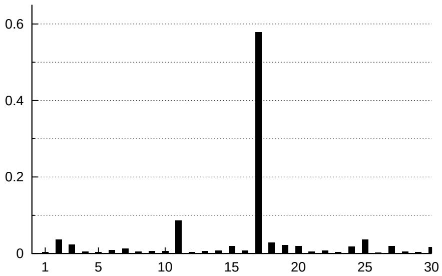
- First, translation rules can be divided into two categories in terms of their topic distributions: *topic-sensitive rules* (i.e., topic-specific rules) and *topic-insensitive rules* (i.e., non-topic specific or generic rules). The former rules, e.g., the translation rule (a), (b) and (d) in Figure 1, have much higher distribution probabilities on a few specific topics than other topics. The latter rules, e.g., the translation rule (c) of Figure 1, have an even distribution over all topics.
- Second, topic information can be used to disambiguate ambiguous source segments. In Figure 1, translation rule (b) and (c) have the same source segment. However their topic distributions are quite different. Rule (b) distributes on the topic about "international relations" with the highest probability, which suggests that rule (b) is much more related to this topic than other topics. In contrast, rule (c) has an even distribution over all topics. Therefore in a document on "international relations", rule (b) will be more appropriate than rule (c) for the source segment "给予 X_1".

These two observations suggest that different translation rules have different topic distributions and document-level topic information can be used to benefit translation rule selection.

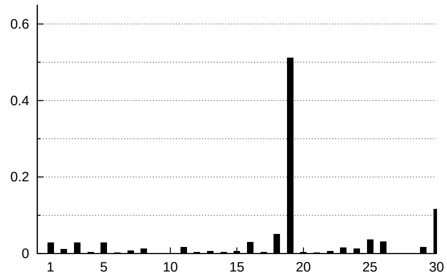
In this article, we propose a framework for translation rule selection that exactly capitalizes on document-level topic information. The proposed topic-based translation rule selection framework associates each translation rule with a topic distribution (rule-topic distribution) on both the source and target side. Each source document is also annotated with its corresponding topic distribution (document-topic distribution). Dissimilarity between the document-topic distribution and rule-topic distribution is calculated and used to help select translation rules that are related to documents in terms of topics. In particular,

- Given a document to be translated, we use a topic dissimilarity model to calculate the dissimilarity of each translation rule to the document based on their topic distributions. Our translation system will penalize candidate translations with high dissimilarities.[2]

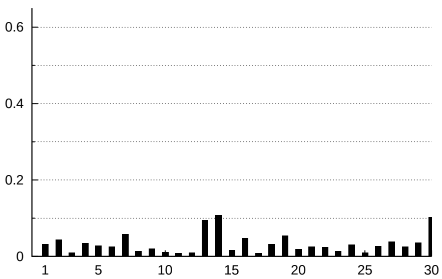
2. Section 6 explains why our system penalizes candidate translations with high dissimilarities.



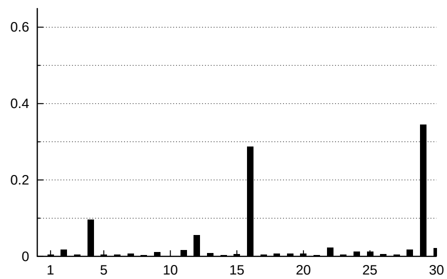
(a) 作战 能力 ⇒ operational capability



(b) 给予 X_1 ⇒ grants 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 shows the topic index and the Y-axis 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 dissimilarity between a topic-insensitive translation rule and a given source document computed by our topic dissimilarity model is often very high as documents are normally topic-sensitive. We don't want to penalize these generic topic-insensitive rules. Therefore we further propose a topic sensitivity model which rewards topic-insensitive rules so as to complement the topic dissimilarity model.

- We associate each translation rule with a rule-topic distribution on both the source and target side. In order to calculate the dissimilarity between target-side rule-topic distributions of translation rules and source-side document-topic distributions of given documents during decoding, we project the target-side rule-topic distributions of translation rules onto the space of source-side document topic model by one-to-many mapping.

We use a hierarchical phrase-based SMT system (Chiang, 2007) to validate the effectiveness of our topic-based models for translation rule selection. Experiments on Chinese-English translation tasks (Section 7) show that our method outperforms the baseline hierarchial phrase-based system by +1.2 BLEU points on large-scale training data.

The use of topic-based dissimilarity and sensitivity models to improve SMT was first presented in our previous paper (Xiao, Xiong, Zhang, Liu, & Lin, 2012). In this article, we provide more detailed comparison to related work and formulations of the two models as well as the integration procedure. More importantly, we carry out large-scale experiments with more bilingual and monolingual training data and incorporate a detailed analysis of the output of topic-based dissimilarity and sensitivity models at both the document and translation hypothesis level.

The rest of this article is organized as follows. Section 2 introduces related work. Section 3 provides background knowledge about statistical machine translation and topic modeling. Section 4 elaborates our topic-based translation rule selection framework, including the topic dissimilarity and topic sensitivity model. Section 5 discusses how we estimate rule-topic and document-topic distributions and how we project target-side rule-topic distributions onto the source-side topic space in a one-to-many mapping fashion. Section 6 presents the integration of the topic-based translation rule selection models into hierarchical phrase-based SMT. Section 7 describes a series of experiments that verify the effectiveness of our approach. Section 8 provides a detailed analysis of the output of our models. Section 9 gives some suggestions for bilingual topic modeling from the perspective of machine translation. Finally, we conclude in Section 10 with future directions.

2. Related Work

Our topic-based dissimilarity and sensitivity models for translation rule selection are related to three categories of work in SMT: translation rule selection, topic models for SMT and document-level translation. In this section, we introduce related approaches of the three categories and highlight the differences of our method from previous work.

2.1 Translation Rule Selection

As we mentioned before, translation rule selection is a very important task in SMT. Several approaches have been proposed for it recently. Carpuat and Wu explore both word and phrase sense disambiguation (WSD and PSD) for translation rule selection in phrase-based SMT (Carpuat & Wu, 2007a, 2007b). Their WSD and PSD system integrate sentence-level local collocation features. Experiments show that multi-word PSD can improve phrase selection. Also following the WSD line, Chan et al. (2007) integrate a WSD system into hierarchical phrase-based SMT for lexical selection or the selection of short phrases of length 1 or 2. Their WSD system also adopts sentence-level features of local collocations, surrounding words and so on.

Different from lexical or phrasal selection using WSD/PSD, He et al. (2008) propose a maximum entropy (MaxEnt) based model for context-dependent synchronous rule selection in hierarchical phrase-based SMT. Local context features such as phrase boundary words and part-of-speech information are incorporated into the model. Liu et al. (2008) extends the selection method of He et al. to integrate a similar MaxEnt-based rule selection model into a tree-to-string syntax-based SMT system (Liu, Liu, & Lin, 2006). Their model uses syntactic information from source parse trees as features.

The significant difference between our topic-based rule selection framework and previous approaches on translation rule selection is that we use global topic information to help select translation rules for ambiguous source segments rather than sentence-level local context features.

2.2 Topic Models for SMT

Topic modeling (Hofmann, 1999; Blei et al., 2003) is a popular technique for discovering underlying topic structures of documents. Recent years have witnessed that topic models have been explored for SMT. Zhao and Xing (2006, 2007) and Tam, Lane, and Schultz (2007) have proposed topic-specific lexicon translation adaptation models to improve translation quality. Such models focus on word-level translations. They first estimate word translation probabilities conditioned on topics, and then adapt lexical translation probabilities of phrases by these topic-conditioned probabilities. Since modern SMT systems use synchronous rules or bilingual phrases to translate sentences, we believe that it is more reasonable to incorporate topic models for phrase or synchronous rule selection than lexical selection.

Gong, Zhang, and Zhou (2010) adopt a topic model to filter out phrase pairs that are not consistent with source documents in terms of their topics. They assign a topic for each document to be translated. Similarly, each phrase pair is also assigned with one topic. A phrase pair will be discarded if its topic mismatches the document topic. The differences from their work are twofold. First, we calculate the dissimilarities of translation rules to documents based on their topic distributions instead of comparing the best topics assigned to translation rules and those of documents. Second, we integrate topic information into SMT in a soft-constraint manner via our topic-based models. They explore topic information in a hard-constraint fashion by discarding translation rules with unmatched topics.

Topic models are also used for domain adaptation on translation and language models in SMT. Foster and Kuhn (2007) describe a mixture model approach for SMT adaptation. They divide a training corpus into different domains, each of which is used to train a domain-specific translation model. During decoding, they combine a general domain translation model with a specific domain translation model that is selected according to various text distances calculated by topic model. Tam et al. (2007) and Ruiz and Federico (2011) use a bilingual topic model to project latent topic distributions across languages. Based on the bilingual topic model, they apply source-side topic weights onto the target-side topic model so as to adapt the target-side n-gram language model.

2.3 Document-Level Machine Translation

Since we incorporate document topic information into SMT, our work is also related to document-level machine translation. Tiedemann (2010) integrates cache-based language and translation models that are built from recently translated sentences into SMT. Gong, Zhang, and Zhou (2011) further extend this cache-based approach by introducing two additional caches: a static cache that stores phrases extracted from documents in training data which are similar to the document in question and a topic cache with target language topic words. Xiao, Zhu, Yao, and Zhang (2011) try to solve the translation consistency issue in document-level translation by introducing a hard constraint where ambiguous source words are required to be consistently translated into the most frequent translation options. Ture, Oard, and Resnik (2012) soften this consistency constraint by integrating three counting features into the decoder. These studies normally focus on the surface structure to capture inter-sentence dependencies for document-level machine translation while we explore the topic structure of a document for document translation.

3. Preliminaries

We establish in this section some background knowledge about both statistical machine translation and topic modeling. Although the introduction here is short, it is sufficient for understanding our

Sub-models	Descriptions
$\sum_1^I \log P(\overline{e}_i \vert \overline{f}_i)$	direct translation probabilities
$\sum_1^I \log P(\overline{f}_i \vert \overline{e}_i)$	inverse translation probabilities
$\sum_1^I \log P_{lex}(\overline{e}_i \vert \overline{f}_i)$	direct lexical translation probabilities
$\sum_1^I \log P_{lex}(\overline{f}_i \vert \overline{e}_i)$	inverse lexical translation probabilities
$\sum_1^{\vert e\vert} \log P(e_i \vert e_1 ... e_{i-1})$	language model
$\sum_1^I \log \psi(\overline{e}_i, \overline{f}_i)$	reordering model
$\vert e\vert$	word count
I	rule count

Table 1: The most widely-used sub-models of statistical machine translation. I is the number of translation rules that are used to generate the target sentence e given the source sentence f. $\overline{e}_i$ and $\overline{f}_i$ are the target and source side of a translation rule r_i.

topic-based dissimilarity and sensitivity models that try to bridge the gap between topic modeling and statistical machine translation.

3.1 Statistical Machine Translation

Given a source sentence f, most SMT systems find the best translation $\hat{e}$ among all possible translations as follows.

$$\begin{aligned}
\hat{e} &= \underset{e}{\operatorname{argmax}} \left\{ \frac{\exp\left[\sum_1^M \lambda_m h_m(f,e)\right]}{\sum_{e'} \exp\left[\sum_1^M \lambda_m h_m(f,e')\right]} \right\} \\
&= \underset{e}{\operatorname{argmax}} \left\{ \exp\left[\sum_{m=1}^{M} \lambda_m h_m(f,e)\right] \right\} \\
&= \underset{e}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f,e) \right\}
\end{aligned} \tag{1}$$

where $h_m(f,e)$ is a feature function defined on the source sentence f and the corresponding translation e, λ_m is the weight of the feature function. Since the normalization $\sum_{e'} \exp\left[\sum_1^M \lambda_m h_m(f,e')\right]$ is constant for all possible translations e', we do not need to calculate it during decoding.

The weighted model in the equation (1) is a log-linear model. The feature functions $h_m(f,e)$ are also referred to as sub-models[3] as they are components of the log-linear model. In Table 1, we show the most widely-used feature functions in SMT. Most of them can be easily factored over translation rules, which facilitates the application of dynamic programming in decoding. We will show that our proposed topic-based dissimilarity and sensitivity models can be also easily factorized in Section 4.

3. This notation is used when we want to emphasize that a sub-model is a component of the log-linear model. Otherwise we just call them models, such as a language model, a reordering model and so on.

In the log-linear model of SMT, the sub-models are trained separately and combined under the assumption that they are independent of each other. The associated weights λs can be tuned using minimum error rate training (MERT) (Och, 2003) or the Margin Infused Relaxed Algorithm (MIRA) (Chiang, Marton, & Resnik, 2008). Note that the normalization factor in the equation (1) is not calculated in these training algorithms. This is because these algorithms directly optimize the log-linear model of SMT towards some translation quality measure such as BLEU. Feature weights that are optimized towards criteria such as Maximum Mutual Information (MMI) are not necessarily optimal with respect to translation quality (Och, 2003).

As we integrate the proposed two models into the log-linear model of a hierarchical phrase-based SMT system (Section 6) in order to validate the effectiveness of the two models, we provide more details about hierarchical phrase-based SMT (Chiang, 2005) in this section. Translation rules in hierarchial phrase-based SMT are synchronous context-free grammar rules, which can be denoted as follows.

$$X \rightarrow \langle \alpha, \beta, \sim \rangle \tag{2}$$

where X is an undifferentiated nonterminal, α and β are strings of terminals and nonterminals[4] on the source and target side respectively, $\sim$ denotes the one-to-one mapping between nonterminals in α and nonterminals in β. These rules can be automatically extracted from word-aligned bilingual training data. In addition to these rules, two special rules are also introduced into hierarchical phrase-based SMT.

$$\begin{aligned} S &\rightarrow \langle X_{\sim 1}, X_{\sim 1} \rangle \\ S &\rightarrow \langle S_{\sim 0} X_{\sim 1}, S_{\sim 0} X_{\sim 1} \rangle \end{aligned} \tag{3}$$

These two rules are used to serially concatenate nonterminal Xs in a monotonic manner to form an initial symbol S, the start symbol of the grammar of hierarchical phrase-based SMT.

The log-linear model of hierarchical phrase-based SMT can be formulated as follows.

$$w(\mathcal{D}) = \exp\left(\sum_{r \in D} \log(t(r)) + \lambda_{lm} \log P_{lm}(e) + \lambda_{wp}|e| + \lambda_{rp} I\right) \tag{4}$$

where $\mathcal{D}$ is a derivation defined as a set of triples (r, i, j), each of which denotes an application of a translation rule that spans words i from j on the source side. I is the number of translation rules in $\mathcal{D}$. The probability of a translation rule r is defined as

$$t(r) = P(\alpha|\beta)^{\lambda_1} P(\beta|\alpha)^{\lambda_2} P_{lex}(\alpha|\beta)^{\lambda_3} P_{lex}(\beta|\alpha)^{\lambda_4} \tag{5}$$

where the lexical translation probabilities $P_{lex}(\alpha|\beta)$ and $P_{lex}(\beta|\alpha)$ estimate the probabilities that the words in α translate the words in β in a word-by-word fashion (Koehn et al., 2003).

3.2 Topic Modeling

Topic modeling is used to discover topics that occur in a collection of documents. Both Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Probabilistic Latent Semantic Analysis (PLSA)

4. In order to simplify the decoder implementation, at most two nonterminals are allowed in hierarchical translation rules.

(Hofmann, 1999) are topic models. As LDA is the most widely used topic model, we exploit it to mine topics for our translation rule selection.

LDA views each document as a mixture of various topics, each of which is a probability distribution over words. More particularly, LDA works in a generative process as follows.

- For each document D_j, sample a document-topic distribution (per-document topic distribution) θ_j from a Dirichlet distribution Dir(α): $\theta_j \sim$ Dir(α);
- for each word $w_{j,i}$ of N_j words in the document D_j,
  - Sample a topic assignment $z_{j,i} \sim$ Multinomial(θ_j);
  - Sample the word $w_{j,i} \sim$ Multinomial$(\varphi_{z_{j,i}})$ where $\varphi_{z_{j,i}}$ is the per-topic word distribution of topic $z_{j,i}$ drawn from Dir(β).

Generally speaking, LDA contains two groups of parameters. The first group of parameters characterizes document-topic distributions (θ_j), which record the distribution of each document over topics. The second group of parameters is used for topic-word distributions (φ_k), which represent each topic as a distribution over words.

Given a document collection with observed words $\mathbf{w} = \{w_{j,i}\}$, the goal of LDA inference is to compute the values for these two sets of parameters θ and φ as well as the latent topic assignments $\mathbf{z} = \{z_{j,i}\}$. The inference is complicated due to the latent topic assignments $\mathbf{z}$. An efficient inference algorithm that has been proposed to address this problem is Collapsed Gibbs Sampling (Griffiths & Steyvers, 2004), where the two sets of parameters θ and φ are integrated out of the LDA model, and only the latent topic assignments $\mathbf{z}$ are sampled from $P(\mathbf{z}|\mathbf{w})$. Once we obtain the values of $\mathbf{z}$, we can estimate θ and φ by recovering their posterior distributions given $\mathbf{z}$ and $\mathbf{w}$. In Section 4, we will use these two sets of estimated parameters and the topic assignments of words to calculate the parameters of our models.

4. Topic-based Dissimilarity and Sensitivity Models

In this section, we elaborate our topic-based models for translation rule selection, including a topic dissimilarity model and a topic sensitivity model.

4.1 Topic Dissimilarity Model

Sentences should be translated in accordance with their topics (Zhao & Xing, 2006, 2007; Tam et al., 2007). Take the translation rule (b) in Figure 1 as an example. If the source side of rule (b) occurs in a document on "international relations", we hope to encourage the application of rule (b) rather than rule (c). This can be achieved by calculating the dissimilarity between probability distributions of a translation rule and a document over topics.

In order to calculate such a topic dissimilarity for translation rule selection, we associate both the source and target side of a translation rule with a *rule-topic distribution* $P(z_\diamond|r_\diamond)$, where $\diamond$ is the placeholder for the source side f or target side e, $r_\diamond$ is the source or target side of a translation rule r, and $z_\diamond$ is the corresponding topic of $r_\diamond$. Therefore each translation rule has two rule-topic distributions: $P(z_f|r_f)$ on the source side and $P(z_e|r_e)$ on the target side.

Supposing there are K topics, the two distributions can be represented by a K-dimension vector. The k-th component $P(z_\diamond = k|r_\diamond)$ denotes the probability of topic k given $r_\diamond$. The source- and target-side rule-topic distributions are separately estimated from training data. The estimation method is described in Section 5, where we also discuss the reason why we estimate them in a separate manner.

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 $P(z = k|d)$ is the topic proportion for topic k in document d. Different from the rule-topic distribution, the document-topic distribution can be directly inferred by an off-the-shelf LDA tool.

Based on the defined rule-topic and document-topic distributions, we can measure the dissimilarity of a translation rule to a document so as to decide whether the rule is suitable for the document in translation. Traditionally, the similarity of two probability distributions is calculated by information measurements such as Jensen-Shannon divergence (Lin, 2006) or Hellinger distance (Blei & Lafferty, 2007).

Here we adopt the Hellinger distance (HD) to measure the topic dissimilarity, which is symmetric and widely used for comparing two probability distributions (Blei & Lafferty, 2007). Given a rule-topic distribution $P(z_\diamond|r_\diamond)$ and a document-topic distribution $P(z|d)$, HD is computed as follows.

$$\text{HD}(P(z|d), P(z_\diamond|r_\diamond)) = \sum_{k=1}^{K} \left(\sqrt{P(z=k|d)} - \sqrt{P(z_\diamond=k|r_\diamond)}\right)^2 \tag{6}$$

Let $\mathcal{D}$ be a derivation as defined in Section 3.1. Let $\mathbf{P}(\mathbf{z}|\mathbf{r})$ represent corresponding rule-topic distributions for all rules in $\mathcal{D}$. Our topic dissimilarity model $\text{Dsim}(P(z|d), \mathbf{P}(\mathbf{z}|\mathbf{r}))$ on a derivation $\mathcal{D}$ is defined on the HD of the equation (6) as follows

$$\text{Dsim}(P(z|d), \mathbf{P}(\mathbf{z}|\mathbf{r})) = \sum_{r\in\mathcal{D}} \text{HD}(P(z|d), P(z_\diamond|r_\diamond)) \tag{7}$$

Obviously, the larger the Hellinger distance between a candidate translation yielded by a derivation and a document, the larger the dissimilarity between them. With the topic dissimilarity model defined above, we aim to select translation rules that are similar to the document to be translated in terms of their topics.

4.2 Topic Sensitivity Model

Before we introduce the topic sensitivity model, let's revisit Figure 1. We can easily find that the probability of rule (c) distributes evenly over all topics. This indicates that it is insensitive to topics, and can be therefore applied on any topics. In contrast, the distributions of the other three rules peak on a few topics. Generally speaking, a topic-insensitive rule has a fairly flat distribution over all topics, while a topic-sensitive rule has a sharp distribution over a few topics.

As a document typically focuses on a few topics, it has a sharp distribution over these topics. In other words, documents are normally topic-sensitive. Since the distribution of a topic-insensitive rule is fairly flat, the dissimilarity between a topic-insensitive rule and a topic-sensitive document will be very low. Therefore, our system with the proposed topic dissimilarity model will punish topic-insensitive rules.

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 a topic of "love", the rule (b) and (c) in Figure 1 are both dissimilar to the document as rule (b) relates to the "international relations" topic and rule (c) is topic-insensitive. Nevertheless, since rule (c) occurs more frequently across various topics, we prefer rule (c) to rule (b) when we translate a document about "love".

To address such issue of the topic dissimilarity model, we further propose a topic sensitivity model. The model employs an entropy based metric to measure the topic sensitivity of a rule as follows

$$H(P(z_\diamond|r_\diamond)) = -\sum_{k=1}^{K} P(z_\diamond = k|r_\diamond) \times \log(P(z_\diamond = k|r_\diamond)) \tag{8}$$

According to this equation, a topic-insensitive rule normally has a large entropy while a topic-sensitive rule has a smaller entropy.

Given a derivation $\mathcal{D}$ and rule-topic distributions $\mathbf{P}(\mathbf{z}|\mathbf{r})$ for rules in $\mathcal{D}$, the topic sensitivity model is defined as follows.

$$\text{Sen}(\mathbf{P}(\mathbf{z}|\mathbf{r})) = \sum_{r \in \mathcal{D}} H(P(z_\diamond|r_\diamond)) \tag{9}$$

Incorporating the topic sensitivity model with the topic dissimilarity model, we enable our SMT system to balance the selection of topic-sensitive and topic-insensitive rules. Given rules with approximately equal values of topic dissimilarity, we prefer topic-insensitive rules.

5. Estimation

Unlike document-topic distributions that can be directly learned by LDA tools, we need to estimate rule-topic distributions for translation rules. As we want to exploit topic information of both the source and target language, we *separately* train two monolingual topic models on the source and target side, and learn correspondences between the two topic models via word alignments in the bilingual training data.

Particularly, we adopt two rule-topic distributions for each translation rule: 1) the source-side rule-topic distribution $P(z_f|r_f)$ and the 2) the target-side rule-topic distribution $P(z_e|r_e)$, both of which are defined in Section 4.1. These two rule-topic distributions are estimated using trained topic models in the same way (Section 5.1). Notably, only source-language documents are available during decoding. In order to compute the dissimilarity between the target-side rule-topic distribution of a translation rule and the source-side document-topic distribution of a given document， we need to project the target-side rule-topic distribution of a translation rule onto the space of the source-side topic model (Section 5.2).

We can also establish alternative approaches to the estimation of rule-topic distributions via multilingual topic models (Mimno, Wallach, Naradowsky, Smith, & McCallum, 2009; Boyd-Graber & Blei, 2009) or bilingual topic models that also infer word-to-word alignments in document pairs (Zhao & Xing, 2006, 2007). The former multilingual topic models only require that documents in different languages are comparable in terms of content similarity. In contrast, the latter bilingual topic models require that documents are parallel, i.e., translations of each other, so as to capture word alignments.

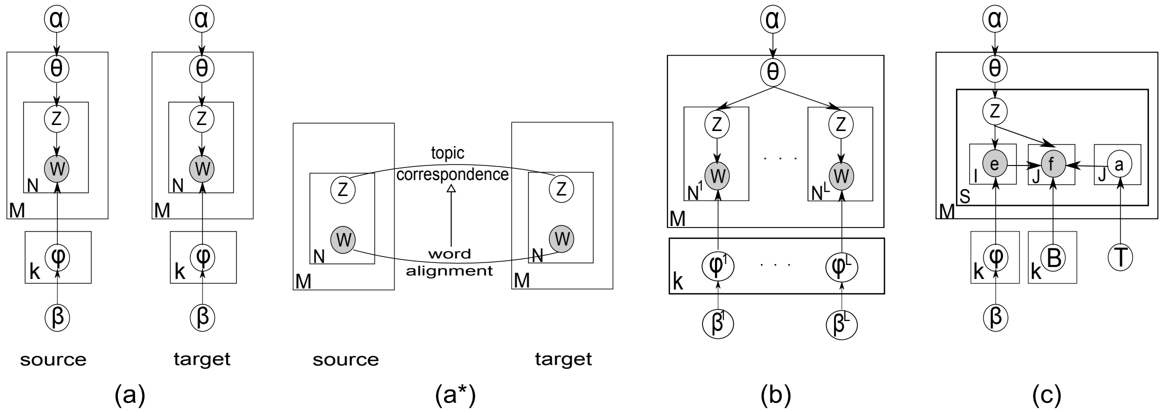


Figure 2: Graphical model representations of (a) our bilingual topic model, (b) polylingual topic model of Mimno et al. (2009), and (c) bilingual topic model of Zhao and Xing (2007) where S is the number of parallel sentence pairs in a document, a is the word alignment between a source and target sentence. For simplicity, we do not display HMM transitions among word alignments a. Subfigure (a*) shows how we build topic correspondences between the source and target language after source and target topics are separately learned as shown in (a).

The biggest difference between our method and these multilingual/bilingual topic models is that they use the same per-tuple topic distribution θ for all documents in the same tuple. Here we define the tuple as a set of documents in different languages. A per-tuple topic distribution is similar to a per-document topic distribution. The only difference between them is that the per-tuple topic distribution is shared by all documents in the tuple.

Topic assignments for words in these languages are naturally connected since they are sampled from the same topic distribution. In contrast, we assume that each document on the source/target side has its own sampled document-specific distribution over topics. Topic correspondences between the source and target document are learned by projection via word alignments. We visualize this difference in Figure 2.

Yet another difference between our models and the topic-specific lexicon translation model of Zhao and Xing (2007) is that they use their bilingual topics to improve SMT at the word level instead of the rule level. 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. In Section 7.2.3, we empirically compare our model with the topic-specific lexicon translation model.

Tam et al. (2007) also construct two monolingual topic models for parallel source and target documents. They build the topic correspondences between source and target documents by enforcing a one-to-one topic mapping constraint. We project target-side topics onto the space of the source-side topic model in a one-to-many fashion. In Section 7.3.1, we compare these two different methods for building topic correspondences.

5.1 Rule-Topic Distribution Estimation

We estimate rule-topic distributions from word-aligned bilingual training corpus with document boundaries explicitly given. The source- and target-side rule-topic distributions are estimated in the same way. Therefore, for simplicity, we only describe the estimation of the source-side rule-topic distribution $P(z_f|r_f)$ of a translation rule in this section.

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

We first extract translation rules from bilingual training data in a traditional way. When the source side of a translation rule r_f is extracted from a source-language document d_f with a document-topic distribution $P(z_f|d_f)$, we obtain an instance $(r_f, P(z_f|d_f), \epsilon)$, where ϵ is the fraction count of an instance as described by Chiang (2007). In this way, we can collect a set of instances $\mathcal{I}$ = $\{(r_f, P(z_f|d_f), \epsilon)\}$ with different document-topic distributions for each translation rule. Using these instances, we calculate the probability $P(z_f = k|r_f)$ of r_f over topic k as follows:

$$P(z_f = k|r_f) = \frac{\sum_{I \in \mathcal{I}} \epsilon \times P(z_f = k|d_f)}{\sum_{k'=1}^{K} \sum_{I \in \mathcal{I}} \epsilon \times P(z_f = k'|d_f)} \tag{10}$$

Based on this equation, we can obtain two rule-topic distributions $P(z_f|r_f)$ and $P(z_e|r_e)$ for each rule using the source- and target-side document-topic distributions $P(z_f|d_f)$ and $P(z_e|d_e)$ respectively.

5.2 Target-Side Rule-Topic Distribution Projection

As described in the previous section, we also estimate target-side rule-topic distributions. However, we can not directly use the equation (6) to calculate the dissimilarity between the target-side rule-topic distribution $P(z_e|r_e)$ of a translation rule and the source-side document-topic distribution $P(z_f|d_f)$ of a source-language document that is to be translated. In order to measure this dissimilarity, we need to project target-side topics onto the source-side topic space. The projection takes the following two steps.

- First, we calculate a correspondence probability $p(z_f|z_e)$ for each pair of a target-side topic z_e and a source-side topic z_f, which are inferred by the two separately trained monolingual topic models respectively.
- Second, we project the target-side rule-topic distribution of a translation rule onto the source-side topic space using the correspondence probabilities learned in the first step.

In the first step, we estimate the topic-to-topic correspondence probabilities using co-occurrence counts of topic assignments of source and target words in the word-aligned corpus. The topic assignments of source/target words are inferred by the two monolingual topic models. With these topic assignments, we characterize a sentence pair (f, e) as $(\mathbf{z}_f, \mathbf{z}_e, \mathbf{a})$, where $\mathbf{z}_f$ and $\mathbf{z}_e$ are two vectors containing topic assignments for words in the source and target sentence f and e respectively, and $\mathbf{a}$ is a set of word alignment links $\{(i, j)\}$ between the source and target sentence. Particularly, a link (i, j) represents that a source-side position i aligns to a target-side position j.

With these notations, we calculate the co-occurrence count of a source-side topic k_f and a target-side topic k_e as follows.

$$\sum_{(\mathbf{z}_f, \mathbf{z}_e, \mathbf{a})} \sum_{(i,j) \in \mathbf{a}} \delta(\mathbf{z}_{f_i}, k_f) * \delta(\mathbf{z}_{e_j}, k_e) \tag{11}$$

where $\mathbf{z}_{f_i}$ and $\mathbf{z}_{e_j}$ are topic assignments for words f_i and e_j respectively, $\delta(x, y)$ is the Kronecker function, which is 1 if $x = y$ and 0 otherwise.

We then compute the topic-to-topic correspondence probability of $P(z_f = k_f | z_e = k_e)$ by normalizing the co-occurrence count as follows.

$$P(z_f = k_f | z_e = k_e) = \frac{\sum_{(\mathbf{z}_f, \mathbf{z}_e, \mathbf{a})} \sum_{(i,j) \in \mathbf{a}} \delta(\mathbf{z}_{f_i}, k_f) * \delta(\mathbf{z}_{e_j}, k_e)}{\sum_{(\mathbf{z}_f, \mathbf{z}_e, \mathbf{a})} \sum_{(i,j) \in \mathbf{a}} \delta(\mathbf{z}_{e_j}, k_e)} \tag{12}$$

Overall, after the first step, we obtain a topic-to-topic correspondence matrix $\mathbf{M}_{K_e \times K_f}$, 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 $P(z_e | r_e)$ to the source-side topic space by multiplication as follows.

$$T(P(z_e | r_e)) = P(z_e | r_e) \cdot \mathbf{M}_{K_e \times K_f} \tag{13}$$

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

Word alignment noises may be introduced in the equation (11), which in turn may flatten the sharpness of the projected topic distributions calculated in the equation (13). In order to decrease the flattening effects of word alignment noises, we take the following action in practice: if the topic-to-topic correspondence probability $P(z_f = k_f | z_e = k_e)$ calculated via word alignments is less than $\frac{1}{K}$ where K is the predefined number of topics, we set it to 0 and then re-normalize all other correspondence probabilities of the target-side topic k_e.

Obviously, our projection method allows one target-side topic z_e to align to multiple source-side topics. This is different from the one-to-one correspondence used by Tam et al. (2007). We investigate the correspondence matrix $\mathbf{M}_{K_e \times K_f}$ obtained from our training data. We find that the topic correspondence between the source and target language is not necessarily one-to-one. Typically, the correspondence probability $P(z_f = k_f | z_e = k_e)$ of a target-side topic mainly distributes over two or three source-side topics. Table 2 shows an example of a target-side topic with its three mainly aligned source-side topics.

6. Integration

We incorporate our topic dissimilarity and sensitivity model as two new features into a hierarchical phrase-based system (Chiang, 2007) under the log-linear discriminative framework (Och & Ney, 2002). The dissimilarity values are positive as Hellinger distances are positive. The weight of this dissimilarity feature tuned by MERT will be negative. Therefore the log-linear model will favor those candidate translations with lower values of the dissimilarity feature (less dissimilar). In other words, translation rules that are more similar to the document to be translated in terms of their topics will be selected.

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 2: An example of topic-to-topic correspondence. The last line shows the correspondence probability. Each column shows a topic represented by its top-10 topical words. The first column is a target-side topic, while the remaining three columns are source-side topics.

One possible side-effect of the integration of such a dissimilarity feature is that our system will favour translations generated by fewer translation rules against those generated by more translation rules because more translation rules result in higher dissimilarity (see the equation (7)). That is to say, the topic-based dissimilarity feature also acts as a translation rule count penalty on derivations. Fortunately, however, we also use a translation rule count feature (see the last row in Table 1) which normally favours translations yielded by a derivation with a large number of translation rules. This feature will balance against the mentioned side-effect of our topic-based dissimilarity feature.

As each translation rule is associated with a source-side rule-topic distribution and a projected target-side rule-topic distribution during decoding, we add four features as follows.[5]

- Dsim$(P(z_f|d), \mathbf{P}(\mathbf{z}_f|\mathbf{r}_f))$ (or DsimSrc): Topic dissimilarity feature on source-side rule-topic distributions.
- Dsim$(P(z_f|d), T(\mathbf{P}(\mathbf{z}_e|\mathbf{r}_e)))$ (or DsimTrg): Topic dissimilarity feature on projected target-side rule-topic distributions.
- Sen$(\mathbf{P}(\mathbf{z}_f|\mathbf{r}_f))$ (or SenSrc): Topic sensitivity feature on source-side rule-topic distributions.
- Sen$(T(\mathbf{P}(\mathbf{z}_e|\mathbf{r}_e))$ (or SenTrg): Topic sensitivity feature on projected target-side rule-topic distributions.

The source-side and projected target-side rule-topic distributions for translation rules can be calculated before decoding as described in the last section. During decoding, we first infer the topic distribution $P(z_f|d)$ for a given document of the source language. When a translation rule is adopted in a derivation, the scores of the four features will be updated correspondingly according to the equation (7) and (9). Obviously, the computational cost of these features is rather small.

5. Since the glue rule and rules of unknown words are not extracted from training data, we just set the values of the four features for these rules to zero.

For topic-specific lexicon translation models (Zhao & Xing, 2007; Tam et al., 2007), they first calculate topic-specific translation probabilities by normalizing the entire lexicon translation table and then adapt the lexical weights of translation rules correspondingly during decoding. This makes the decoder run slower. Therefore, comparing with previous topic-specific lexicon translation methods, our method provides a more efficient way for incorporating topic models into SMT.

7. Experiments

In this section, we conducted two groups of experiments to validate the effectiveness of our topic-based translation rule selection framework. In the first group of experiments, we use medium-scale bilingual data to train our SMT system and topic models. The purpose of this group of experiments is to quickly answer the following questions:

- Is our topic dissimilarity model able to improve translation rule selection in terms of BLEU? Furthermore, are the source-side and target-side rule-topic distributions complementary to each other?
- Is it helpful to introduce the topic sensitivity model to distinguish topic-insensitive and topic-sensitive rules?
- Is our topic-based method better than previous topic-specific lexicon translation method (Zhao & Xing, 2007) in terms of both BLEU and decoding speed?

After we confirm the efficacy of our topic-based dissimilarity and sensitivity model on medium-scale training data, we conducted a second group of experiments on large-scale training data to further investigate the following questions:

- Is our one-to-many target-side rule-topic projection method better than previous methods proposed by Zhao and Xing (2007) or Tam et al. (2007)?
- What are the effects of our models on various types of rules, such as phrase rules and rules with non-terminals?
- What else can we achieve if we use more monolingual data to train topic models?

7.1 Setup

We carried out our experiments on NIST Chinese-to-English translation. We used the NIST evaluation set of 2005 (MT05) as our development set, and sets of MT06/MT08 as the test sets. The numbers of documents in MT05, MT06, MT08 are 100, 79, and 109 respectively. Case-insensitive NIST BLEU (Papineni, Roukos, Ward, & Zhu, 2002) was used to measure translation performance. We used minimum error rate training (Och, 2003) to optimize the feature weights.

In our medium-scale experiments, we used the FBIS corpus as our bilingual training data, which contains 10,947 documents, 239K sentence pairs with 6.9M Chinese words and 9.14M English words. In our large-scale experiments, the bilingual training data consists of LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hong Kong Hansards/Laws/News).

These selected corpora contain 103,236 documents and 2.80M sentences. On average, each document has 28.4 sentences.

We obtained symmetric word alignments of training data by first running GIZA++ (Och & Ney, 2003) in both directions and then applying the refinement rule "grow-diag-final-and" (Koehn et al., 2003). Our hierarchical phrase translation rules were extracted from word-aligned training data. We used the SRILM toolkit (Stolcke, 2002) to train language models on the Xinhua portion of the GIGAWORD corpus, which contains 238M English words. We trained a 4-gram language model for our medium-scale experiments and a 5-gram language model for our large-scale experiments.

In order to train the two monolingual topic models on the source and target side of our bilingual training data, we used the open source LDA tool GibbsLDA++.[6] GibssLDA++ is an implementation of LDA using gibbs sampling for parameter estimation and inference. The source- and target-side topic models were separately estimated from the Chinese and English part of the bilingual training data. We set the number of topic $K = 30$ for both the source- and target-side topic models, and used the default setting of the tool for training and inference.[7] During decoding, we inferred the document-topic distribution for each document in the dev/test sets before translation using the trained source-side topic model. Note that the topic inference on the dev/test sets was performed after all parameters of the two topic models were estimated on the training data.

The case-insensitive BLEU-4 was used as our evaluation metric. We performed the statistical significance in BLEU differences using the paired bootstrap re-sampling (Koehn, 2004). In order to alleviate the impact of the instability of MERT, we ran the tuning process three times for all our large scale experiments and presented the average BLEU scores on the three runs following the suggestion by Clark, Dyer, Lavie, and Smith (2011)

7.2 Medium-Scale Experiments

In this section, we conducted medium-scale experiments to investigate the effectiveness of our two topic-based models for translation rule selection.

7.2.1 EFFECT OF TOPIC DISSIMILARITY MODEL

We quickly investigated the effectiveness of our topic dissimilarity and sensitivity model using medium-scale training data. Results are shown in Table 3. From the table, we can observe that

- If we use the topic dissimilarity model only with the source-side or projected target-side rule-topic distributions ("DsimSrc/DsimTrg" in the table, see descriptions in Section 5), we can obtain an absolute improvement of 0.48/0.38 BLEU points over the baseline.
- If we combine the two topic dissimilarity features together, we can achieve a further improvement of 0.16 BLEU points over "DsimSrc".

These two observations show that our topic dissimilarity model is able to improve translation quality in terms of BLEU.

6. http://gibbslda.sourceforge.net/
7. 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. In order to improve the stability of the topic estimation, we run the tool multiple times and use the best model with respect to the log-likelihood.

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

Table 3: Results of our topic dissimilarity and sensitivity model in terms of BLEU and speed (words per second), comparing with the traditional hierarchical system ("Baseline") and the system with the topic-specific lexicon translation model ("TopicLex"). "DsimSrc" and "DsimTrg" are topic dissimilarity features on the source-side and projected target-side rule-topic distributions respectively. "Dsim+Sen" activates both the two dissimilarity features and the two sensitivity features as described in Section 6. "Avg" denotes average BLEU scores on the two test sets. Scores in bold are significantly better than *Baseline* ($p < 0.01$). "Speed" denotes the number of words translated per second.

Rule Type	Count	Src-Sen(%)	Trg-Sen(%)
Phrase	3.9M	83.4	84.4
Monotone	19.2M	85.3	86.1
Reordering	5.7M	85.9	86.8
All	28.8M	85.1	86.0

Table 4: Percentages of topic-sensitive rules listed by rule types according to entropies of their source-side ("Src") and target-side ("Trg") rule-topic distributions. Phrase rules are fully lexicalized, while monotone and reordering rules contain nonterminals.

In order to gain insights into why the topic dissimilarity model is helpful for translation rule selection, we further investigate how many rules are topic-sensitive. As described in Section 4.2, we use entropy to measure whether a translation rule is topic-sensitive based on its rule-topic distribution. If the entropy of a translation rule calculated by the equation (8) is smaller than a certain threshold, the rule is topic-sensitive. Since documents often focus on a few topics, we use the average entropy of document-topic distributions of all training documents as the threshold. We compare entropies of source-side and target-side rule-topic distributions against this threshold. Our findings are shown in Table 4. 85.5% translation rules are topic-sensitive rules if we compare entropies of their source-side rule-topic distributions against the threshold. If we compare entropies of target-side rule-topic distributions against the threshold, topic-sensitive rules account for 86%. These strongly suggest that most rules only occur in documents with specific topics and topic information can be used to improve translation rule selection.

7.2.2 EFFECT OF TOPIC SENSITIVITY MODEL

As we can see from Table 4, there are still about 15% translation rules which are generic, not sensitive to any topics. These rules are also widely used in documents. As mentioned before, our topic dissimilarity model always punishes such rules as documents are normally topic-specific. We therefore introduce a topic sensitivity model to complement the topic dissimilarity model. The experiment result of this model is show in the last line of Table 3. We obtain a further improvement of 0.23 BLEU points when incorporating the topic sensitivity model. This indicates that it is necessary to distinguish topic-insensitive and topic-sensitive rules.

7.2.3 COMPARISON WITH TOPIC-SPECIFIC LEXICON TRANSLATION MODEL

We also compared our topic models against the topic-specific lexicon translation model proposed by Zhao and Xing (2007). They introduce a framework to combine Hidden Markov Model (HMM) and LDA topic model for SMT, which is shown in Figure 2. In their framework, each bilingual sentence pair has a single topic assignment sampled from the document-pair topic distribution θ. Then all words of the target language (e.g., English) are sampled given the sentence-pair topic assignment and a monolingual per-topic word distribution φ. After that, word alignments and words of the source language are sampled from a first-order Markov process and a topic-specific translation lexicon respectively.

Zhao and Xing integrate the topic-specific word-to-word translation lexicons estimated from their bilingual topic model described above into the topic-specific lexicon translation model, which is formulated as follows.

$$\begin{aligned} P(w_e|w_f, d_f) &\propto P(w_f|w_e, d_f)P(w_e|d_f) \\ &= \sum_k P(w_f|w_e, z=k)P(w_e|z=k)P(z=k|d_f) \end{aligned} \tag{14}$$

In this model, the probability of a candidate translation w_e for a source word w_f in a source document d_f is calculated by marginalizing over all topics and corresponding topic-specific translation lexicons. We simplify the estimation of $p(w_f|w_e, z=k)$ by directly computing these probabilities on our word-aligned corpus associated with target-side topic assignments that are inferred from the target-side topic model. Despite this simplification, the improvement of our implementation is comparable with the improvement obtained by Zhao and Xing (2007). Given a new document, we need to adapt the lexical translation weights of rules. The adapted lexicon translation model is integrated as a new feature into the log-linear discriminative framework.

We show the comparison results in Table 3. The topic-specific lexicon translation model is better than the baseline by 0.4 BLEU points. However, our topic-based method (the combination of topic dissimilarity and sensitivity models) outperforms the baseline by 0.87 BLEU points.

We also compare these two methods in terms of the decoding speed (words/second). The baseline translates 12.6 words per second, while the system with the topic-specific lexicon translation model only translates 3.3 words in one second. The overhead of the topic-specific lexicon translation model mainly comes from the adaptation of lexical weights. It takes 72.8% of the time to do the adaptation. 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.

7.3 Large-Scale Experiments

In this section, we investigated deeper into our models with the second group of experiments on large-scale training data.

7.3.1 EFFECT OF ONE-TO-MANY PROJECTION

As we discussed in Section 5.2, we need to project target-side topics onto source-side topic space so as to calculate the dissimilarity between a target-side rule-topic distribution and a source-side document-topic distribution. We propose a one-to-many projection method for this issue. In order to investigate the effectiveness of this method, we conducted experiments with large-scale training data to compare it with the following 3 other methods.

- *One-to-One Mapping* We enforce a one-to-one mapping between source-side and target-side topics, similar to the method by Tam et al. (2007). We achieve this by aligning a target-side topic to the corresponding source-side topic with the largest correspondence probability as calculated in Section 5.2.
- *Marginalization over Word Alignments* Following Zhao and Xing (2007), we first obtain topics on the target side using LDA and then retrieve topics of the source language through a marginalization over word alignments as follows.

$$P(w_f|k) = \sum_{w_e} P(w_f|w_e)P(w_e|z = k) \tag{15}$$

- *Combination of the source and target language documents* We concatenate each target document and its aligned source document into one document. We then run the LDA tool on these combined documents to train one topic model with mixed-language words. During decoding, we use the trained topic model to infer topics only on source documents.

In order to compare our one-to-many projection method with the three methods described above, we only add the target-side topic dissimilarity feature (DsimTrg) to the log-linear translation model. The experiment results are reported in Table 5. Clearly, all the four methods achieve improvements over the baseline. However, our one-to-many projection method performs better than all three other methods. In particular,

- Our method outperforms the one-to-one topic mapping method, which indicates that source-side and target-side topics do not exactly match in a one-to-one correspondence manner.
- The reason that the marginalization method performs the worse among the four methods may be that the topic model is trained only on target documents.
- Surprisingly, the combination method performs quite well. This shows that the LDA model can find hidden topics even on mixed-language documents.

7.3.2 EFFECT OF THE TOPIC-BASED RULE SELECTION FRAMEWORK ON VARIOUS TYPES OF RULES

We conducted experiments to further investigate the effect of our topic-based models for various types of rules selection. Particularly, we divide translation rules in hierarchical phrase-based SMT into three types: 1) phrase rules, which only contain terminals and are the same as bilingual phrase pairs used in phrase-based system; 2) monotone rules, which contain non-terminals and produce

System	MT06	MT08	Avg
Baseline	31.77	24.89	28.33
One-to-One	32.15	25.32	28.73
Marginalization	32.23	24.99	28.61
Combination	32.17	25.56	28.86
One-to-Many	32.44	25.54	28.99

Table 5: Effect of our one-to-many topic projection method against other methods. Marginalization: Marginalization over Word Alignments; Combination: Combination of the source and target language documents.

System	MT06	MT08	Avg
Baseline	31.77	24.89	28.33
Phrase rule	32.43	25.53	28.98
Monotone rule	32.24	25.62	28.93
Reordering rule	31.82	25.15	28.48
All	32.77	26.29	29.53

Table 6: Effect of our topic-based rule selection models on three types of rules. Phrase rules are fully lexicalized, while monotone and reordering rules contain nonterminals.

monotone translations; and finally 3) 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).

When we study the impact of our topic-based models on translation rule type A, we activate all of the four features described in Section 6 only on those rules of type A. Topic dissimilarity and sensitivity features on the other two types of translation rules are deactivated.

Table 6 shows the experiment results. From the table, we can observe that

- Our topic-based models achieve the highest improvement of 0.65 BLEU points over the baseline on phrase rules among the three types of translation rules. This is reasonable as phrase rules consist of topical words.
- We also obtain improvements of 0.6 and 0.15 BLEU points over the baseline on the monotone and reordering rules respectively. This shows that our models are also able to help select appropriate translation rules with non-terminals.
- When we activate the topic dissimilarity and sensitivity models on all translation rules, we can still achieve an additional improvement of 0.55 BLEU points. In total, our models outperform the baseline by an absolute improvement of 1.2 BLEU points.

7.3.3 EFFECT OF MORE MONOLINGUAL DATA

Comparing Table 6 and Table 3, we find that our topic-based dissimilarity and sensitivity models trained with medium-scale data (about 10K documents) collectively achieve an improvement of 0.87

System	MT06	MT08	Avg
Baseline	31.77	24.89	28.33
DsimSrc + SenSrc + DsimTrg + SenTrg	32.77	26.29	29.53
DsimSrc + **SenSrc** + DsimTrg + SenTrg	32.70	25.91	29.31
DsimSrc + SenSrc + **DsimTrg** + **SenTrg**	32.37	25.80	29.09
DsimSrc + **SenSrc** + **DsimTrg** + **SenTrg**	32.61	25.66	29.13

Table 7: Effect of using more monolingual data to train topic models. The features in bold are the topic-based dissimilarity/sensitivity model where the LDA topic model is trained using the combination of source/target part of the large-scale bilingual data and the corresponding monolingual corpus.

BLEU points over the baseline while the two models trained with large-scale data (about 100K documents) obtain an improvement of 1.2 BLEU points. This suggests that further performance gains may be obtained if we have more data. As parallel bilingual data with document boundaries provided is not easily accessible, we try to collect monolingual data of the source or/and target language. Our interest is to study whether we can gain further improvements by using more monolingual data to train our topic models.

We used a Chinese monolingual corpus where documents were collected from the Chinese Sohu weblog in 2009.[8] The collected Chinese corpus contains 500K documents with 273.8M Chinese words. We also used an English monolingual corpus where documents were collected from the English Blog Authorship corpus (Schler, Koppel, Argamon, & Pennebaker, 2006). The English monolingual corpus consists of 371K documents with 98M English words. We combined this new Chinese corpus with the source part of our large-scale bilingual data to train a source-side LDA topic model ST. The English monolingual corpus is also combined with the target part of the large-scale bilingual data to train a target-side LDA topic model TT.

We then used the two topic models ST and TT to infer topics for the test sets. Topic information on the source and target part of the large-scale bilingual training data inferred by ST and TT was used to estimate source-side rule-topic distributions and projected target-side rule-topic distributions. In this way, we can obtain a new topic-based dissimilarity and sensitivity model on the source/target side.

Experiment results are shown in Table 7. Unfortunately, we can not obtain any further improvements by training topic models on larger data, such as the combination of Chinese monolingual corpus and the source part of our bilingual training data. Instead, the performance drops from 29.53 to 29.31 if we use the topic model ST to build the source-side dissimilarity and sensitivity features and to 29.09 if we adopt the topic model TT to build the target-side dissimilarity and sensitivity features.

One reason for the lower performance with larger topic model training data may be that we only use 30 topics. Using more topics may improve our models on these larger corpora. In order to investigate this, we conducted new experiments with more topics than 30. We trained our source-side topic model using the combination of source part of the large-scale bilingual data and the Sohu weblog data. Based on this topic model, we built our source-side topic dissimilarity model and

8. http://blog.sohu.com/.

System	MT06	MT08	Avg
Baseline	31.77	24.89	28.33
$K = 30$	32.32	25.41	28.86
$K = 50$	31.96	25.73	28.85
$K = 100$	32.26	25.53	28.90
$K = 200$	32.16	25.28	28.72

Table 8: Experiment results with different number of topics (K). Only the source-side topic dissimilarity model (DsimSrc) is integrated into the SMT system.

Test Set	MonoSrc	BiSrc	MonoBiSrc
MT06	0.359	0.238	0.297
MT08	0.232	0.136	0.261

Table 9: The Hellinger distances of the MT06/08 test sets to the Chinese monolingual corpus (MonoSrc) and the source part of the bilingual training data (BiSrc) as well as their combination (MonoBiSrc) in terms of their average document-topic distributions.

integrated it into our SMT system. Experiment results are shown in Table 8. From this table, we find that using more topics is not able to improve our model on these corpora.

Yet another reason may be that the additional monolingual corpus is not similar to the test sets in terms of their topic distributions. In order to examine this hypothesis, we inferred document-topic distributions for all documents in the test sets, the Chinese monolingual corpus and the source part of the bilingual corpus using the topic model ST. We then average these document-topic distributions and obtain four average document-topic distributions for MT06, MT08, the Chinese monolingual corpus and the source part of the bilingual corpus respectively. These average topic distributions can be approximated as the corpus-topic distributions over the four corpora. We calculate the Hellinger distances between the corpus-topic distributions of the test sets and those of the Chinese monolingual corpus and the source part of the bilingual training data, which are shown in Table 9.

From the table, we can clearly find that the additional monolingual corpus is much less similar to the test sets comparing with the bilingual training corpus. The Hellinger distance of the test set MT08 to the MonoBiSrc corpus is almost twice as large as that to the bilingual training data (0.261 vs. 0.136). A topic model trained on such an enlarged corpus will make our topic-based models select translation rules that are not similar to the documents of the test sets in terms of topic distributions. This suggests that we should select additional monolingual data that are similar to the test sets if we want to obtain further improvements.

We further conducted a new group of experiments to empirically examine this hypothesis by translating a web-domain test set that is similar to the additional weblog corpus in terms of their topics. We used the web portion of the NIST MT06 set as our new development set and the web portion of the NIST MT08 as the new test set. Results are displayed in Table 10, which show that the additional monolingual data can improve the performance this time. This again suggests that we should select monolingual corpus that is similar to our test sets to learn topics for our topic-based dissimilarity and sensitivity models.

System	MT08-web
Baseline	20.45
DsimSrc + SenSrc + DsimTrg + SenTrg	21.42
DsimSrc + **SenSrc** + DsimTrg + SenTrg	21.77

Table 10: Results of translating a web-domain test set with our topic-based models trained on the data augmented with the monolingual weblog corpus. The features in bold are the topic-based dissimilarity/sensitivity model where the LDA topic model is trained using the combination of source/target part of the large-scale bilingual data and the corresponding monolingual corpus. MT08-web is the web portion of the NIST MT08 test set.

8. Analysis

In this section, we will study more details of our topic-based models for translation rule selection by looking at the differences that they make on target documents and individual translation hypotheses. These differences will help us gain some insights into how the presented models improve translation quality. In the analysis, both the baseline system and the system that is enhanced with the proposed topic-based models (all four features in Section 6 activated) are trained with the large-scale bilingual data as described in Section 7.1. For notational convenience, hereafter we refer to the baseline system as BASE and the system enhanced with our topic-based dissimilarity and sensitivity models as TOPSEL.

8.1 Differences on Target Documents

In order to measure the impact that our topic-based models have on target documents, we calculate the Hellinger distances between target documents generated by the BASE/TOPSEL system and reference documents generated by human in terms of their topics inferred by the target-side LDA topic model according to the following 4 steps. The target-side LDA topic model is trained on the target part of the large-scale bilingual data described in Section 7.1.

- Use the target-side LDA topic model to infer document-topic distribution for each document in reference translations (called reference distribution).
- Use the target-side LDA topic model to infer document-topic distribution for each target document generated by the BASE system (called BASE distribution).
- Similarly, we can obtain TOPSEL distribution on each target document generated by the TOPSEL system.
- Calculate the dissimilarity between the BASE and reference distribution as well as that between the TOPSEL and reference distribution according to the equation (6). These dissimilarities are first averaged on all documents and then averaged on four reference translations.

Table 11 shows the calculated dissimilarities. According to the equation (6), the smaller the Hellinger distance between two items, the more similar they are. The average Hellinger distance between TOPSEL and reference documents is 0.123 while the distance between BASE and reference

System	MT06	MT08	Avg
BASE	0.119	0.137	0.128
TOPSEL	0.116	0.129	0.123

Table 11: Dissimilarities (measured by Hellinger distance) between reference documents and target documents generated by the BASE and TOPSEL system in terms of their topics according to the equation (6).

	MT06	MT08
More similar (+)	49	72
Less similar (-)	30	37
$p <$	0.05	0.01

Table 12: The number of target documents generated by TOPSEL that are more/less similar to reference documents than those by BASE.

documents is 0.128. Therefore, the target documents generated by the TOPSEL system in both MT06 and MT08 are more similar to the documents in reference translations than those by the baseline system. We further calculate the number of target documents generated by TOPSEL that are more/less similar to reference documents than those by BASE based on the average Hellinger distances. These numbers are shown in Table 12. According to a sign test using these numbers, TOPSEL is statistically significantly better than the baseline system in terms of the similarity of translations generated by the two systems to human-generated translations.

8.2 Differences on Translation Hypotheses

We now look deeper into translation hypotheses to understand how our models select translation rules. Table 13 shows three translation examples that compare the baseline against the system enhanced with our topic-based models. In order to conduct a quantitative comparison, we calculate dissimilarity values (measured by Hellinger distance) of all underlined phrases in Table 13 using our topic-based dissimilarity model. The dissimilarity values are computed between the projected target-side rule-topic distributions of the underlined phrases and the source-side document-topic distributions of the corresponding documents where these phrase are used. The values are shown in Table 14.

From the two tables, we can easily observe that the system with the topic-based dissimilarity model prefers those target phrases that have smaller Hellinger distances to the documents where they occur in terms of topic distributions. In contrast, the baseline is not able to use this document-level topic information for translation rule selection. Figure 3 further shows the topic distributions of the source-side document, the TOPSEL phrase "allow" and the BASE phrase "permit" in Eg. 2. The major topics of the source-side document are topic 12 and 36. The TOPSEL phrase "allow" mainly distributes over 12 different topics[9] including topic 12 and 36 while the BASE phrase "permit" mainly over 10 different topics which do not include topic 12.

9. The distribution probability over these topics is larger than 0.03.

Eg. 1	Source	他 指 「 北 限 线 」 并 **不 合法** 。
	BASE	he described the “ northern limit line ” and unlawful .
	TOPSEL	he referred to the “ northern limit line ” is not legitimate .
	Reference	he pointed out that the “ northern limit line ” is not legitimate .
Eg. 2	Source	怎么 会 **允许** 自己 爱 的 人 也 同时 接受 别人 的 爱
	BASE	how would permit its love others also accepted by the people .
	TOPSEL	will allow their love of love others also accepted by the people
	Reference	how would someone allow the person he loves to accept other people’s love at the same time
Eg. 3	Source	目前 互联网 不 **享有** 这种 法定 许可 的 权利
	BASE	at present , the internet is not entitled to such a statutory right to leave
	TOPSEL	at present the internet does not enjoy such a statutory right to leave
	Reference	at present the internet does not enjoy these rights

Table 13: Translation examples from the NIST MT06/08 test sets, comparing the baseline with the system enhanced with the topic-based models. The underlined words highlight the difference between the enhanced models and the baseline.

Phrase	HD
unlawful	3.08
not legitimate	2.27
permit	3.75
allow	3.47
entitled to	3.45
enjoy	3.24

Table 14: Dissimilarity values (measured by Hellinger distance) of the underlined phrases in Table 13 between their projected target-side rule-topic distributions and the corresponding source-side document-topic distributions of documents calculated by our topic-based dissimilarity model.

9. Discussion on Bilingual Topic Modeling

Although topic models are widely adopted in monolingual text analysis, bilingual or multilingual topic models are less explored, especially those tailored for multilingual tasks such as machine translation. In this section we try to provide some suggestions for bilingual topic modeling from the perspective of statistical machine translation as well as our practice on the integration of topic models into SMT. These suggestions are listed as follows, some of which are also our future directions.

- *Investigation on Topic divergences across different languages* Cross-language divergences are pervasive and become one of big challenges for machine translation (Dorr, 1994). Such language-level divergences hint that divergences at the topic or concept level may also exist across languages. This may explain why our one-to-many topic projection from the target side

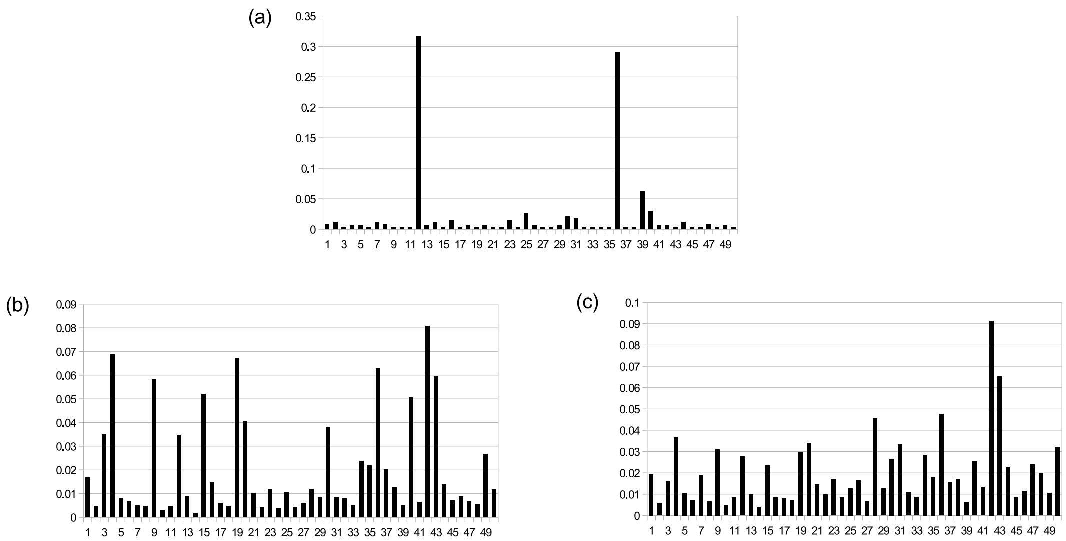


Figure 3: Topic distributions of the source-side document (a), the TOPSEL phrase "allow" (b) and the BASE phrase "permit" shown in Eg. 2 of Table 13.

to the source side is better than the one-to-one mapping. Although Mimno et al. (2009) have studied on topic divergences using Wikipedia articles, we believe that a deeper and wider investigation on topic divergence is needed as it will shed new light on how we can build better bilingual topic models.

- *Adding more linguistic assumptions into topic modeling* Practices in SMT show that integrating more linguistic knowledge into machine translation normally generates better translations (Chiang et al., 2008). We believe that adding more linguistic assumptions beyond bag-of-words will also improve topic modeling. A flexible topic modeling framework that allows us to integrate rich linguistic knowledge in the form of features will definitely further facilitate the application of topic models in natural language processing.

- *Joint modeling of topic induction and synchronous grammar induction* Synchronous grammar induction for machine translation is a task of automatically learning translation rules from bilingual data (Blunsom, Cohn, Dyer, & Osborne, 2009; Xiao & Xiong, 2013). As Bayesian approaches are successfully used in both topic modeling and synchronous grammar induction, joint modeling of them is an very interesting direction, which will also benefit grammar adaptation from one domain to another domain in machine translation.

10. Conclusions

In this article we have presented a topic-based translation rule selection framework which incorporates the topic information from both the source and target language for translation rule disambiguation. Particularly, we use a topic dissimilarity model to select appropriate translation rules for documents according to the similarities between translation rules and documents. We also adopt a topic sensitivity model to complement the topic dissimilarity model in order to balance translation rule selection between topic-sensitive and topic-insensitive rules. In order to calculate dissimilarities between source- and target-side topic distributions, we project topic distributions on the target side onto the source-side topic model space in a new and efficient way.

We have integrated our topic-based rule selection models into a hierarchical phrase-based SMT system. Experiments on medium/large-scale training data show that

- Our topic dissimilarity and sensitivity model are able to substantially improve translation quality in terms of BLEU and improve translation rule selection on various types of rules (i.e., phrase/monotone/reordering rules).
- Our method is better than previous topic-specific lexicon translation method in both translation quality and decoding speed.
- The proposed one-to-many projection method also outperforms various other methods such as one-to-one mapping, marginalization via word alignments and so on.
- If we want to use additional monolingual corpus to train topic models, we should first investigate whether the new monolingual corpus is similar to the test data in terms of topic distributions.

Topic models can provide global and document-level information for machine translation. In the future, we would like to use topic models to address document-level machine translation issues, such as coherence and cohesion (Barzilay & Lee, 2004; Hardmeier, Nivre, & Tiedemann, 2012). We also want to integrate our topic-based models into linguistically syntax-based machine translation for syntactic translation rule selection (Liu et al., 2006).

Acknowledgments

The work was sponsored by the National Natural Science Foundation of China under projects 61373095 and 61333018. Qun Liu's work was partially supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the CNGL at Dublin City University. We would like to thank three anonymous reviewers for their insightful comments. The corresponding author of this article is Deyi Xiong.

References

Barzilay, R., & Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In Susan Dumais, D. M., & Roukos, S. (Eds.), *HLT-NAACL 2004: Main Proceedings*, pp. 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *AAS*, *1*(1), 17–35.

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *JMLR*, *3*, 993–1022.

Blunsom, P., Cohn, T., Dyer, C., & Osborne, M. (2009). A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 782–790, Suntec, Singapore. Association for Computational Linguistics.

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 75–82, Arlington, Virginia, United States. AUAI Press.

Carpuat, M., & Wu, D. (2007a). How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 43–52.

Carpuat, M., & Wu, D. (2007b). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 61–72, Prague, Czech Republic. Association for Computational Linguistics.

Chan, Y. S., Ng, H. T., & Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 33–40, Prague, Czech Republic. Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proc. ACL 2005*.

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

Chiang, D., Marton, Y., & Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 224–233, Honolulu, Hawaii. Association for Computational Linguistics.

Clark, J. H., Dyer, C., Lavie, A., & Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 176–181, Portland, Oregon, USA.

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, *20*(4), 597–633.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pp. 128–135, Prague, Czech Republic.

Gong, Z., Zhang, M., & Zhou, G. (2011). Cache-based document-level statistical machine translation. In *Proc. EMNLP 2011*.

Gong, Z., Zhang, Y., & Zhou, G. (2010). Statistical machine translation based on LDA. In *Proc. IUCS 2010*, p. 286 – 290.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228–5235.

Hardmeier, C., Nivre, J., & Tiedemann, J. (2012). Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1179–1190, Jeju Island, Korea. Association for Computational Linguistics.

He, Z., Liu, Q., & Lin, S. (2008). Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 321–328, Manchester, UK. Coling 2008 Organizing Committee.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proc. of UAI 1999*, pp. 289–296.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pp. 388–395, Barcelona, Spain.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proc. HLT-NAACL 2003*.

Lin, J. (2006). Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theor.*, *37*(1), 145–151.

Liu, Q., He, Z., Liu, Y., & Lin, S. (2008). Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 89–97, Honolulu, Hawaii. Association for Computational Linguistics.

Liu, Y., Liu, Q., & Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proc. ACL 2006*.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proc. of EMNLP 2009*.

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL 2002*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. ACL 2003*.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL 2002*.

Ruiz, N., & Federico, M. (2011). Topic adaptation for lecture translation through bilingual latent semantic models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 199–205.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proc. ICSLP 2002*.

Tam, Y.-C., Lane, I. R., & Schultz, T. (2007). Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, *21*(4), 187–207.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for*

Natural Language Processing, pp. 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Ture, F., Oard, D. W., & Resnik, P. (2012). Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 417–426, Montréal, Canada. Association for Computational Linguistics.

Xiao, T., Zhu, J., Yao, S., & Zhang, H. (2011). Document-level consistency verification in machine translation. In *Proceedings of the 2011 MT summit XIII*, pp. 131–138, Xiamen, China.

Xiao, X., & Xiong, D. (2013). Max-margin synchronous grammar induction for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 255–264, Seattle, Washington, USA. Association for Computational Linguistics.

Xiao, X., Xiong, D., Zhang, M., Liu, Q., & Lin, S. (2012). A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 750–758, Jeju Island, Korea. Association for Computational Linguistics.

Xiong, D., Zhang, M., & Li, H. (2012). Modeling the translation of predicate-argument structure for SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 902–911, Jeju Island, Korea. Association for Computational Linguistics.

Zhao, B., & Xing, E. P. (2007). HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Proc. NIPS 2007*.

Zhao, B., & Xing, E. P. (2006). BiTAM: Bilingual topic admixture models for word alignment. In *Proc. ACL 2006*.