A Context-Aware Topic Model for Statistical Machine Translation

Jinsong Su[1], **Deyi Xiong**[2*], **Yang Liu**[3], **Xianpei Han**[4], **Hongyu Lin**[1],
Junfeng Yao[1], **Min Zhang**[2]

Xiamen University, Xiamen, China[1]
Soochow University, Suzhou, China[2]
Tsinghua University, Beijing, China[3]
Institute of Software, Chinese Academy of Sciences, Beijing, China[4]
{jssu, hylin, yao0010}@xmu.edu.cn
{dyxiong, minzhang}@suda.edu.cn
liuyang2011@tsinghua.edu.cn
xianpei@nfs.iscas.ac.cn

Abstract

Lexical selection is crucial for statistical machine translation. Previous studies separately exploit sentence-level contexts and document-level topics for lexical selection, neglecting their correlations. In this paper, we propose a context-aware topic model for lexical selection, which not only models local contexts and global topics but also captures their correlations. The model uses target-side translations as hidden variables to connect document topics and source-side local contextual words. In order to learn hidden variables and distributions from data, we introduce a Gibbs sampling algorithm for statistical estimation and inference. A new translation probability based on distributions learned by the model is integrated into a translation system for lexical selection. Experiment results on NIST Chinese-English test sets demonstrate that 1) our model significantly outperforms previous lexical selection methods and 2) modeling correlations between local words and global topics can further improve translation quality.

1 Introduction

Lexical selection is a very important task in statistical machine translation (SMT). Given a sentence in the source language, lexical selection statistically predicts translations for source words, based on various translation knowledge. Most conventional SMT systems (Koehn et al., 2003; Galley et al., 2006; Chiang, 2007) exploit very limited context information contained in bilingual rules for lexical selection.
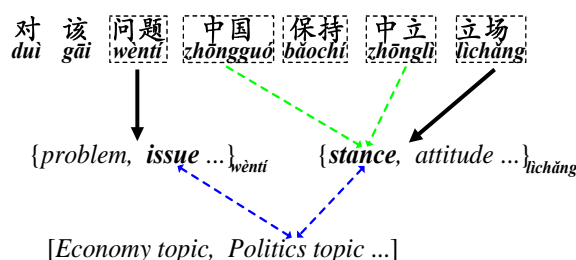
---

*Corresponding author.

Figure 1: A Chinese-English translation example to illustrate the effect of local contexts and global topics as well as their correlations on lexical selection. Each black line indicates a set of translation candidates for a Chinese content word (within a dotted box). Green lines point to translations that are favored by local contexts while blue lines show bidirectional associations between global topics and their consistent target-side translations.

Previous studies that explore richer information for lexical selection can be divided into two categories: 1) incorporating sentence-level contexts (Chan et al., 2007; Carpuat and Wu, 2007; Hasan et al., 2008; Mauser et al., 2009; He et al., 2008; Shen et al., 2009) or 2) integrating document-level topics (Xiao et al., 2011; Ture et al., 2012; Xiao et al., 2012; Eidelman et al., 2012; Hewavitharana et al., 2013; Xiong et al., 2013; Hasler et al., 2014a; Hasler et al., 2014b) into SMT. The methods in these two strands have shown their effectiveness on lexical selection.

However, correlations between sentence- and document-level contexts have never been explored before. It is clear that local contexts and global topics are often highly correlated. Consider a Chinese-English translation example presented in Figure 1. On the one hand, if local contexts suggest that the source word "立场/lìchǎng" should be translated in-

to "*stance*", they will also indicate that the topic of the document where the example sentence occurs is about *politics*. The *politics* topic can be further used to enable the decoder to select a correct translation "*issue*" for another source word "问题/*wèntí*", which is consistent with this topic. On the other hand, if we know that this document mainly focuses on the *politics* topic, the candiate translation "*stance*" will be more compatible with the context of "立场/*lìchǎng*" than the candiate translation "*attitude*". This is because neighboring source-side words "中国/*zhōnguó*" and "中立/*zhōnglì*" often occur in documents that are about international politics. We believe that such correlations between local contextual words and global topics can be used to further improve lexical selection.

In this paper, we propose a unified framework to jointly model local contexts, global topics as well as their correlations for lexical selection. Specifically,

- First, we present a context-aware topic model (CATM) to exploit the features mentioned above for lexical selection in SMT. To the best of our knowledge, this is the first work to jointly model both local and global contexts for lexical selection in a topic model.

- Second, we present a Gibbs sampling algorithm to learn various distributions that are related to topics and translations from data. The translation probabilities derived from our model are integrated into SMT to allow collective lexical selection with both local and global informtion.

We validate the effectiveness of our model on a state-of-the-art phrase-based translation system. Experiment results on the NIST Chinese-English translation task show that our model significantly outperforms previous lexical selection methods.

2 Context-Aware Topic Model

In this section, we describe basic assumptions and elaborate the proposed context-aware topic model.

2.1 Basic Assumptions

In CATM, we assume that each source document d consists of two types of words: *topical words* which are related to topics of the document and *contextual*

words which affect translation selections of topical words.

As topics of a document are usually represented by content words in it, we choose source-side *nouns*, *verbs*, *adjectives* and *adverbs* as topical words. For contextual words, we use all words in a source sentence as contextual words. We assume that they are generated by target-side translations of other words than themselves. Note that a source word may be both topical and contextual. For each topical word, we identify its candidate translations from training corpus according to word alignments between the source and target language. We allow a target translation to be a phrase of length no more than 3 words. We refer to these translations of source topical words as *target-side topical items*, which can be either words or phrases. In the example shown in Figure 1, all source words within dotted boxes are topical words. Topical word "立场/*lìchǎng*" is supposed to be translated into a target-side topical item "*stance*", which is collectively suggested by neighboring contextual words " 中国/*zhōngguó*", "中立/*zhōnglì*" and the topic of the corresponding document.

In our model, all target-side topical items in a document are generated according to the following two assumptions:

- *Topic consistency assumption*: All target-side topical items in a document should be consistent with the topic distribution of the document. For example, the translations "*issue*", "*stance*" tend to occur in documents about *politics* topic.

- *Context compatibility assumption*: For a topical word, its translation (i.e., the counterpart target-side topical item) should be compatible with its neighboring contextual words. For instance, the translation "*stance*" of "立场/*lìchǎng*" is closely related to contextual words "中国/*zhōnguó*" and "中立/*zhōnglì*".

2.2 Model

The graphical representation of CATM, which visualizes the generative process of training data $\boldsymbol{D}$, is shown in Figure 2. Notations of CATM are presented in Table 1. In CATM, each document d can be generated in the following three steps[1]:

---

[1]In the following description, $Dir(.)$, $Mult(.)$ and $Unif(.)$ denote *Dirichlet*, *Multinomial* and *Uniform* distributions, re-

Symbol	Meaning
α	hyperparameter for θ
β	hyperparameter for ϕ
γ	hyperparameter for ψ
δ	hyperparameter for ξ
f	topical word
c	contextual word
$\tilde{e}$	target-side topical item
$\tilde{e}'$	a sampled target-side topical item used to generate a source-side contextual word
θ	the topic distribution of document
ϕ	the distribution of a topic over target-side topical items
ψ	the translation probability distribution of a target-side topical item over source-side topical words
ξ	the generation probability distribution of a target-side topical item over source-side contextual words
N_z	topic number
N_d	document number
N_f	the number of topical words
N_c	the number of contextual words
$N_{\tilde{e}}$	the number of target-side topical items
$N_{f,d}$	the number of topical words in d
$N_{c,d}$	the number of contextual words in d

Table 1: Notations in CATM.

1. Sample a topic distribution $\theta_d \sim Dir(\alpha)$.
2. For each position i that corresponds to a topical word f_i in the document:
   (a) Sample a topic $z_i \sim Mult(\theta_d)$.
   (b) Conditioned on the topic z_i, sample a target-side topical item $\tilde{e}_i \sim Mult(\phi_{z_i})$.
   (c) Conditioned on the target-side topical item $\tilde{e}_i$, sample the topical word $f_i \sim Mult(\psi_{\tilde{e}_i})$.
3. For each position j that corresponds to a contextual word c_j in the document:
   (a) Collect all target-side topical items $\tilde{\boldsymbol{e}}_s$ that are translations of neighboring topical words within a window centered at c_j (window size w_s).
   (b) Randomly sample an item from $\tilde{\boldsymbol{e}}_s$, $\tilde{e}'_j \sim Unif(\tilde{\boldsymbol{e}}_s)$.
   (c) Conditioned on the sampled target-side topical item $\tilde{e}'_j$, sample the contextual word $c_j \sim Mult(\xi_{\tilde{e}'_j})$.

To better illustrate CATM, let us revisit the example in Figure 1. We describe how CATM generates top-
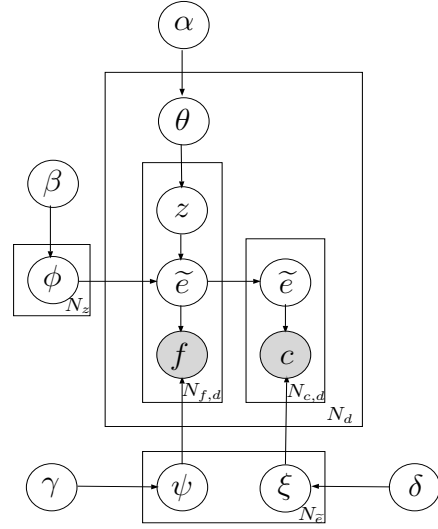
spectively.

Figure 2: Graphical representation of our model.

ical words "问题/*wèntí*", "立场/*lìchǎng*", and contextual word "中立/*zhōnglì*" in the following steps:

Step 1: The model generates a topic distribution for the corresponding document as $\{economy^{0.25}, politics^{0.75}\}$.

Step 2: Based on the topic distribution, we choose "*economy*" and "*politics*" as topic assignments for "问题/*wèntí*" and "立场/*lìchǎng*" respectively; Then, according to the distributions of the two topics over target-side topical items, we generate target-side topical items "*issue*" and "*stance*"; Finally, according to the translation probability distributions of these two topical items over source-side topical words, we generate source-side topical words "问题/*wèntí*" and "立场/*lìchǎng*" for them respectively.

Step 3: For the contextual word "中立/*zhōnglì*", we first collect target-side topical items of its neighboring topical words such as "问题/*wèntí*", "保持/*bǎochí*" and "立场/*lìchǎng*" to form a target-side topical item set $\{$"*issue*","*keep*", "*stance*"$\}$, from which we randomly sample one item "*stance*". Next, according to the generation probability distribution of "*stance*" over source contextual words, we finally generate the source contextual word "中立/*zhōnglì*".

In the above generative process, all target-side topical items are generated from the underlying topics of a source document, which guarantees that selected target translations are topic-consistent. Ad-

ditionally, each source contextual word is derived from a target-side topical item given its generation probability distribution. This makes selected target translations also compatible with source-side local contextual words. In this way, global topics, topical words, local contextual words and target-side topical items are highly correlated in CATM that exactly captures such correlations for lexical selection.

3 Parameter Estimation and Inference

We propose a Gibbs sampling algorithm to learn various distributions described in the previous section. Details of the learning and inference process are presented in this section.

3.1 The Probability of Training Corpus

According to CATM, the total probability of training data $\boldsymbol{D}$ given hyperparameters α, β, γ and δ is computed as follows:

$$
\begin{aligned}
P(\boldsymbol{D}; \alpha, \beta, \gamma, \delta) &= \prod_d P(\boldsymbol{f}_d, \boldsymbol{c}_d; \alpha, \beta, \gamma, \delta) \\
&= \prod_d \sum_{\tilde{\boldsymbol{e}}_d} P(\tilde{\boldsymbol{e}}_d|\alpha, \beta) P(\boldsymbol{f}_d|\tilde{\boldsymbol{e}}_d, \gamma) P(\boldsymbol{c}_d|\tilde{\boldsymbol{e}}_d, \delta) \\
&= \int_\phi P(\phi|\beta) \int_\psi P(\psi|\gamma) \prod_d \sum_{\tilde{\boldsymbol{e}}_d} P(\boldsymbol{f}_d|\tilde{\boldsymbol{e}}_d, \psi) \\
&\quad \times \int_\xi P(\xi|\delta) \sum_{\tilde{\boldsymbol{e}}'_d} P(\tilde{\boldsymbol{e}}'_d|\tilde{\boldsymbol{e}}_d) p(\boldsymbol{c}_d|\tilde{\boldsymbol{e}}'_d, \xi) \\
&\quad \times \int_\theta P(\theta|\alpha) P(\tilde{\boldsymbol{e}}_d|\theta, \phi) d\theta d\xi d\psi d\phi \quad (1)
\end{aligned}
$$

where $\boldsymbol{f}_d$ and $\tilde{\boldsymbol{e}}_d$ denote the sets of topical words and their target-side topical item assignments in document d, $\boldsymbol{c}_d$ and $\tilde{\boldsymbol{e}}'_d$ are the sets of contextual words and their target-side topical item assignments in document d.

3.2 Parameter Estimation via Gibbs Sampling

The joint distribution in Eq. (1) is intractable to compute because of coupled hyperparameters and hidden variables. Following Han et al, (2012), we adapt the well-known Gibbs sampling algorithm (Griffiths and Steyvers, 2004) to our model. We compute the joint posterior distribution of hidden variables, denoted by $P(\boldsymbol{z}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}'|\boldsymbol{D})$, and then use this distribution to 1) estimate θ, ϕ, ψ and ξ, and 2) predict translations and topics of all documents in $\boldsymbol{D}$.

Specifically, we derive the joint posterior distribution from Eq. (1) as:

$$
P(\boldsymbol{z}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}'|\boldsymbol{D}) \propto P(\boldsymbol{z}) P(\tilde{\boldsymbol{e}}|\boldsymbol{z}) P(\boldsymbol{f}|\tilde{\boldsymbol{e}}) P(\tilde{\boldsymbol{e}}'|\tilde{\boldsymbol{e}}) P(\boldsymbol{c}|\tilde{\boldsymbol{e}}') \quad (2)
$$

Based on the equation above, we construct a Markov chain that converges to $P(\boldsymbol{z}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}'|\boldsymbol{D})$, where each state is an assignment of a hidden variable (including topic assignment to a topical word, target-side topical item assignment to a source topical or contextual word.). Then, we sequentially sample each assignment according to the following three conditional assignment distributions:

1. $P(z_i = z|\boldsymbol{z}_{-i}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}', \boldsymbol{D})$: topic assignment distribution of a topical word given $\boldsymbol{z}_{-i}$ that denotes all topic assignments but z_i, $\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{e}}'$ that are target-side topical item assignments. It is updated as follows:

$$
\begin{aligned}
&P(z_i = z|\boldsymbol{z}_{-i}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}', \boldsymbol{D}) \propto \\
&\frac{C^{DZ}_{(-i)dz} + \alpha}{C^{DZ}_{(-i)d*} + N_z \alpha} \times \frac{C^{Z\tilde{E}}_{(-i)z\tilde{e}} + \beta}{C^{Z\tilde{E}}_{(-i)z*} + N_{\tilde{e}} \beta} \quad (3)
\end{aligned}
$$

where the topic assignment to a topical word is determined by the probability that this topic appears in document d (the 1st term) and the probability that the selected item $\tilde{e}$ occurs in this topic (the 2nd term).

2. $P(\tilde{e}_i = \tilde{e}|\boldsymbol{z}, \tilde{\boldsymbol{e}}_{-i}, \tilde{\boldsymbol{e}}', \boldsymbol{D})$: target-side topical item assignment distribution of a source topical word given the current topic assignments $\boldsymbol{z}$, the current item assignments of all other topical words $\tilde{\boldsymbol{e}}_{-i}$, and the current item assignments of contextual words $\tilde{\boldsymbol{e}}'$. It is updated as follows:

$$
\begin{aligned}
&P(\tilde{e}_i = \tilde{e}|\boldsymbol{z}, \tilde{\boldsymbol{e}}_{-i}, \tilde{\boldsymbol{e}}', \boldsymbol{D}) \propto \frac{C^{Z\tilde{E}}_{(-i)z\tilde{e}} + \beta}{C^{Z\tilde{E}}_{(-i)z*} + N_{\tilde{e}} \beta} \\
&\times \frac{C^{\tilde{E}F}_{(-i)\tilde{e}f} + \gamma}{C^{\tilde{E}F}_{(-i)\tilde{e}*} + N_f \gamma} \times \Big(\frac{C^{W\tilde{E}}_{(-i)w\tilde{e}} + 1}{C^{W\tilde{E}}_{(-i)w\tilde{e}}}\Big) C^{W\tilde{E}'}_{w\tilde{e}} \quad (4)
\end{aligned}
$$

where the target-side topical item assignment to a topical word is determined by the probability that this item is from the topic z (the 1st term), the probability that this item is translated into the topical word f (the 2nd term) and the probability of contextual words within a w_s word window centered at the topical word f, which influence the selection of the target-side topical item $\tilde{e}$ (the 3rd term). It is very important to note that we use a parallel corpus to train the model. Therefore we directly identify target-side topical items for source topical words via word alignments rather than sampling.

3. $P(\tilde{e}'_i = \tilde{e}|\boldsymbol{z}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}'_{-i}, \boldsymbol{D})$: target-side topical item assignment distribution for a contextual word given the current topic assignments $\boldsymbol{z}$, the current item assignments of topical words $\tilde{\boldsymbol{e}}$, and the current item assignments of all other contextual words $\tilde{\boldsymbol{e}}'_{-i}$. It is updated as follows:

$$P(\tilde{e}'_i = \tilde{e}|\boldsymbol{z}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}'_{-i}, \boldsymbol{D}) \propto$$

$$\frac{C_{w\tilde{e}}^{W\tilde{E}}}{C_{w*}^{W\tilde{E}}} \times \frac{C_{(-i)\tilde{e}c}^{\tilde{E}C} + \delta}{C_{(-i)\tilde{e}*}^{\tilde{E}C} + N_c\,\delta} \qquad (5)$$

where the target-side topical item assignment used to generate a contextual word is determined by the probability of this item being assigned to generate contextual words within a surface window of size w_s (the 1st term) and the probability that contextual words occur in the context of this item (the 2nd term).

In all above formulas, C_{dz}^{DZ} is the number of times that topic z has been assigned for all topical words in document d, $C_{d*}^{DZ}=\sum_z C_{dz}^{DZ}$ is the topic number in document d, and $C_{z\tilde{e}}^{Z\tilde{E}}$, $C_{\tilde{e}f}^{\tilde{E}F}$, $C_{w\tilde{e}}^{W\tilde{E}}$, $C_{w\tilde{e}}^{W\tilde{E}'}$ and $C_{\tilde{e}c}^{\tilde{E}C}$ have similar explanations. Based on the above marginal distributions, we iteratively update all assignments of corpus $\boldsymbol{D}$ until the constructed Markov chain converges. Model parameters are estimated using these final assignments.

3.3 Inference on Unseen Documents

For a new document, we first predict its topics and target-side topical items using the incremental Gibbs sampling algorithm described in (Kataria et al., 2011). In this algorithm, we iteratively update topic assignments and translation assignments of an unseen document following the same process described in Section 3.2, but with estimated model parameters.

Once we obtain these assignments, we estimate lexical translation probabilities based on the sampled counts of target-side topical items. Formally, for the position i in the document corresponding to the content word f, we collect the sampled count that translation $\tilde{e}$ generates f, denoted by $C_{sam}(\tilde{e}, f)$. This count can be normalized to form a new translation probability in the following way:

$$p(\tilde{e}|f) = \frac{C_{sam}(\tilde{e}, f) + k}{C_{sam} + k \cdot N_{\tilde{e},f}} \qquad (6)$$

where C_{sam} is the total number of samples during inference and $N_{\tilde{e},f}$ is the number of candidate translations of f. Here we apply add-k smoothing to refine this translation probability, where k is a tunable global smoothing constant. Under the framework of log-linear model (Och and Ney, 2002), we use this translation probability as a new feature to improve lexical selection in SMT.

4 Experiments

In order to examine the effectiveness of our model, we carried out several groups of experiments on Chinese-to-English translation.

4.1 Setup

Our bilingual training corpus is from the FBIS corpus and the Hansards part of LDC2004T07 corpus ($1M$ parallel sentences, $54.6K$ documents, with $25.2M$ Chinese words and $29M$ English words). We first used *ZPar* toolkit[2] and Stanford toolkit[3] to preprocess (i.e., word segmenting, PoS tagging) the Chinese and English parts of training corpus, and then word-aligned them using *GIZA++* (Och and Ney, 2003) with the option "*grow-diag-final-and*". We chose the NIST evaluation set of MT05 as the development set, and the sets of MT06/MT08 as test sets. On average, these three sets contain 17.2, 13.9 and 14.1 content words per sentence, respectively. We trained a 5-gram language model on the Xinhua portion of Gigaword corpus using the *SRILM* Toolkit (Stolcke, 2002).

Our baseline system is a state-of-the-art SMT system, which adapts bracketing transduction grammars (Wu, 1997) to phrasal translation and equips itself with a maximum entropy based reordering model (MEBTG) (Xiong et al., 2006). We used the toolkit[4] developed by Zhang (2004) to train the reordering model with the following parameters: iteration number $iter$=200 and Gaussian prior g=1.0. During decoding, we set the *ttable-limit* as 20, the *stack-size* as 100. The translation quality is evaluated by case-insensitive BLEU-4 (Papineni et al., 2002) metric. Finally, we conducted *paired bootstrap sampling* (Koehn, 2004) to test the significance in BLEU score differences.

Model	MT05
CATM ($\pm$ 6w)	33.35
CATM ($\pm$ 8w)	33.43
CATM ($\pm$ 10w)	33.42
CATM ($\pm$ 12w)	33.49
CATM ($\pm$ 14w)	33.30

Table 2: Experiment results on the development set using different window sizes w_s.

To train CATM, we set the topic number N_z as 25.[5] For hyperparameters α and β, we empirically set α=50/N_z and β=0.1, as implemented in (Griffiths and Steyvers, 2004). Following Han et al. (2012), we set γ and δ as $1.0/N_f$ and $2000/N_c$, respectively. During the training process, we ran 400 iterations of the Gibbs sampling algorithm. For documents to be translated, we first ran 300 rounds in a *burn-in* step to let the probability distributions converge, and then ran 1500 rounds where we collected independent samples every 5 rounds. The longest training time of CATM is less than four days on our server using 4GB RAM and one core of 3.2GHz CPU. As for the smoothing constant k in Eq. (6), we set its values to 0.5 according to the performance on the development set in additional experiments.

4.2 Impact of Window Size w_s

Our first group of experiments were conducted on the development set to investigate the impact of the window size w_s. We gradually varied window size from 6 to 14 with an increment of 2.

Experiment results are shown in Table 2. We achieve the best performance when w_s=12. This suggests that a $\pm$12-word window context is sufficient for predicting target-side translations for ambiguous source-side topical words. We therefore set w_s=12 for all experiments thereafter.

4.3 Overall Performance

In the second group of experiments, in addition to the conventional MEBTG system, we also compared CATM with the following two models:

Word Sense Disambiguation Model (**WSDM**) (Chan et al., 2007). This model improves lexical selection in SMT by exploiting local contexts. For

each content word, we construct a MaxEnt-based classifier incorporating local collocation and surrounding word features, which are also adopted by Chan et al. (2007). For each candidate translation $\tilde{e}$ of topical word f, we use WSDM to estimate the context-specific translation probability $P(\tilde{e}|f)$, which is used as a new feature in SMT system.

Topic-specific Lexicon Translation Model (**TLTM**) (Zhao and Xing, 2007). This model focuses on the utilization of document-level context. We adapted it to estimate a lexicon translation probability as follows:

$$p(f|\tilde{e}, d) \propto p(\tilde{e}|f, d) \cdot p(f|d)$$
$$= \sum_z p(\tilde{e}|f, z) \cdot p(f|z) \cdot p(z|d) \quad (7)$$

where $p(\tilde{e}|f, z)$ is the lexical translation probability conditioned on topic z, which can be calculated according to the principle of maximal likelihood, $p(f|z)$ is the generation probability of word f from topic z, and $p(z|d)$ denotes the posterior topic distribution of document d.

Note that our CATM is proposed for lexical selection on content words. To show the strong effectiveness of our model, we also compared it against the full-fledged variants of the above-mentioned two models that are built for all source words. We refer to them as **WSDM (All)** and **TLTM (All)**, respectively.

Table 3 displays BLEU scores of different lexical selection models. All models outperform the baseline. Although we only use CATM to predict translations for content words, CATM achieves an average BLEU score of **26.77** on the two test sets, which is higher than that of the baseline by **1.18** BLEU points. This improvement is statistically significant at p<0.01. Furthermore, we also find that our model performs better than WSDM and TLTM with significant improvements. Finally, even if WSDM (All) and TLTM (all) are built for all source words, they are still no better than than CATM that selects desirable translations for content words. These experiment results strongly demonstrate the advantage of CATM over previous lexical selection models.

5 Analysis

In order to investigate why CATM is able to outperform previous models that explore only local contex-

---

[5]We try different topic numbers from 25 to 100 with an increment of 25 each time. We find that N_z=25 produces a slightly better performance than other values on the development set.

Model	Local Context	Global Topic	MT06	MT08	Avg
Baseline	$\times$	$\times$	$29.66^{\downarrow\downarrow}$	$21.52^{\downarrow\downarrow}$	25.59
WSDM	$\checkmark$	$\times$	$30.62^{\downarrow}$	$22.05^{\downarrow\downarrow}$	26.34
WSDM (All)	$\checkmark$	$\times$	30.92	22.27	26.60
TLTM	$\times$	$\checkmark$	$30.27^{\downarrow\downarrow}$	$21.70^{\downarrow\downarrow}$	25.99
TLTM (All)	$\times$	$\checkmark$	$30.33^{\downarrow\downarrow}$	$21.58^{\downarrow\downarrow}$	25.96
CATM	$\checkmark$	$\checkmark$	**30.97**	**22.56**	**26.77**

Table 3: Experiment results on the test sets. **Avg** = average BLEU scores. **WSDM (All)** and **TLTM (All)** are models built for all source words. $\downarrow$: significantly worse than CATM ($p<0.05$), $\downarrow\downarrow$: significantly worse than CATM ($p<0.01$)

.

tual words or global topics, we take a deep look into topics, topical items and contextual words learned by CATM and empirically analyze the effect of modeling correlations between local contextual words and global topics on lexical selection.

5.1 Outputs of CATM

We present some examples of topics learned by CATM in Table 4. We also list five target-side topical items with the highest probabilities for each topic, and the most probable five contextual words for each target-side topical item. These examples clearly show that target-side topical items tightly connect global topics and local contextual words by capturing their correlations.

5.2 Effect of Correlation Modeling

Compared to previous lexical selection models, CATM jointly models both local contextual words and global topics. Such a joint modeling also enables CATM to capture their inner correlations at the model level. In order to examine the effect of correlation modeling on lexical selection, we compared CATM with its three variants: ① **CATM (Context)** that only uses local context information. We determined target-side topical items for content words in this variant by setting the probability distribution that a topic generates a target-side topical item to be uniform; ② **CATM (Topic)** that explores only global topic information. We identified target-side topical items for content words in the model by setting w_s as 0, i.e., no local contextual words being used at all. ③ **CATM (Log-linear)** is the combination of the above-mentioned two variants (① and ②) in a log-linear manner, which does not capture correlations between local contextual words and global topics at the model level.

Model	MT06	MT08	Avg
CATM (Context)	$30.46^{\downarrow\downarrow}$	$22.02^{\downarrow\downarrow}$	26.24
CATM (Topic)	$30.20^{\downarrow\downarrow}$	$21.90^{\downarrow\downarrow}$	26.05
CATM (Log-linear)	$30.59^{\downarrow}$	$22.24^{\downarrow}$	26.42
CATM	**30.97**	**22.56**	**26.77**

Table 5: Experiment results on the test sets. CATM (Log-linear) is the combination of CATM (Context) and CATM (Topic) in a log-linear manner.

Results in Table 5 show that CATM performs significantlly better than both CATM (Topic) and CATM (Context). Even compared with CATM (Log-linear), CATM still achieves a significant improvement of 0.35 BLEU points ($p<0.05$). This validates the effectiveness of capturing correlations for lexical selection at the model level.

6 Related Work

Our work is partially inspired by (Han and Sun, 2012), where an entity-topic model is presented for entity linking. We successfully adapt this work to lexical selection in SMT. The related work mainly includes the following two strands.

(1) *Lexical Selection in SMT*. In order to explore rich context information for lexical selection, some researchers propose *trigger-based lexicon models* to capture long-distance dependencies (Hasan et al., 2008; Mauser et al., 2009), and many more researchers build classifiers to select desirable translations during decoding (Chan et al., 2007; Carpuat and Wu, 2007; He et al., 2008; Liu et al., 2008). Along this line, Shen et al. (2009) introduce four new linguistic and contextual features for translation selection in SMT. Recently, we have witnessed an increasing efforts in exploiting document-level context information to improve lexical selection. Xiao et al. (2011) impose a hard constraint to guarantee

Topic	Target-side Topical Items	Source-side Contextual Words
refugee	*UNHCR*	难民(refugee) 办事处(office) 专员(commissioner) 事务(affair) 高级(high-level)
	republic	联盟(union) 民主(democracy) 政府(government) 伊斯兰(Islam) 中非(Central Africa)
	refugee	难民(refugee) 返回(return) 流离失所(displaced) 遣返(repatriate) 保护(protect)
	Kosovo	梅托希亚(Metohija) 境内(territory) 危机(crisis) 局势(situation) 塞尔维亚(Serbia)
	federal	共和国(republic) 南斯拉夫(Yugoslavia) 科索沃(Kosovo) 政府(government) 当局(authority)
military	*military*	观察员(observer) 行动(action) 美国(USA) 人员(personnel) 部队(army)
	missile	防御(defense) 系统(system) 美国(USA) 发射(launch) 枚(*)
	United States	中国(China) 日本(Japan) 台湾(Taiwan) 军事(military) NMD(National Missile Defense)
	system	联合国(United Nations) 建立(build) 国(country) 国家(country) 信息(information)
	war	战争(war) 场(*) 世界(world) 发动(wage) 海湾(gulf)
economy	*country*	发展中(developing) 发达(developed) 非洲(Africa) 发展(development) 中(China)
	development	可持续(sustainable) 经济(economy) 促进(promote) 社会(society) 事态(situation)
	international	社会(society) 组织(organization) 合作(coorporation) 国家(country) 联合国(United Nations)
	economic	社会(society) 发展(development) 增长(growth) 国家(country) 全球化(globalization)
	trade	发展(development) 国际(international) 世界(world) 投资(investment) 点(point)
cross-strait relation	*Taiwan*	中国(China) 大陆(mainland) 当局(authority) 美国(USA) 同胞(compatriot)
	China	说(say) 美国(USA) 台湾(Taiwan) 原则(principle) 两(*)
	relation	发展(development) 岸(*) 中(China) 两(*) 国(country)
	cross-strait	两(*) 关系(relation) 台湾(Taiwan) 岸(*) 交流(exchange)
	issue	解决(settlement) 讨论(discuss) 问题(issue) 重要(important) 台湾(Taiwan)

Table 4: Examples of topics, topical items and contextual words learned by CATM with N_z=25 and W_s=12. Chinese words that do not have direct English translations are denoted with "*". Here "枚" and "场" are Chinese quantifiers for *missile* and *war*, respectively; "两" and "岸" together means *cross-starit*.

the document-level translation consistency. Ture et al. (2012) soften this consistency constraint by integrating three counting features into decoder. Also relevant is the work of Xiong et al.(2013), who use three different models to capture lexical cohesion for document-level SMT.

(2) *SMT with Topic Models*. In this strand, Zhao and Xing (2006, 2007) first present a bilingual topical admixture formalism for word alignment in SMT. Tam et al. (2007) and Ruiz et al. (2012) apply topic model into language model adaptation. Su et al. (2012) conduct translation model adaptation with monolingual topic information. Gong et al. (2010) and Xiao et al. (2012) introduce topic-based similarity models to improve SMT system. Axelrod et al. (2012) build topic-specific translation models from the TED corpus and select topic-relevant data from the UN corpus to improve coverage. Eidelman et al. (2012) incorporate topic-specific lexical weights into translation model. Hewavitharana et al. (2013) propose an incremental topic based translation model adaptation approach that satisfies the causality constraint imposed by spoken conversations. Hasler et al. (2014) present a new bilingual variant of LDA to compute topic-adapted, probabilistic phrase translation features. They also use a topic model to learn latent distributional representations of different context levels of a phrase pair (Hasler et al., 2014b).

In the studies mentioned above, those by Zhao and Xing (2006), Zhao and Xing (2007), Hasler et al. (2014a), and Hasler et al. (2014b) are most related to our work. However, they all perform dynamic translation model adaptation with topic models. Significantly different from them, we propose a new topic model that exploits both local contextual words and global topics for lexical selection. To the best of our knowledge, this is first attempt to capture correlations between local words and global topics for better lexical selection at the model level.

7 Conclusion and Future Work

This paper has presented a novel context-aware topic model for lexical selection in SMT. Jointly modeling local contexts, global topics and their correlations in a unified framework, our model provides an effective way to capture context information at different levels for better lexical selection in SMT. Experiment results not only demonstrate the effectiveness of the proposed topic model, but also show that lexical selection benefits from correlation modeling.

In the future, we want to extend our model from the word level to the phrase level. We also plan to

improve our model with monolingual corpora.

Acknowledgments

References

Amittai Axelrod, Xiaodong He, Li Deng, Alex Acero, and Mei-Yuh Hwang. 2012. New methods and evaluation experiments on translating TED talks in the I-WSLT benchmark. In *Proc. of ICASSP 2012*, pages 4945-4648.

Rafael E. Banchs and Marta R. Costa-jussà. 2011. A Semantic Feature for Statistical Machine Translation. In *Proc. of SSSST-5 2011*, pages 126-134.

David M. Blei. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning*, pages 993-1022.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proc. of EMNLP 2007*, pages 61-72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proc. of ACL 2007*, pages 33-40.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, pages 201-228.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proc. of ACL 2011, short papers*, pages 176-181.

George Doddington. 2002. Translation Quality Using N-gram Cooccurrence Statistics. In *Proc. of HLT 2002*, 138-145.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. In *Proc. of ACL 2012, Short Papers*, pages 115-119.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proc. of ACL 2006*, pages 961-968.

Zhengxian Gong and Guodong Zhou. 2010. Improve SMT with Source-side Topic-Document Distributions. In *Proc. of SUMMIT 2010*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. In *Proc. of the National Academy of Sciences 2004*.

Xianpei Han and Le Sun. 2012. An Entity-Topic Model for Entity Linking. In *Proc. of EMNLP 2012*, pages 105-115.

Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Jesús Andrés-Ferrer 2008. Triplet Lexicon Models for Statistical Machine Translation. In *Proc. of EMNLP 2008*, pages 372-381.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic Topic Adaptation for Phrase-based MT. In *Proc. of EACL 2014*, pages 328-337.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic Topic Adaptation for SMT using Distributional Profiles. In *Proc. of WMT 2014*, pages 445-456.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving Statistical Machine Translation using Lexicalized Rule Selection. In *Proc. of COLING 2008*, pages 321-328.

Sanjika Hewavitharana, Dennis Mehay, Sankaranarayanan Ananthakrishnan, and Prem Natarajan. 2013. Incremental Topic-based TM Adaptation for Conversational SLT. In *Proc. of ACL 2013, Short Papers*, pages 697-701.

Saurabh S. Kataria, Krishnan S. Kumar, and Rajeev Rastogi. 2011. Entity Disambiguation with Hierarchical Topic Models. In *Proc. of KDD 2011*, pages 1037-1045.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proc. of NAACL-HLT 2003*, pages 127-133.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of EMNLP 2004*, pages 388-395.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. Maximum Entropy based Rule Selection Model for Syntax-based Statistical Machine Translation. In *Proc. of EMNLP 2008*, pages 89-97.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models. In *Proc. of EMNLP 2009*, pages 210-218.

Franz Joseph Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of ACL 2002*, pages 295-302.

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 2003(29), pages 19-51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL 2003*, pages 160-167.

Franz Joseph Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 2004(30), pages 417-449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2007. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL 2002*, pages 311-318.

Nick Ruiz and Marcello Federico. 2012. Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 294-302.

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. of EMNLP 2009*, pages 72-80.

Andreas Stolcke. 2002. Srilm - An Extensible Language Modeling Toolkit. In *Proc. of ICSLP 2002*, pages 901-904.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proc. of ACL 2012*, pages 459-468.

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

Ferhan Ture, DouglasW. Oard, and Philip Resnik. 2012. Encouraging Consistent Translation Choices. In *Proc. of NAACL-HLT 2012*, pages 417-426.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377-403.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. In *Proc. of MT SUMMIT 2011*, pages 131-138.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A Topic Similarity Model for Hierarchical Phrase-based Translation. In *Proc. of ACL 2012*, pages 750-758.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proc. of ACL 2006*, pages 521-528.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü, and Qun Liu. 2013. Modeling Lexical Cohesion for Document-Level Machine Translation. In *Proc. of IJCAI 2013*, pages 2183-2189.

Deyi Xiong and Min Zhang. 2014. A Sense-Based Translation Model for Statistical Machine Translation. In *Proc. of ACL 2014*, pages 1459-1469.

Bing Zhao and Eric P.Xing. 2006. BiTAM: Bilingual Topic AdMixture Models for Word Alignment. In *Proc. of ACL/COLING 2006*, pages 969-976.

Bing Zhao and Eric P.Xing. 2007. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In *Proc. of NIPS 2007*, pages 1-8.