Joint and Coupled Bilingual Topic Model Based Sentence Representations for Language Model Adaptation

Shixiang Lu, Xiaoyin Fu, Wei Wei, Xingyuan Peng, Bo Xu
Interactive Digital Media Technology Research Center (IDMTech)
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{*shixiang.lu,xiaoyin.fu,wei.wei.media,xingyuan.peng,xubo*}*@ia.ac.cn*

Abstract

This paper is concerned with data selection for adapting language model (LM) in statistical machine translation (SMT), and aims to find the LM training sentences that are topic similar to the translation task. Although the traditional approaches have gained significant performance, they ignore the topic information and the distribution information of words when selecting similar training sentences. In this paper, we present two bilingual topic model (BLTM) (joint and coupled BLTM) based sentence representations for cross-lingual data selection. We map the data selection task into cross-lingual semantic representations that are language independent, then rank and select sentences in the target language LM training corpus for a sentence in the translation task by the semantics-based likelihood. The semantic representations are learned from the parallel corpus, with the assumption that the bilingual pair shares the same or similar distribution over semantic topics. Large-scale experimental results demonstrate that our approaches significantly outperform the state-of-the-art approaches on both LM perplexity and translation performance, respectively.

1 Introduction

In recent years, LM adaptation for SMT tries to improve a generic LM by using smaller amounts of training data [Ruiz and Federico, 2011]. Selecting training data which are similar to the translation task from the large corpus has become an important approach to improve the performance of LM [Eck *et al.*, 2004; Zhao *et al.*, 2004; Masskey and Sethy, 2010; Axelrod *et al.*, 2011; Lu *et al.*, 2012b]. The bias LM, which is estimated with the similar training data, would empirically provide more accurate lexical probabilities, thus better match with the translation task [Lu *et al.*, 2012b] and promote more fluent translations.

To select similar training data for LM adaptation in SMT, many researchers proposed various approaches, such as TF-IDF [Eck *et al.*, 2004; Zhao *et al.*, 2004; Foster and Kuhn, 2007], centroid similarity [Masskey and Sethy, 2010], cross-entropy difference [Axelrod *et al.*, 2011], phrase-based similarity [Lu *et al.*, 2012a], cross-lingual similarity (CLS) [Ananthakrishnan *et al.*, 2011], and cross-lingual information retrieval [Snover *et al.*, 2008; Lu *et al.*, 2012b]. They perform at the word level, exact only term matching schemes. Unfortunately, they all do not consider the topic information and the distribution information of words in the whole LM training corpus. These information have been successfully used for LM adaptation in SMT [Tam *et al.*, 2007; Zhao and Xing, 2007; Ruiz and Federico, 2011] and proved very useful. These approaches apply the topic posterior distribution to the target language LM via marginal adaptation or minimum discrimination information (MDI) adaptation. However, they focus on modifying the LM parameter itself, which is different from data selection based LM adaptation in the category.

In this paper, we argue that it is beneficial to capture word-topic information and word-distribution information for data selection based LM adaptation. To this end, we present two more principled bilingual topic models (BLTM), joint BLTM (JBLTM) and coupled BLTM (CBLTM), and map the data selection task into cross-lingual semantic representations that are language independent. We apply these BLTM based sentence representations to cross-lingual data selection for LM adaptation in SMT. Unlike the general topic models, these two BLTMs work at the sentence level, represent a sentence as a distribution of semantic topics, and assume that a word in the sentence is generated from a mixture of these topics. They learn a semantic representation from bilingual parallel corpus which shares the same or similar topic fractions as much as possible. Compared with the traditional approaches, our approaches are potentially more effective because they consider the topic information and the distribution of words into similar data selection. It is thus reasonable to expect that using such information as ranking features is likely to further improve the quality of selected sentences and adapted LMs, as we will show in the experiments. To the best of our knowledge, this is the first extensive and empirical study of applying BLTM based sentence representations into cross-lingual data selection for LM Adaptation.

Specifically, we make following contributions:

- We present two BLTMs, JBLTM and CBLTM, and formulate the data selection task into cross-lingual semantic representations at the sentence level that are language independent (in section 4.1 and 4.2). JBLTM assigns a pair

of corresponding sentences having the same topic distributions, while CBLTM assigns the similar topic distributions.

- We apply JBLTM and CBLTM based sentence semantic representations to cross-lingual data selection (in section 4.3). Then, we rank and select the candidate sentences in the target LM training corpus for a sentence in the translation task by the semantics-based likelihood.
- To further improve the quality of selected sentences, we introduce a linear ranking model framework in which other information (e.g., word-based translation model [Lu *et al.*, 2012b]) is incorporated as features (in section 4.4).
- Finally, we conduct large-scale experiments on LM perplexity and translation performance, respectively (in section 5). The results demonstrate that our approach significantly outperforms the state-of-the-art approaches.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the framework of cross-lingual data selection for LM adaptation. Section 4 presents our BLTM based sentence representations for cross-lingual data selection. Section 5 gives large-scale experiments, and followed by the conclusions in section 6.

2 Related Work

Most previous work focus on monolingual LM adaptation in SMT based on multi-pass translation. They select the sentences which are similar to the translation hypotheses. However, the noisy translation hypotheses would mislead data selection process, and degrades the performance of adapted LM. To address this problem, Lu et al. (2012b) propose cross-lingual data selection based LM adaptation for SMT, which models data selection based on the translation task directly. Following this convincing idea, we present BLTM based cross-lingual data selection for LM adaptation.

Recently, topic models have been extended from monolingual to handle cross-lingual or multi-lingual cases, where there are pairs or tuples of corresponding documents in different languages.

Tam et al. (2007) propose a bilingual-LSA approach for LM adaptation. This model consists of two hierarchical LDA models, constructed from parallel corpora, and assumes that the topic distributions of the source and target documents are identical. A one-to-one correspondence between LDA models is enforced by learning the hyperparameters of the variational Dirichlet posteriors in one LDA model and bootstrapping the second model by fixing the hyperparameters. HM-BiTAM [Zhao and Xing, 2007] constructs a generative model in which words from a target language are sampled from a mixture of topics drawn from a Dirichlet distribution. It generates unigram LMs for both the source and target language and thus can be used for LM adaptation through MDI. Ruiz and Federico (2011) presents a bilingual latent semantic model for LM adaptation by combining text in the source and target language into very short documents and performing PLSA during model training. During inference, documents containing only the source language can be used to infer a full topic-word distribution on all words in the target language's vocabulary, from which they perform MDI adaptation on a background LM. However, the above approaches focus on modifying the LM parameter itself, which is different from data selection based LM adaptation in the category.

Except the application for LM adaptation, some cross-lingual topic models are proposed for other NLP tasks. Polylingual topic model (PLTM) [Mimno *et al.*, 2009] is an extension to LDA that views documents in a tuple as having a shared topic distribution. Joint PLSA (JPLSA) is a variant of PLTM when documents of different languages share the same word-topic distribution, and coupled PLSA (CPLSA) extends JPLSA by constraining corresponding documents to have similar fractions of words assigned to each topic according to the posterior distribution of topic assignments, instead of sharing the prior topic distributions [Platt *et al.*, 2010]. Clickthrough-based bilingual latent semantic model is proposed for web search [Gao *et al.*, 2011], and it can be viewed as a special case of PLTM, where search queries and web documents are assumed to be written in two different languages.

Our presented two BLTMs are the variants and extensions to these previous models. For the first time, we work the topic model at the sentence level, and strive to effectively learn model parameters from bilingual parallel corpus for the application of LM adaptation in SMT.

3 Cross-Lingual Data Selection Based Language Model Adaptation

Our LM adaptation is an unsupervised similar data selection guided by BLTM based bilingual sentence representations. For the source sentences in the translation task, we estimate a new bias LM, from the corresponding target LM training sentences which are selected as the similar sentences. Following [Zhao *et al.*, 2004; Snover *et al.*, 2008], the generic LM $P_g(w_i|h)$ and the bias LM $P_b(w_i|h)$ are combined as the adapted LM $P_a(w_i|h)$, as follows,

$$P_a(w_i|h) = \mu P_g(w_i|h) + (1-\mu)P_b(w_i|h) \quad (1)$$

where the interpolation factor μ is estimated using the Powell Search algorithm [Press *et al.*, 1992].

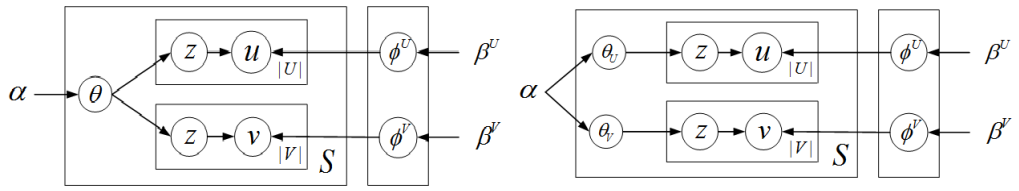
Our work focuses on the two BLTM based bilingual sentence representations and their application for cross-lingual data selection in SMT.

4 BLTM Based Sentence Representations for Cross-Lingual Data Selection

In this section, we will introduce two BLTMs (JBLTM and CBLTM) in detail. Furthermore, we will introduce JBLTM and CBLTM based sentence representations for cross-lingual data selection for LM adaptation in SMT.

4.1 JBLTM Based Sentence Representations

The first JBLTM model (Figure 1(a)), which is a close variant of JPLSA [Platt *et al.*, 2010], assumes that a pair of parallel sentences have a common topic distribution θ, and uses this assumption to optimize the probability of the data.



(a) JBLTM (b) CBLTM

Figure 1: Graphical model for bilingual topic model.

We assume that a sentence $U = u_1, \cdots, u_{|U|}$ in the translation task and its paired similar sentence $V = v_1, \cdots, v_{|V|}$ in the LM training corpus share a common topic distribution, but use different vocabularies to express these topics. Formally, inspired by [Gao *et al.*, 2011], JBLTM assumes the following process of generating a cross-lingual sentence pair.

For each topic z, a pair of different word distributions (ϕ_z^U, ϕ_z^V) are selected from a Dirichlet prior with concentration parameter β, where ϕ_z^U is a topic-specific word distribution of U, and ϕ_z^V a topic-specific word distribution of V. Assuming there are T topics, we have two sets of distributions $\phi^U = (\phi_1^U, \cdots, \phi_T^U)$ and $\phi^V = (\phi_1^V, \cdots, \phi_T^V)$. Then, for each U and its paired V, a topic distribution θ is drawn from a Dirichlet prior with concentration parameter α. Each word in U is then generated by first selecting a topic z according to θ, and drawing a word from ϕ_z^U. The paired sentence V has the corresponding derivation.

Thus, the semantics-based log-likelihood of a similar cross-lingual sentence pair, together with the paired word-topic vectors, is

$$\log\left(P(\phi^U|\beta^U)P(\phi^V|\beta^V)\prod_{(U,V)}P(\theta|\alpha)P((U,V)|\theta,\phi^U,\phi^V)\right) \tag{2}$$

where

$$P\left((U,V)|\theta,\phi^U,\phi^V\right) = \prod_{u\in U}\sum_z P(u|\phi_z^U)P(z|\theta)\cdot\prod_{v\in V}\sum_z P(v|\phi_z^V)P(z|\theta) \tag{3}$$

Recently, MAP inference has shown to perform comparably to the best inference method for LDA , if the hyper-parameters are chosen optimally for the inference method [Asuncion *et al.*, 2009]. Therefore, we use the standard EM algorithm [Dempster *et al.*, 1977] to estimate the parameter (θ, ϕ^U, ϕ^V) of JBLTM by maximizing the joint log-likelihood of the parallel corpus and the parameters for LM adaptation.

E-Step: computing the posterior probability for each word u in U and each word v in its paired sentence V for the latent variables z, as follows,

$$P(z|u,\theta) = \frac{P(u|\phi_z^U)P(z|\theta)}{\sum_{z'}P(u|\phi_{z'}^U)P(z'|\theta)} \tag{4}$$

$$P(z|v,\theta) = \frac{P(v|\phi_z^V)P(z|\theta)}{\sum_{z'}P(v|\phi_{z'}^V)P(z'|\theta)} \tag{5}$$

M-Step: updating the parameters for given posterior probability computed in the previous E-step, as follows,

$$P(u|\phi_z^U) = \frac{\beta^U - 1 + \sum_{U,V}N_{u,z}^{U,V}}{D_U\beta^U - D_U + \sum_{(U,V),u'}N_{u',z}^{U,V}} \tag{6}$$

$$P(v|\phi_z^V) = \frac{\beta^V - 1 + \sum_{U,V}N_{v,z}^{U,V}}{D_V\beta^V - D_V + \sum_{(U,V),v'}N_{v',z}^{U,V}} \tag{7}$$

$$P(z|\theta) = \frac{\alpha - 1 + (\sum_u N_{u,z}^{U,V} + \sum_v N_{v,z}^{U,V})}{T\alpha - \alpha + \sum_{z'}(\sum_u N_{u,z'}^{U,V} + \sum_v N_{v,z'}^{U,V})} \tag{8}$$

where, α, β^U and β^V are hyper-parameters, each corresponding to one Dirichlet prior. D_U and D_V are the size of the source and target vocabulary, respectively. Let $n(u,U)$ be the frequency of u in U, $n(v,V)$ be the frequency of v in V, and we have the following definitions,

$$N_{u,z}^{U,V} = n(u,U)P(z|u,\theta) \tag{9}$$

$$N_{v,z}^{U,V} = n(v,V)P(z|v,\theta) \tag{10}$$

4.2 CBLTM Based Sentence Representations

For the task of similar sentence selection based LM adaptation, we also want our BLTM to assign similar topic distributions θ to a pair of corresponding sentences, which is different from the first JBLTM model. This difference becomes especially apparent when corresponding sentences have different lengths in our task for LM adaptation. In this case, the model will tend to derive a topic vector θ which explains the longer sentence best, making the sum of the two sentences' log-likelihoods higher. Modeling the shorter sentence's best topic carries little weight.

To address the above problems, we present the second BLTM, CBLTM (Figure 1(b)). CBLTM models both sentences equally, and the topic vectors of a pair of sentences in two languages are shown completely independent, which is somewhat similar to CPLSA [Platt *et al.*, 2010]. Compared with [Gao *et al.*, 2011], this modification not only solves the above problems, but also significantly improves the selected sentences for LM adaptation.

CBTLM is also trained using a modified EM algorithm, and we use posterior regularization [Graca *et al.*, 2008; Ganchev *et al.*, 2010] to place linear constraints on the expectations of closer topic assignments to two corresponding sentences.

E-Step: computing the posterior distributions of topics from parallel corpus like Equation (4) and (5), but with some corresponding changes, as follows,

$$P(z^U|u,\theta^U) = \frac{P(u|\phi_z^U)P(z^U|\theta^U)}{\sum_{z'}P(u|\phi_{z'}^U)P(z'|\theta^U)} \tag{11}$$

$$P(z^V|v,\theta^V) = \frac{P(v|\phi_z^V)P(z^V|\theta^V)}{\sum_{z'} P(v|\phi_{z'}^V)P(z'|\theta^V)} \quad (12)$$

Then, the posterior distributions are projected onto a constrained set of distributions, for which the expected fraction of tokens in U that are assigned topic t is similar with the expected fraction of tokens in V.

M-Step: updating the expected counts with respect to this projected posterior distribution like Equation (6) to (8), but with some corresponding changes for Equation (8) to (10), as follows,

$$P(z^U|\theta^U) = \frac{\alpha - 1 + \sum_u N_{u,z}^{U,V}}{T\alpha - \alpha + \sum_{z'}\sum_u N_{u,z'}^{U,V}} \quad (13)$$

$$P(z^V|\theta^V) = \frac{\alpha - 1 + \sum_v N_{v,z}^{U,V}}{T\alpha - \alpha + \sum_{z'}\sum_v N_{v,z'}^{U,V}} \quad (14)$$

$$N_{u,z}^{U,V} = n(u,U)P(z^U|u,\theta^U) \quad (15)$$

$$N_{v,z}^{U,V} = n(v,V)P(z^U|v,\theta^V) \quad (16)$$

Next, we describe how the projection is performed. For a sentence pair (U,V), we would like CBLTM to be such that the expected fraction of tokens in U that get assigned topic t is approximately the same as the expected fraction of tokens in V that get assigned the same topic t, for each topic $t = 1 \cdots T$. This is exactly what we need to make each pair of corresponding sentences close. Let (U,V) be a pair of sequences of tokens and their topic assignments, where

$$U = \{(u_1,\cdots,u_{|U|}),(z_1^U,\cdots,z_{|U|}^U)\} \quad (17)$$

$$V = \{(v_1,\cdots,v_{|V|}),(z_1^V,\cdots,z_{|V|}^V)\} \quad (18)$$

Let Q denotes the posterior distribution set over the hidden topic assignments,

$$Q = \{P(z^U|u,\theta^U), P(z^V|v,\theta^V)\} \quad (19)$$

Q' is an ideal distribution set that has the desired property,

$$Q' = \{P'_U(z^U|u), P'_V(z^V|v)\} \quad (20)$$

such that, the expected fraction of each topic is similar or closer in U and V,

$$\left| E_{P'_U}\left[\frac{1}{|U|}\sum_{j=1}^{|U|} 1(z_j^U = t)\right] - E_{P'_V}\left[\frac{1}{|V|}\sum_{j=1}^{|V|} 1(z_j^V = t)\right]\right| \leq \epsilon_t \quad (21)$$

Then, the projection minimizes the KL divergence between two sets of distributions Q and Q', $KL(Q'||Q)$. The projection can be formulated as a constrained optimization problem, where we seek an ideal set of distributions Q' that is closest to Q,

$$\min_{Q' \in Q} KL(Q'||Q) \quad (22)$$

The valid ideal distribution space Q is non-empty, and the problem of Equation (22) can be solved efficiently in its dual form. The final corpus-wide objective is summed over sentence-pairs, and also contains terms for the probabilities of the parameters θ and ϕ given the Dirichlet priors. The norm of ϵ is minimized, which makes the expected proportions of topics in two sentences as close as possible.

We initialize the models deterministically by assigning each word to exactly one topic to begin with, such that all topics have roughly the same number of words. Words are sorted by frequency and thus words of similar frequency are more likely to be assigned to the same topic.

4.3 Ranking Candidate Sentences

In our two BLTMs, the topics are generated from V written in the target language, and the word u is generated from topic-specific word distributions in the source language. Therefore, it can be considered as performing a translation of two corresponding sentences from the translation task to the large LM training corpus via hidden topics. Then, we can rank the candidate sentences in the target LM training corpus for a sentence in the translation task by the semantics-based likelihood, and further select the top-N ranked similar sentences to adapt LM.

In our experiments for LM adaptation, we use the following sentence ranking function,

$$P(U|V) = \prod_{u \in U} P(u|V) \quad (23)$$

$$P(u|V) = \gamma P(u|C_U) + (1-\gamma)P_{BLTM}(u|V) \quad (24)$$

$$P(u|C_U) = \frac{n(u,C_U)}{|C_U|} \quad (25)$$

$$P_{BLTM}(u|V) = \begin{cases} \sum_z P(u|\phi_z^U)P(z|\theta) & \text{JBLTM} \\ \sum_z P(u|\phi_z^U)P(z|\theta^V) & \text{CBLTM} \end{cases} \quad (26)$$

where, γ is the tuning parameter. $P(u|C_U)$ is the un-smoothed background model. C_U refers to the translation task, $|C_U|$ refers to its size, and $n(u,C_U)$ refers to the frequency of u in C_U, respectively.

4.4 Linear Ranking Model

To further improve the performance of ranking the candidate sentences, we introduce a linear ranking model framework for cross-lingual data selection in which different models are incorporated as features.

We consider the linear ranking model as follows,

$$\begin{aligned} S(U,V) &= \lambda^T \cdot H(U,V) \\ &= \sum_{n=1}^{N} \lambda_n h_n(U,V) \end{aligned} \quad (27)$$

where the model has a set of N features. Each feature is an arbitrary function that maps (U,V) to a real value, and λ_n is the corresponding parameter. We optimize the parameter using the Powell Search algorithm [Press *et al.*, 1992] via cross-validation.

In our experiments, we use word translation model based cross-lingual data selection (CLWTM) [Lu *et al.*, 2012b] as the features, listed as follows,

$$P_{CLWTM}(u|V) = \sum_{v \in V} P(u|v)P(v|V) \quad (28)$$

$$P(v|V) = \frac{n(v,V)}{|V|} \quad (29)$$

where, $P(u|v)$ is the word-to-word based translation model. $P(v|V)$ is the un-smoothed sentence model, and $n(v,V)$ refer to the frequency of v in V.

5 Experiments and Results

We measure the utility of our approach in two ways: (a) comparing reference translations based perplexity of adapted LMs with the generic LM, and (b) comparing translation performance of adapted LMs with the generic LM.

5.1 Experimental Setup

We conduct experiments on the following NIST Chinese-English translation tasks. The bilingual corpus comes from LDC[1], which consists of 3.4M sentence pairs with 64M/70M Chinese/English words. The LM training corpus is the English Gigaword corpus[2], which consists of 11.3M sentences. Our first test set (test-1) is NIST 2006 MT Evaluation test set, our second test set (test-2) is NIST 2008 MT Evaluation test set, and our tuning set is NIST 2005 MT Evaluation test set.

To improve the efficiency of BLTM training, we consider the data sets in our task are constructed by the key words or important words. We adopt a variant of TextRank algorithm [Mihalcea and Tarau, 2004] for key word extraction which achieves state-of-the-art accuracy, and manually remain 75% of total words as the important words.

For the two BLTMs training, we both use 100 topics. We use $\alpha = 1.1$ and $\beta = 1.01$ in JBLTM, and $\alpha = 1.1$ and $\beta^U = \beta^V = 1.01$ in CBLTM for the values of the concentration parameters, respectively. We perform learning through MAP inference using EM (with a projection step for CBLTM). We do up to 300 iterations, and do early stopping based on task performance on the tuning set. The JBLTM model requires more iterations before reaching its peak accuracy, tending to require around 200 to 250 iterations for convergence. CBLTM requires fewer iterations, but each iteration is slower due to the projection step.

We randomly divide the tuning set into five subsets and conduct 5-fold cross-validation experiments. In each trial, we tune the parameter μ in Equation (1) and parameter λ in Equation (27) with four of five subsets, respectively. Then, we apply these parameters to one remaining subset.

For comparable experiments, we choose TF-IDF as a foundation since it has gain the state-of-the-art performance for monolingual LM adaptation [Eck *et al.*, 2004; Zhao *et al.*, 2004; Foster and Kuhn, 2007]. CLS_s [Lu *et al.*, 2012b] is the improved algorithm on CLS [Ananthakrishnan *et al.*, 2011] with optimization measure like TF-IDF. CLWTM refers to word-to-word translation model based cross-lingual data selection [Lu *et al.*, 2012b], and it is the state-of-the-art cross-lingual data selection approach for LM adaptation. JBLTM+CLWTM and CBLTM+CLWTM are our presented approaches which are incorporated with CLWTM by the linear ranking model.

5.2 Perplexity Analysis

We estimate the generic 4-gram LM with the entire LM training corpus as the baseline. Then, we select the top-N ranked sentences which are similar to the tuning set, estimate the bias 4-gram LMs with these selected sentences, and interpolate with the generic 4-gram LM as the adapted LMs. All the LMs are estimated by the SRILM toolkit [Stolcke, 2002] with interpolated modified Kneser-Ney discounting.

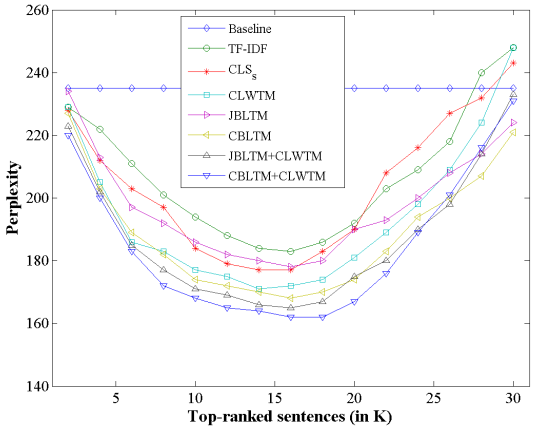


Figure 2: English reference translations based perplexity of adapted LMs vs. the size of selected training data with different approaches on the tuning set.

Figure 2 shows the English reference translations based perplexity of adapted LMs vs. the size of selected data. Obviously, proper size of similar sentences with the translation task makes the adapted LM perform well, but if too many noisy data are taken into the selected sentences, the performance significantly becomes worse. Furthermore, our approaches have two clear advantages. JBLTM and CBLTM are more stable than other approaches, and the increasing trends of perplexity after the certain size are relatively slow. JBLTM+CLWTM and CBLTM+CLWTM have more obvious perplexity reductions than other approaches.

According to the perplexity results in Figure 2, we select the top 16K sentences for test-1 and test-2 which are similar to the test set for adapting LM, respectively. From Table 1, we can see that our approaches have significant perplexity reduction on the test set compared with other approaches, and the results indicate that adapted LMs are significantly better predictors of the corresponding translation task at hand than the generic LM. Next, we use these adapted LMs for translation experiments to show the detailed performance of selected training data for LM adaptation.

5.3 Translation Experiments

We conduct translation experiments by hierarchical phrase-based (HPB) [Chiang, 2007] translation system, and evaluated the translation quality by BLEU-4 score [Papineni *et al.*, 2002]. The generic LM and adapted LMs are estimated as above in perplexity analysis experiments. We use minimum error rate training [Och, 2003] to tune the feature weights of HPB on the tuning set.

Table 2 shows the main translation results, and the improvements are statistically significant at the 95% confidence

[1]The corpus includes LDC2002E18, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006T04, and LDC2007T09.

[2]LDC2007T07

Method	Test-1		Test-2	
	Perplexity	Reduction	Perplexity	Reduction
Baseline	398.3	–	569.9	–
TF-IDF	346.2	13.08%	512.0	10.16%
CLS_s	340.9	14.41%	515.4	9.57%
CLWTM	332.7	16.47%	490.7	13.89%
JBLTM	**340.6**	**14.49%**	**494.6**	**13.21%**
CBLTM	**331.2**	**16.85%**	**492.5**	**13.57%**
JBLTM+CLWTM	**322.7**	**18.99%**	**480.0**	**15.78%**
CBLTM+CLWTM	**320.1**	**19.64%**	**471.1**	**17.34%**

Table 1: English reference translations based perplexity of adapted LMs on two test sets, with the top 16K similar sentences.

#	Method	BLEU		
		Tune	Test-1	Test-2
1	Baseline	32.45	29.15	25.69
2	TF-IDF	32.94	29.78	26.24
3	CLS_s	33.16	29.84	26.21
4	CLWTM	33.49	29.93	26.40
5	**JBLTM**	**33.41**	**29.87**	**26.38**
6	**CBLTM**	**33.54**	**30.01**	**26.47**
7	**JBLTM+CLWTM**	**33.63**	**30.13**	**26.56**
8	**CBLTM+CLWTM**	**33.72**	**30.22**	**26.64**

Table 2: The compared results of translation performance ($p < 0.05$) for LM adaptation with different approaches.

interval with respect to the baseline. From the results, we get some clear trends:

(1) CLS_s outperforms TF-IDF (row 3 vs. row2), so data selection based LM adaptation in SMT should be performed by the cross-lingual model.

(2) CBLTM outperforms JBLTM (row 6 vs. row 5). Compared to JBLTM, CBLTM assigns similar topic distributions θ to a pair of corresponding sentences and models both corresponding sentences equally when they have different lengths. Thus, CBLTM is more suitable for cross-lingual data selection based LM adaptation, and it obtains better performance.

(3) CBLTM outperforms CLWTM (row 6 vs. row 4). CLWTM is the state-of-the-art cross-lingual data selection method for LM adaptation, it learns the word-to-word translation probability from parallel corpus. However, it ignores the word-topic information and the word-distribution information for similar data selection, and these information are proved useful for LM adaptation.

(4) JBLTM+CLWTM significantly outperforms JBLTM and CLWTM (row 7 vs. row 5 and 4), CBLTM+CLWTM significantly outperforms CBLTM and CLWTM (row 8 vs. row 6 and 4), respectively. This demonstrates that the word-topic information learned from BLTM (JBLTM and CBLTM) and the word-to-word translation probability learned from CLWTM are complementary to each other for cross-lingual data selection, and the performance of adapted LMs are further improved by incorporating them together with the linear ranking model.

5.4 The Effectiveness of Data Selection

Previous work for LM adaptation can be divided into two categories: one focuses on selecting similar training data (e.g., all approaches in the above experiments) and the other focuses on modifying LM parameter itself (e.g., bilingual-LSA [Tam *et al.*, 2007]). In this section, we choose bilingual-LSA as the foundation since it has gained the state-of-the-art performance, and compare these two categories' effectiveness. Table 3 shows the effectiveness of data selection to the performance of LM adaptation. We can see that CBLTM and JBLTM significantly outperform bilingual-LSA for LM adaptation. Data selection based LM adaptation introduces similar training data, and this is more directly for the translation task at hand. We suspect this leads to the poor performance of bilingual-LSA. However, it will be instructive to explore the detail of these two categories' effectiveness in future.

Method	BLEU		
	Tune	Test-1	Test-2
Bilingual-LSA	32.96	29.67	26.13
JBLTM	33.41	29.87	26.38
CBLTM	33.54	30.01	26.47

Table 3: The effectiveness of data selection to the performance of LM adaptation.

6 Conclusions

In this paper, we present two more principled BLTMs, JBLTM and CBLTM, and apply the BLTM based sentence representations into cross-lingual data selection for LM adaptation in SMT. Unlike the general topic models, our two BLTMs work at the sentence level, and represent a sentence as a distribution of semantic topics. Compared to the traditional approaches, our approaches are potentially more effective because they consider the word-topic and word-distribution information into the similar data selection. Furthermore, we introduce a linear ranking framework which can incorporate different models for cross-lingual data selection to further improve the performance of adapted LMs. Large-scale experiments are conducted on LM perplexity and translation performance, respectively, and the results demonstrate that our approaches significantly outperform the state-of-the-art approaches for LM adaptation.

Acknowledgments

This work was supported by 863 program in China (No. 2011AA01A207). We thank Dr. Zhenbiao Chen and Dr. Hongyan Li for his helpful discussions and suggestions, and also thank the anonymous reviewers for their insightful and helpful comments.

References

[Ananthakrishnan *et al.*, 2011] Sankaranarayanan Ananthakrishnan, Rohit Prasad, and Prem Natarajan. On-line language model biasing for statistical machine translation. In *Proceedings of ACL-HLT*, pages 445–449, 2011.

[Asuncion *et al.*, 2009] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of UAI*, pages 27–34, 2009.

[Axelrod *et al.*, 2011] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*, pages 355–362, 2011.

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

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, pages 39: 1–38, 1977.

[Eck *et al.*, 2004] Matthias Eck, Stephan Vogel, and Alex Waibel. Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of LREC*, pages 327–330, 2004.

[Foster and Kuhn, 2007] George Foster and Roland Kuhn. Mixture-model adaptation for smt. In *Proceedings of ACL*, pages 128–135, 2007.

[Ganchev *et al.*, 2010] Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, pages 11: 2001–2049, 2010.

[Gao *et al.*, 2011] Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. Clickthrough-based latent semantic models for web search. In *Proceedings of SIGIR*, pages 675–684, 2011.

[Graca *et al.*, 2008] Joao Graca, Kuzman Ganchev, and Ben Taskar. Expectation maximization and posterior constraints. In *Proceedings of NIPS*, pages 569–576, 2008.

[Lu *et al.*, 2012a] Shixiang Lu, Wei Wei, Xiaoyin Fu, Lichun Fan, and Bo Xu. Phrase-based data selection for language model adaptation in spoken language translation. In *Proceedings of ISCSLP*, pages 193–196, 2012.

[Lu *et al.*, 2012b] Shixiang Lu, Wei Wei, Xiaoyin Fu, and Bo Xu. Translation model based cross-lingual language model adaptation: from word models to phrase models. In *Proceedings of EMNLP-CoNLL*, pages 512–522, 2012.

[Masskey and Sethy, 2010] Sameer Masskey and Abhinav Sethy. Resampling auxiliary data for language model adaptation in machine translation for speech. In *Proceedings of ICASSP*, pages 4817–4820, 2010.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, pages 404–411, 2004.

[Mimno *et al.*, 2009] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of EMNLP*, pages 880–889, 2009.

[Och, 2003] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, 2003.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

[Platt *et al.*, 2010] John Platt, Kristina Toutanova, and Wentau Yih. Translingual document representations from discriminative projections. In *Proceedings of EMNLP*, pages 251–261, 2010.

[Press *et al.*, 1992] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Numerical recipes in C. *Cambridge University Press*, 1992.

[Ruiz and Federico, 2011] Nick Ruiz and Marcello Federico. Topic adaptation for lecture translation through bilingual latent semantic models. In *Porceedings of WMT*, pages 294–302, 2011.

[Snover *et al.*, 2008] Matthew Snover, Bonnie Dorr, and Richard Schwartz. Language and translation model adaptation using comparable corpora. In *Proceedings of EMNLP*, pages 857–866, 2008.

[Stolcke, 2002] Andreas Stolcke. SRILM - An extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 901–904, 2002.

[Tam *et al.*, 2007] Yik-Cheung Tam, Ian Lane, and Tanja Schultz. Bilingual-LSA based LM adaptation for spoken language translation. In *Proceedings of ACL*, pages 520–527, 2007.

[Zhao and Xing, 2007] Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Proceedings of NIPS*, pages 1689–1696, 2007.

[Zhao *et al.*, 2004] Bing Zhao, Matthias Eck, and Stephan Vogel. Language model adaptation for statistical machine translation via structured query models. In *Proceedings of COLING*, pages 411–417, 2004.