Translation Model Based Cross-Lingual Language Model Adaptation: from Word Models to Phrase Models

Shixiang Lu, Wei Wei, Xiaoyin Fu, and Bo Xu
Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Haidian District, Beijing 100190, China
{shixiang.lu,wei.wei.media,xiaoyin.fu,xubo}@ia.ac.cn

Abstract

In this paper, we propose a novel translation model (TM) based cross-lingual data selection model for language model (LM) adaptation in statistical machine translation (SMT), from word models to phrase models. Given a source sentence in the translation task, this model directly estimates the probability that a sentence in the target LM training corpus is similar. Compared with the traditional approaches which utilize the first pass translation hypotheses, cross-lingual data selection model avoids the problem of noisy proliferation. Furthermore, phrase TM based cross-lingual data selection model is more effective than the traditional approaches based on bag-of-words models and word-based TM, because it captures contextual information in modeling the selection of phrase as a whole. Experiments conducted on large-scale data sets demonstrate that our approach significantly outperforms the state-of-the-art approaches on both LM perplexity and SMT performance.

1 Introduction

Language model (LM) plays a critical role in statistical machine translation (SMT). It seems to be a universal truth that LM performance can always be improved by using more training data (Brants et al., 2007), but only if the training data is reasonably well-matched with the desired output (Moore and Lewis, 2010). It is also obvious that among the large training data the topics or domains of discussion will change (Eck et al., 2004), which causes the mismatch problems with the translation task. For this reason, most researchers preferred to select similar training data from the large training corpus in the past few years (Eck et al., 2004; Zhao et al., 2004; Kim, 2005; Masskey and Sethy, 2010; Axelrod et al., 2011). This would empirically provide more accurate lexical probabilities, and thus better match the translation task at hand (Axelrod et al., 2011).

Many previous data selection approaches for LM adaptation in SMT depend on the first pass translation hypotheses (Eck et al., 2004; Zhao et al., 2004; Kim, 2005; Masskey and Sethy, 2010), they select the sentences which are similar to the translation hypotheses. These schemes are overall limited by the quality of the translation hypotheses (Tam et al., 2007 and 2008), and better initial translation hypotheses lead to better selected sentences (Zhao et al., 2004). However, while SMT has achieved a great deal of development in recent years, the translation hypotheses are still far from perfect (Wei and Pal, 2010), which have many noisy data. The noisy translation hypotheses mislead data selection process (Xu et al., 2001; Tam et al., 2006 and 2007; Wei and Pal, 2010), and thus take noisy data into the selected training data, which causes *noisy proliferation* and degrades the performance of adapted LM.

Furthermore, traditional approaches for LM adaptation are based on bag-of-words models and considered to be *context independent*, despite of their state-of-the-art performance, such as TF-IDF (Eck et al., 2004; Zhao et al., 2004; Hildebrand et al., 2005; Kim, 2005; Foster and Kuhn, 2007), centroid similarity (Masskey and Sethy, 2010), and cross-lingual similarity (CLS) (Ananthakrishnan et al., 2011a). They all perform at the word level, exact only ter-

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 512–522, Jeju Island, Korea, 12–14 July 2012.

m matching schemes, and do not take into account any contextual information when modeling the selection by single words in isolation, which degrade the quality of selected sentences.

In this paper, we argue that it is beneficial to model the data selection based on the source translation task directly and capture the contextual information for LM adaptation. To this end, we propose a more principled translation model (TM) based cross-lingual data selection model for LM adaptation, from word models to phrase models. We assume that the data selection should be performed by the cross-lingual model and at the phrase level. Given a source sentence in the translation task, this model directly estimates the probability before translation that a sentence in the target LM training corpus is similar. Therefore, it does not require the translation task to be pre-translation as in monolingual adaptation, and can address the problem of noisy proliferation.

To the best of our knowledge, this is the first extensive and empirical study of using phrase TM based cross-lingual data selection for LM adaptation. This model learns the transform probability of a multi-term phrase in a source sentence given a phrase in the target sentence of LM training corpus. Compared with bag-of-words models and word-based TM that account for selecting single words in isolation, this model performs at the phrase level and captures some *contextual information* in modeling the selection of phrase as a whole, thus it is potentially more effective. More precise data selection can be determined for phrases than for words. In this model, we propose a linear ranking model framework to further improve the performance, referred to the linear discriminant function (Duda et al., 2001; Collins, 2002; Gao et al., 2005) in pattern classification and information retrieval (IR), where different models are incorporated as features, as we will show in our experiments.

Unlike the general TM in SMT, we explore the use of TextRank algorithm (Mihalcea et al., 2004) to identify and eliminate unimportant words (e.g., non-topical words, common words) for corpus preprocessing, and construct TM by *important words*. This reduces the average number of words in cross-lingual data selection model, thus improving the efficiency. Moreover, TextRank utilizes the context information of words to assign term weights (Lee et al., 2008), which makes phrase TM based cross-lingual data selection model play its advantage of capturing the contextual information, thus further improving the performance.

The remainder of this paper is organized as follows. Section 2 introduces the related work of LM adaptation. Section 3 presents the framework of cross-lingual data selection for LM adaptation. Section 4 describes our proposed TM based cross-lingual data selection model: from word models to phrase models. In section 5 we present large-scale experiments and analyses, and followed by conclusions and future work in section 6.

2 Related Work

TF-IDF and cosine similarity have been widely used for LM adaptation (Eck et al., 2004; Zhao et al., 2004; Hildebrand et al., 2005; Kim, 2005; Foster and Kuhn, 2007). Masskey and Sethy (2010) selected the auxiliary data by computing centroid similarity score to the centroid of the in-domain data. The main idea of these methods is to select the sentences which are similar to the first pass translation hypotheses or in-domain corpus from the large LM training corpus, and estimate the bias LM for SMT system to improve the translation quality.

Tam et al. (2007 and 2008) proposed a bilingual-LSA model for LM adaptation. They integrated the LSA marginal into the target generic LM using marginal adaptation which minimizes the Kullback-Leibler divergence between the adapted LM and the generic LM. Ananthakrishnan et al. (2011a) proposed CLS to bias the count and probability of corresponding n-gram through weighting the LM training corpus. However, these two cross-lingual approaches focus on modify LM itself, which are different from data selection method for LM adaptation. In our comparable experiments, we apply CLS for the first time to the task of cross-lingual data selection for LM adaptation. Due to lack of smoothing measure for sparse vector representation in CLS, the similarity computation is not accurate which degrades the performance of adapted LM. To avoid this, we add smoothing measure like TF-IDF, called CLS_s, as we will discuss in the experiments.

Snover et al. (2008) used a word TM based CLIR

system (Xu et al., 2001) to select a subset of target documents comparable to the source document for adapting LM. Because of the data sparseness in the document state and it operated at the document level, this model selected large quantities of irrelevant text, which may degrade the adapted LM (Eck et al., 2004; Ananthakrishnan et al., 2011b). In our word TM based cross-lingual data selection model, we operate at the sentence level and add the smoothing mechanism by integrating with the background word frequency model, and these can significantly improve the performance. Axelrod et al. (2011) proposed a bilingual cross-entropy difference to select data from parallel corpus for domain adaptation which captures the contextual information slightly, and outperformed monolingual cross-entropy difference (Moore and Lewis, 2010), which first shows the advantage of bilingual data selection. However, its performance depends on the parallel in-domain corpus which is usually hard to find, and its application is assumed to be limited.

3 Cross-Lingual Data Selection for Language Model Adaptation

Our LM adaptation is an unsupervised similar training data selection guided by TM based cross-lingual data selection model. For the source sentences in the translation task, we estimate a new LM, the bias LM, from the corresponding target LM training sentences which are selected as the similar sentences. Since the size of the selected sentences is small, the corresponding bias LM is specific and more effective, giving high probabilities to those phrases that occur in the desired output translations.

Following the work of (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 using linear interpolation as the adapted LM $P_a(w_i|h)$, which is shown to improve the performance over individual model,

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

where the interpolation factor μ can be simply estimated using the Powell Search algorithm (Press et al., 1992) via cross-validation.

Our work focuses on TM based cross-lingual data selection model, from word model to phrase models, and the quality of this model is crucial to the performance of adapted LM.

4 Translation Model for Cross-Lingual Data Selection (CLTM)

Let $Q = \mathbf{q}_1, \ldots, \mathbf{q}_j$ be a source sentence in the translation task and $S = \mathbf{w}_1, \ldots, \mathbf{w}_i$ be a sentence in the general target LM training corpus, thus cross-lingual data selection model can be framed probabilistically as maximizing the $P(S|Q)$. By Bayes' rule,

$$P(S|Q) = \frac{P(S)P(Q|S)}{P(Q)} \quad (2)$$

where the prior probability $P(S)$ can be viewed as uniform, and the $P(Q)$ is constant across all sentences. Therefore, selecting a sentence to maximize $P(S|Q)$ is equivalent to selecting a sentence that maximizes $P(Q|S)$.

4.1 Word-Based Translation Model for Cross-Lingual Data Selection (CLWTM)

4.1.1 Cross-Lingual Sentence Selection Model

Following the work of (Xu et al., 2001; Snover et al., 2008), CLWTM can be described as

$$P(Q|S) = \prod_{q \in Q} P(q|S) \quad (3)$$

$$P(q|S) = \alpha P(q|C_q) + (1-\alpha) \sum_{w \in S} P(q|w)P(w|S) \quad (4)$$

where α is the interpolation weight empirically set as a constant[1], $P(q|w)$ is the word-based TM which is estimated by IBM Model 1 (Brown et al., 1993) from the parallel corpus, $P(q|C_q)$ and $P(w|S)$ are the un-smoothed background and sentence model, respectively, estimated using maximum likelihood estimation (MLE) as

$$P(q|C_q) = \frac{freq(q, C_q)}{|C_q|} \quad (5)$$

$$P(w|S) = \frac{freq(w, S)}{|S|} \quad (6)$$

where C_q refers to the translation task, $freq(q, C_q)$ refers to the number of times q occurs in C_q, $freq(w, S)$ refers to the number of times w occurs in S, and $|C_q|$ and $|S|$ are the sizes of the translation task and the current target sentence, respectively.

[1] As in Xu et al. (2001), a value of 0.3 was used for α.

4.1.2 Ranking Candidate Sentences

Because of the data sparseness in the sentence state which degrades the model, Equation (6) does not perform well in our data selection experiments. Inspired by the work of (Berger et al., 1999) in IR, we make the following smoothing mechanism:

$$P(q|S) = \alpha P(q|C_q) + (1-\alpha) \sum_{w \in S} P(q|w) P_s(w|S) \quad (7)$$

$$P_s(w|S) = \beta P(w|C_s) + (1-\beta) P(w|S) \quad (8)$$

$$P(w|C_s) = \frac{freq(w, C_s)}{|C_s|} \quad (9)$$

where $P(w|C_s)$ is the un-smoothed background model, estimated using MLE as Equation (5), C_s refers to the LM training corpus and $|C_s|$ refers to its size. Here, β is interpolation weight; notice that letting $\beta = 0$ in Equation (8) reduces the model to the un-smoothed model in Equation (4).

4.2 Phrase-Based Translation Model for Cross-Lingual Data Selection (CLPTM)

4.2.1 Cross-Lingual Sentence Selection Model

The phrase-based TM (Koehn et al., 2003; Och and Ney, 2004) has shown superior performance compared to the word-based TM. In this paper, the goal of phrase-based TM is to transfer S into Q. Rather than transferring single words in isolation, the phrase model transfers one sequence of words into another sequence of words, thus incorporating contextual information. Inspired by the work of web search (Gao et al., 2010) and question retrieval in community question answer (Q&A) (Zhou et al., 2011), we assume the following generative process: first the sentence S is broken into K non-empty word sequences $\mathbf{w}_1, \ldots, \mathbf{w}_k$, then each is transferred into a new non-empty word sequences $\mathbf{q}_1, \ldots, \mathbf{q}_k$, and finally these phrases are permutated and concatenated to form the sentence Q, where q and w denote the phrases or consecutive sequence of words.

To formulate this generative process, let U denote the segmentation of S into K phrases $\mathbf{w}_1, \ldots, \mathbf{w}_k$, and let V denote the K phrases $\mathbf{q}_1, \ldots, \mathbf{q}_k$, we refer to these $(\mathbf{w}_i, \mathbf{q}_i)$ pairs as bi-phrases. Finally, let M denote a permutation of K elements representing the final ranking step.

Next we place a probability distribution over rewrite pairs. Let $B(S, Q)$ denote the set of U, V, M triples that transfer S into Q. Here we assume a uniform probability over segmentations, so the phrase-based selection probability can be formulated as

$$P(Q|S) \propto \sum_{\substack{(U,V,M) \in \\ B(S,Q)}} P(V|S,U) \cdot P(M|S,U,V) \quad (10)$$

Then, we use the maximum approximation to the sum:

$$P(Q|S) \approx \max_{\substack{(U,V,M) \in \\ B(S,Q)}} P(V|S,U) \cdot P(M|S,U,V) \quad (11)$$

Although we have defined a generative model for transferring S into Q, our goal is to calculate the ranking score function over existing Q and S. However, this model can not be used directly for sentence ranking because Q and S are often of different lengths, the length of S is almost 1.5 times to that of Q in our corpus, leaving many words in S unaligned to any word in Q. This is another key difference between our task and SMT. As pointed out by the previous work (Berger and Lafferty, 1999; Gao et al., 2010; Zhou et al., 2011), sentence-query selection requires a distillation of the sentence, while selection of natural language tolerates little being thrown away. Thus we restrict our attention to those *key sentence words* that form the distillation of S, do not consider the unaligned words in S, and assume that Q is transfered only from the key sentence words.

In this paper, the key sentence words are identified via word alignment. Let $A = a_1 \ldots a_J$ be the "hidden" word alignment, which describes a mapping from a term position j in Q to a word position a_j in S. We assume that the positions of the key sentence words are determined by the Viterbi alignment $\hat{A}$, which can be obtained using IBM Model 1 (Brown et al., 1993) as follows:

$$\begin{aligned} \hat{A} &= \arg\max_A P(Q, A|S) \\ &= \arg\max_A \Big\{ P(J|I) \prod_{j=1}^{J} P(q_j|w_{a_j}) \Big\} \\ &= \Big[\arg\max_{a_j} P(q_j|w_{a_j}) \Big]_{j=1}^{J} \end{aligned} \quad (12)$$

Given $\hat{A}$, when scoring a given Q/S pair, we restrict our attention to those U, V, M triples that are consistent with $\hat{A}$, which we denote as $B(S, Q, \hat{A})$. Here, consistency requires that if two words are aligned in $\hat{A}$, then they must appear in the same bi-phrase $(\mathbf{w}_i, \mathbf{q}_i)$. Once the word alignment is fixed, the final permutation is uniquely determined, so we can safely discard that factor. Then Equation (11) can be written as

$$P(Q|S) \approx \max_{\substack{(U,V,M)\in \\ B(S,Q,\hat{A})}} P(V|S,U) \quad (13)$$

For the sole remaining factor $P(V|S,U)$, we assume that a segmented queried question $V = \mathbf{q}_1, \ldots, \mathbf{q}_k$ is generated from left to right by transferring each phrase $\mathbf{w}_1, \ldots, \mathbf{w}_k$ independently, as follows:

$$P(V|S,U) = \prod_{k=1}^{K} P(\mathbf{q}_k|\mathbf{w}_k) \quad (14)$$

where $P(\mathbf{q}_k|\mathbf{w}_k)$ is a phrase translation probability computed from the parallel corpus, which can be estimated in two ways (Koehn et al., 2003; Och and Ney, 2004): relative frequency and lexical weighting, and has two format: phrase translation probability and lexical weight probability.

In order to find the maximum probability assignment $P(Q|S)$ efficiently, we use a dynamic programming approach, somewhat similar to the monotone decoding algorithm described in the work (Och, 2002). We consider quantity a_j as the maximal probability of the most likely sequence of phrases in S covering the first j words in Q, therefore the probability can be calculated using the following recursion:

step (1). Initialization:

$$\alpha_0 = 1 \quad (15)$$

step (2). Induction:

$$\alpha_j = \sum_{j'<j, \mathbf{q}=q_{j'+1}\cdots q_j} \left\{ \alpha_{j'} P(\mathbf{q}|\mathbf{w_q}) \right\} \quad (16)$$

step (3). Total:

$$P(Q|S) = \alpha_J \quad (17)$$

4.2.2 Ranking Candidate Sentences

However, directly using the phrase-based TM, computed in Equations (15) to (17), to rank the candidate sentences does not perform well. Inspired by the linear discriminant function (Duda et al., 2001; Collins, 2002; Gao et al., 2005) in pattern classification and IR, we therefore propose a linear ranking model framework for cross-lingual data selection model in which different models are incorporated as features.

We consider the linear ranking model as follows:

$$\begin{aligned} Score(Q,S) &= \lambda^T \cdot H(Q,S) \\ &= \sum_{n=1}^{N} \lambda_n h_n(Q,S) \end{aligned} \quad (18)$$

where the model has a set of N features, and each feature is an arbitrary function that maps $(Q|S)$ to a real value, i.e., $H(Q,S) \in \mathbf{R}$. λ_n for $n = 1 \ldots N$ is the corresponding parameters of each feature, and we optimize these parameters using the Powell Search algorithm (Press et al., 1992) via cross-validation.

The used features in the linear ranking model are as follows:

- Phrase translation feature (PT): $h_{PT}(Q,S,A) = \log P(Q|S)$, where $P(Q|S)$ is computed using Equations (15) to (17), and $P(\mathbf{q}_k|\mathbf{w}_k)$ is phrase translation probability.
- Inverted phrase translation feature (IPT): $h_{IPT}(S,Q,A) = \log P(S|Q)$, where $P(S|Q)$ is computed using Equations (15) to (17), and $P(\mathbf{w}_k|\mathbf{q}_k)$ is inverted phrase translation probability.
- Lexical weight feature (LW): $h_{LW}(Q,S,A) = \log P(Q|S)$, where $P(Q|S)$ is computed using Equations (15) to (17), and $P(\mathbf{q}_k|\mathbf{w}_k)$ is lexical weight probability.
- Inverted lexical weight feature (ILW): $h_{ILW}(S,Q,A) = \log P(S|Q)$, where $P(S|Q)$ is computed using Equations (15) to (17), and $P(\mathbf{w}_k|\mathbf{q}_k)$ is inverted lexical weight probability.
- Unaligned word penalty feature (UWP): $h_{UWP}(Q,S,A)$, which is defined as the ratio between the number of unaligned terms and the total number of terms in Q.

- Word-based translation feature (WT): $h_{WT}(Q,S,A) = \log P(Q|S)$, where $P(Q|S)$ is the word-based TM defined by Equations (3) and (7).

4.3 Eliminating Unimportant Words (EUW)

To improve the efficiency of cross-lingual data selection process, we consider the translation task, the LM training corpus and the parallel corpus in our task are constructed by the key words or important words, and thus construct TM by the key words or important words, which is another key difference between our task and SMT. We identify and eliminate unimportant words, somewhat similar to Q&A retrieval (Lee et al., 2008; Zhou et al., 2011). Thus, the average number of words (the total word number in Q and S) in cross-lingual sentence selection model would be minimized naturally, and the efficiency of cross-lingual data selection would be improved.

In this paper, we adopt a variant of TextRank algorithm (Mihalcea and Tarau, 2004), a graph-based ranking model for key word extraction which achieves state-of-the-art accuracy. It identifies and eliminates unimportant words from the corpus, and assumes that a word is unimportant if it holds a relatively low significance in the corpus. Compared with the traditional approaches, such as TF-IDF, TextRank utilizes the context information of words to assign term weights (Lee et al., 2008), so it further improves the performance of CLPTM, as we will show in the experiments.

Following the work of (Lee et al., 2008), the ranking algorithm proceeds as follows. First, all the words in a given document are added as vertices in a graph. Then edges are added between words (vertices) if the words co-occur in a fixed-sized window. The number of co-occurrences becomes the weight of an edge. When the graph is constructed, the score of each vertex is initialized as 1, and the PageRank based ranking algorithm is run on the graph iteratively until convergence. The TextRank score $R^k_{w_i,D}$ of a word w_i in document D at kth iteration is defined as follows:

$$R^k_{w_i,D} = (1-d) + d \cdot \sum_{\forall j:(i,j)\in G} \frac{e_{i,j}}{\sum_{\forall l:(j,l)\in G} e_{j,l}} R^{k-1}_{w_j,D} \quad (19)$$

where d is a damping factor usually set as a constant[2], and $e_{i,j}$ is an edge weight between w_i and w_j.

In our experiments, we manually set the proportion to be removed as 25%, that is to say, 75% of total words in the documents would be remained as the important words.

5 Experiments

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

5.1 Corpus and Tasks

We conduct experiments on two Chinese-to-English translation tasks: IWSLT-07 (dialogue domain) and NIST-06 (news domain).

IWSLT-07. The bilingual training corpus comes from BTEC[3] and CJK[4] corpus, which contains 3.82K sentence pairs with 3.0M/3.1M Chinese/English words. The LM training corpus is from the English side of the parallel data (BTEC, CJK, and CWMT2008[5]), which consists of 1.34M sentences and 15.2M English words. The test set is IWSLT-07 test set which consists of 489 sentences, and the development set is IWSLT-05 test set which consists of 506 sentences.

NIST-06. The bilingual training corpus comes from Linguistic Data Consortium (LDC)[6], which consists of 3.4M sentence pairs with 64M/70M Chinese/English words. The LM training corpus is from the English side of the parallel data as well as the English Gigaword corpus[7], which consists of 11.3M sentences. The test set is 2006 NIST MT Evaluation test set which consists of 1664 sentences, and the development set is 2005 NIST MT Evaluation test set which consists of 1084 sentences.

[2] As in Lee et al. (2008), a value of 0.85 was used for d.

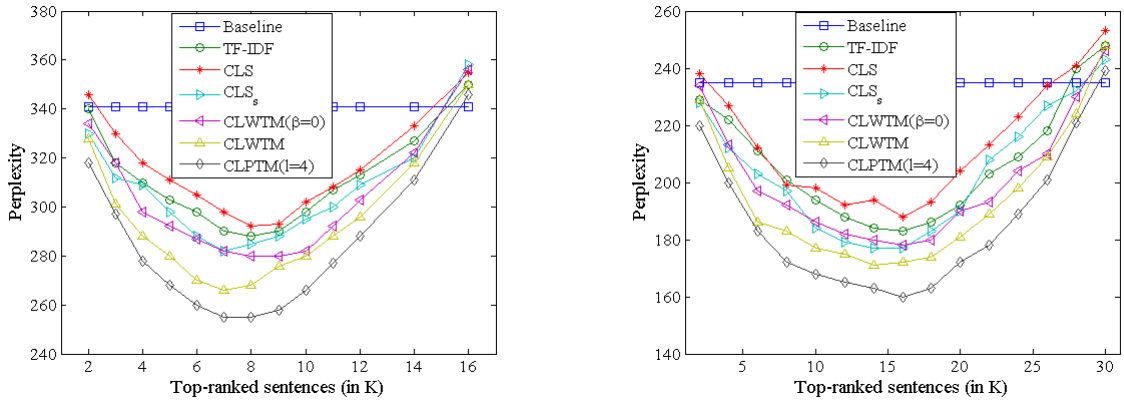
[3] Basic Traveling Expression Corpus

[4] China-Japan-Korea

[5] The 4th China Workshop on Machine Translation

[6] LDC2002E18, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006T04, LDC2007T09

[7] LDC2007T07



(a)IWSLT-07 (b)NIST-06

Figure 1: English reference translations based perplexity of adapted LMs vs. the size of selected training data with different approaches on two development sets.

5.2 Perplexity Analysis

We randomly divide the development 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 (18) with four of five subsets and then apply it to one remaining subset. The experiments reported below are those averaged over the five trials.

We estimate the generic 4-gram LM with the entire LM training corpus as the baseline. Then, we select the top-N sentences which are similar to the development set, estimate the bias 4-gram LMs (with n-gram cutoffs tuned as above) 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). Perplexity is a metric of LM performance, and the lower perplexity value indicates the better performance. Therefore, we estimate the perplexity of adapted LMs according to English reference translations.

Figure 1 shows the perplexity of adapted LMs vs. the size of selected data. In this paper, we choose TF-IDF as the foundation of our solution since TF-IDF has gained the state-of-the-art performance for LM adaptation (Eck et al., 2004; Hildebrand et al., 2005; Kim, 2005; Foster and Kuhn, 2007). CLS refers to the cross-lingual similarity of (Ananthakrishnan et al., 2011a), and CLS$_s$ is our proposed improved algorithm on CLS with optimization measure like TF-IDF. CLWTM($\beta = 0$) refers to Snover et al. (2008), which is the un-smooth version of our proposed CLWTM in the document state. CLPTM($l = 4$) is our proposed CLPTM with a maximum phrase length of four, and we score the target sentences by the highest scoring Q/S pair.

Task	Method	Perplexity	Reduction
IWSLT-07	Baseline	524.1	–
	TF-IDF	471.4	10.06%
	CLS	475.7	9.23%
	CLS$_s$	468.9	10.53%
	CLWTM($\beta = 0$)	463.5	11.56%
	CLWTM	**451.5**	**13.85%**
	CLPTM($l = 4$)	**435.3**	**16.94%**
NIST-06	Baseline	398.3	–
	TF-IDF	346.2	13.08%
	CLS	351.6	11.72%
	CLS$_s$	340.9	14.41%
	CLWTM($\beta = 0$)	341.1	14.36%
	CLWTM	**332.7**	**16.47%**
	CLPTM($l = 4$)	**319.2**	**19.86%**

Table 1: English reference translations based perplexity of adapted LMs with different approaches on two test sets, with the top 8K sentences on IWSLT-07 and top 16K sentences on NIST-06, respectively.

The results in Figure 1 indicate that English reference translations based perplexity of adapted LMs decreases consistently with increase of the size of selected top-N sentences, and increases consistently after a certain size in all approaches. Therefore, 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 becomes worse. Similar observations have been done by (Eck et al., 2004;

Task	#	Method	BLEU
IWSLT-07	1	Baseline	33.60
	2	TF-IDF	34.14
	3	CLS	34.08
	4	CLS$_s$	34.18
	5	CLWTM($\beta = 0$)	34.22
	6	**CLWTM**	**34.30**
	7	**CLPTM**($l = 4$)	**34.49**
NIST-06	8	Baseline	29.15
	9	TF-IDF	29.78
	10	CLS	29.73
	11	CLS$_s$	29.84
	12	CLWTM($\beta = 0$)	29.87
	13	**CLWTM**	**29.93**
	14	**CLPTM**($l = 4$)	**30.17**

Table 2: Comparison of SMT performance ($p < 0.05$) with different approaches for LM adaptation on two test sets.

Axelrod et al., 2011). Furthermore, it is comforting that our approaches (CLWTM and CLPTM($l = 4$)) performs better and are more stable than other approaches.

According to the perplexity results in Figure 1, we select the top 8K sentences on IWSLT-07 and top 16K sentences on NIST-06 which are similar to the test set for adapting LM, respectively. Table 1 shows English reference translations based perplexity of adapted LMs on two test sets. Our approaches have significantly reduction in perplexity 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. We use these adapted LMs for next translation experiments to show the detailed performance of selected training data for LM adaptation.

5.3 Translation Experiments

We carry out translation experiments on the test set by hierarchical phrase-based (HPB) SMT (Chiang, 2005 and 2007) system to demonstrate the utility of LM adaptation on improving SMT performance by BLEU 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 for maximum BLEU score on the development set with serval groups of different start weights.

Table 2 shows the main translation results on two test sets, and the improvements are statistically significant at the 95% confidence interval with respect to the baseline. From the comparison results, we get some clear trends:

(1) Cross-lingual data selection model outperforms the traditional approaches which utilize the first pass translation hypotheses (row 4 vs. row2; row 11 vs. row 9), but the detailed impact of noisy data in the translation hypotheses on data selection will be shown in the next section (section 5.4).

(2) CLWTM significantly outperforms CLS$_s$ (row 6 vs. row 4; row 13 vs. row 11), we suspect that word-based TM makes more accurate cross-lingual data selection model than single cross-lingual projection (Ananthakrishnan et al., 2011a).

(3) Compared with (Snover et al., 2008), adding the smoothing mechanism in the sentence state for CLWTM significantly improves the performance (row 6 vs. row 5; row 13 vs. row 12).

(4) Phrase-based TM (CLPTM) significantly outperforms the state-of-the-art approaches based on bag-of-words models and word-based TM (row 7 vs. row 2, row 4, row 5 and row 6; row 14 vs. row 9, row 11, row 12 and row 13).

Task	Translation Hypotheses	BLEU
IWSLT-07	First Pass	34.14
	Second Pass	34.31
NIST-06	First Pass	29.78
	Second Pass	29.91

Table 3: The impact of noisy data in the translation hypotheses on the performance of LM adaptation.

5.4 Impact of Noisy Data in the Translation Hypotheses

The experiment results in Table 2 indicate the second pass translation hypotheses (row 2 and row 9) made by TF-IDF are better than the first pass translation hypotheses (row 1 and row 8), so we consider that these translations have less noisy data. Thus, they were considered as the new translation hypotheses (the second pass) to select the similar sentences for LM adaptation by TF-IDF.

Table 3 shows the impact of noisy data in the translation hypotheses on the performance of adapted LMs. The observed improvement suggests that better initial translations which have less noisy data

Task	Phrase Length	BLEU
IWSLT-07	$l = 1$	34.33
	$l = 2$	34.44
	$l = 3$	34.49
	$l = 4$	34.49
NIST-06	$l = 1$	29.97
	$l = 2$	30.07
	$l = 3$	30.14
	$l = 4$	30.17

Table 4: The impact of phrase length in CLPTM on the performance of LM adaptation, and the maximum phrase length is four.

lead to better adapted LMs, and thereby better second iteration translations. Therefore, it is advisable to use cross-lingual data selection for LM adaptation in SMT, which can address the problem of noisy proliferation.

5.5 Impact of Phrase Length

The results in Table 4 show that longer phrases do yield some visible improvement up to the maximum length of four. This may suggest that some properties captured by longer phrases are also captured by other features. The performances when the phrase length is 1 are better than that of single word-based TM (row 6 and row 13 in Table 2), this suspect that the features in our linear ranking model are useful. However, it will be instructive to explore the methods of preserving the improvement generated by longer phrase when more features are incorporated in the future work.

5.6 Impact of Eliminating Unimportant Words

Table 5 shows the results of EUW by TextRank algorithm on the performance of CLTM for LM adaptation. Initial represents that we do not eliminate unimportant words. Average number represents the average number of words (the total word number in Q and S) in cross-lingual data selection model. The average number is reduced when unimportant words are eliminated, from 19 to 12 on IWSLT-07 and from 37 to 24 on NIST-06, respectively. This makes the cross-lingual data selection process become more efficient. In CLWTM, the performance with EUW is basically the same with that of the initial state; but in CLPTM, EUW outperforms the initial state because TextRank algorithm utilizes the context information of words when assigning term weights, thus makeing CLPTM play its advantage of capturing the contextual information.

Task	Method	Average Number	BLEU	
			CLWTM	CLPTM ($l = 4$)
IWSLT-07	Initial	19	34.31	34.47
	EUW	12	34.30	34.49
NIST-06	Initial	37	29.91	30.12
	EUW	24	29.93	30.17

Table 5: The impact of eliminating unimportant words by TextRank algorithm on the performance of CLTM for LM adaptation.

6 Conclusions and Future Work

In this paper, we propose a novel TM based cross-lingual data selection model for LM adaptation in SMT, from word models to phrase models, and aims to find the LM training corpus which are similar to the translation task at hand. Unlike the general TM in SMT, we explore the use of TextRank algorithm to identify and eliminate unimportant words for corpus preprocessing, and construct TM by important words. Compared with the traditional approaches which utilize the first pass translation hypotheses, cross-lingual data selection avoids the problem of noisy proliferation. Furthermore, phrase TM based cross-lingual data selection is more effective than the traditional approaches based on bag-of-words models and word-based TM, because it captures contextual information in modeling the selection of phrase as a whole. Large-scale experiments are conducted on LM perplexity and SMT performance, and the results demonstrate that our approach solves the two aforementioned disadvantages and significantly outperforms the state-of-the-art methods for LM adaptation.

There are some ways in which this research could be continued in the future. First, we will utilize our approach to mine large-scale corpora by distributed infrastructure system, and investigate the use of our approach for other domains, such as speech translation system. Second, the significant improvement of LM adaptation based on cross-lingual data selection is exciting, so it will be instructive to explore other knowledge based cross-lingual data selection for LM adaptation, such as latent semantic model.

Acknowledgments

This work was supported by 863 program in China (No. 2011AA01A207). We thank Guangyou Zhou for his helpful discussions and suggestions. We also thank the anonymous reviewers for their insightful comments.

References

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

Sankaranarayanan Ananthakrishnan, Stavros Tsakalidis, Rohit Prasad, and Prem Natarajan. 2011b. On-line language model biasing for multi-pass automatic speech recognition. In *Proceedings of INTERSPEECH*, pages 621-624.

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

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of SIGIR*, pages 222-229.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of EMNLP*, pages 858-867.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pages 263-270.

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

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with the perceptron algorithm. In *Proceedings of EMNLP*, pages 1-8.

Richard O. Duda, Peter E. Hart, and David G. Stork. 2001. Pattern classification. John Wiley & Sons, Inc.

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

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

Jianfeng Gao, Haoliang Qi, Xinsong Xia, and Jian-Yun Nie. 2005. Linear discriminative model for information retrieval. In *Proceedings of SIGIR*, pages 290-297.

Jianfeng Gao, Xiaodong He, and Jian-Yun Nie. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In *Proceedings of CIKM*, pages 1139-1148.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based information retrieval. In *Proceedings of EAMT*, pages 133-142.

Woosung Kim. 2005. Language model adaptation for automatic speech recognition and statistical machine translation. *Ph.D. thesis*, The Johns Hopkins University.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48-54.

Jung-Tae Lee, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. 2008. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. In *Proceedings of EMNLP*, pages 410-418.

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

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of EMNLP*, pages 404-411.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of ACL*, pages 220-224.

Franz Josef Och. 2002. Statistical mahcine translation: from single word models to alignment templates. *Ph.D thesis*, RWTH Aachen.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311-318.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C*. Cambridge University Press.

Matthew Snover, Bonnie Dorr, and Richard Marcu. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of EMNLP*, pages 857-866.

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

Yik-Cheung Tam and Tanja Schultz. 2006. Unsupervised language model adaptation using latent semantic marginals. In *Proceedings of ICSLP*, pages 2206-2209.

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

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2008. Bilingual-LSA based adaptation for statistical machine translation. *Machine Translation*, 21:187-207.

Bin Wei and Christopher Pal. 2010. Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of ACL*, pages 258-262.

Jinxi Xu, Ralpha Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of SIGIR*, pages 105-110.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, pages 475-482.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of COLING*, pages 411-417.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL*, pages 653-662.