Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information*

Jinsong Su[1,2], **Hua Wu**[3], **Haifeng Wang**[3], **Yidong Chen**[1], **Xiaodong Shi**[1], **Huailin Dong**[1], and **Qun Liu**[2]
Xiamen University, Xiamen, China[1]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[2]
Baidu Inc., Beijing, China[3]
{jssu, ydchen, mandel, hldong}@xmu.edu.cn
{wu_hua, wanghaifeng}@baicu.com
liuqun@ict.ac.cn

Abstract

To adapt a translation model trained from the data in one domain to another, previous works paid more attention to the studies of parallel corpus while ignoring the in-domain monolingual corpora which can be obtained more easily. In this paper, we propose a novel approach for translation model adaptation by utilizing in-domain monolingual topic information instead of the in-domain bilingual corpora, which incorporates the topic information into translation probability estimation. Our method establishes the relationship between the out-of-domain bilingual corpus and the in-domain monolingual corpora via a topic mapping and phrase-topic distribution probability estimation from in-domain monolingual corpora. Experimental result on the NIST Chinese-English translation task shows that our approach significantly outperforms the baseline system.

1 Introduction

In recent years, statistical machine translation(SMT) has been rapidly developing with more and more novel translation models being proposed and put into practice (Koehn et al., 2003; Och and Ney, 2004; Galley et al., 2006; Liu et al., 2006; Chiang, 2007; Chiang, 2010). However, similar to other natural language processing(NLP) tasks, SMT systems often suffer from domain adaptation problem during practical applications. The simple reason is that the underlying statistical models always tend to closely approximate the empirical distributions of the training data, which typically consist of bilingual sentences and monolingual target language sentences. When the translated texts and the training data come from the same domain, SMT systems can achieve good performance, otherwise the translation quality degrades dramatically. Therefore, it is of significant importance to develop translation systems which can be effectively transferred from one domain to another, for example, from *newswire* to *weblog*.

According to adaptation emphases, domain adaptation in SMT can be classified into translation model adaptation and language model adaptation. Here we focus on how to adapt a translation model, which is trained from the large-scale out-of-domain bilingual corpus, for domain-specific translation task, leaving others for future work. In this aspect, previous methods can be divided into two categories: one paid attention to collecting more sentence pairs by information retrieval technology (Hildebrand et al., 2005) or synthesized parallel sentences (Ueffing et al., 2008; Wu et al., 2008; Bertoldi and Federico, 2009; Schwenk and Senellart, 2009), and the other exploited the full potential of existing parallel corpus in a *mixture-modeling* (Foster and Kuhn, 2007; Civera and Juan, 2007; Lv et al., 2007) framework. However, these approaches focused on the studies of bilingual corpus synthesis and exploitation while ignoring the monolingual corpora, therefore limiting the potential of further translation quality improvement.

In this paper, we propose a novel adaptation method to adapt the translation model for domain-specific translation task by utilizing in-domain

*Part of this work was done during the first author's internship at Baidu.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 459–468,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

monolingual corpora. Our approach is inspired by the recent studies (Zhao and Xing, 2006; Zhao and Xing, 2007; Tam et al., 2007; Gong and Zhou, 2010; Ruiz and Federico, 2011) which have shown that a particular translation always appears in some specific topical contexts, and the topical context information has a great effect on translation selection. For example, "*bank*" often occurs in the sentences related to the *economy* topic when translated into "*yínháng*", and occurs in the sentences related to the *geography* topic when translated to "*héàn*". Therefore, the co-occurrence frequency of the phrases in some specific context can be used to constrain the translation candidates of phrases. In a monolingual corpus, if "*bank*" occurs more often in the sentences related to the *economy* topic than the ones related to the *geography* topic, it is more likely that "*bank*" is translated to "*yínháng*" than to "*héàn*". With the out-of-domain bilingual corpus, we first incorporate the topic information into translation probability estimation, aiming to quantify the effect of the topical context information on translation selection. Then, we rescore all phrase pairs according to the phrase-topic and the word-topic posterior distributions of the additional in-domain monolingual corpora. As compared to the previous works, our method takes advantage of both the in-domain monolingual corpora and the out-of-domain bilingual corpus to incorporate the topic information into our translation model, thus breaking down the corpus barrier for translation quality improvement. The experimental results on the NIST data set demonstrate the effectiveness of our method.

The reminder of this paper is organized as follows: Section 2 provides a brief description of translation probability estimation. Section 3 introduces the adaptation method which incorporates the topic information into the translation model; Section 4 describes and discusses the experimental results; Section 5 briefly summarizes the recent related work about translation model adaptation. Finally, we end with a conclusion and the future work in Section 6.

2 Background

The statistical translation model, which contains phrase pairs with bi-directional *phrase probabilities* and bi-directional *lexical probabilities*, has a great effect on the performance of SMT system. Phrase probability measures the co-occurrence frequency of a phrase pair, and lexical probability is used to validate the quality of the phrase pair by checking how well its words are translated to each other.

According to the definition proposed by (Koehn et al., 2003), given a source sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$, a target sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, e_I$, and its word alignment $\mathbf{a}$ which is a subset of the Cartesian product of word positions: $\mathbf{a} \subseteq (j, i) : j = 1, \ldots, J; i = 1, \ldots, I$, the phrase pair $(\tilde{f}, \tilde{e})$ is said to be *consistent* (Och and Ney, 2004) with the alignment if and only if: (1) there must be at least one word inside one phrase aligned to a word inside the other phrase and (2) no words inside one phrase can be aligned to a word outside the other phrase. After all consistent phrase pairs are extracted from training corpus, the phrase probabilities are estimated as *relative frequencies* (Och and Ney, 2004):

$$\phi(\tilde{e}|\tilde{f}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{e}'} count(\tilde{f}, \tilde{e}')} \quad (1)$$

Here $count(\tilde{f}, \tilde{e})$ indicates how often the phrase pair $(\tilde{f}, \tilde{e})$ occurs in the training corpus.

To obtain the corresponding lexical weight, we first estimate a lexical translation probability distribution $w(e|f)$ by *relative frequency* from the training corpus:

$$w(e|f) = \frac{count(f, e)}{\sum_{e'} count(f, e')} \quad (2)$$

Retaining the alignment $\tilde{a}$ between the phrase pair $(\tilde{f}, \tilde{e})$, the corresponding lexical weight is calculated as

$$p_w(\tilde{e}|\tilde{f}, \tilde{a}) = \prod_{i=1}^{|\tilde{e}|} \frac{1}{|\{j|(j, i) \in \tilde{a}\}|} \sum_{\forall (j,i) \in \tilde{a}} w(e_i|f_j) \quad (3)$$

However, the above-mentioned method only counts the co-occurrence frequency of bilingual phrases, assuming that the translation probability is independent of the context information. Thus, the statistical model estimated from the training data is not suitable for text translation in different domains, resulting in a significant drop in translation quality.

3 Translation Model Adaptation via Monolingual Topic Information

In this section, we first briefly review the principle of Hidden Topic Markov Model(HTMM) which is the basis of our method, then describe our approach to translation model adaptation in detail.

3.1 Hidden Topic Markov Model

During the last couple of years, topic models such as Probabilistic Latent Semantic Analysis (Hofmann, 1999) and Latent Dirichlet Allocation model (Blei, 2003), have drawn more and more attention and been applied successfully in NLP community. Based on the "bag-of-words" assumption that the order of words can be ignored, these methods model the text corpus by using a co-occurrence matrix of words and documents, and build generative models to infer the latent aspects or topics. Using these models, the words can be clustered into the derived topics with a probability distribution, and the correlation between words can be automatically captured via topics.

However, the "bag-of-words" assumption is an unrealistic oversimplification because it ignores the order of words. To remedy this problem, Gruber et al.(2007) propose HTMM, which models the topics of words in the document as a Markov chain. Based on the assumption that all words in the same sentence have the same topic and the successive sentences are more likely to have the same topic, HTMM incorporates the local dependency between words by Hidden Markov Model for better topic estimation.

HTMM can also be viewed as a soft clustering tool for words in training corpus. That is, HTMM can estimate the probability distribution of a topic over words, i.e. the topic-word distribution $P(word|topic)$ during training. Besides, HTMM derives inherent topics in sentences rather than in documents, so we can easily obtain the sentence-topic distribution $P(topic|sentence)$ in training corpus. Adopting maximum likelihood estimation(MLE), this posterior distribution makes it possible to effectively calculate the word-topic distribution $P(topic|word)$ and the phrase-topic distribution $P(topic|phrase)$ both of which are very important in our method.

3.2 Adapted Phrase Probability Estimation

We utilize the additional in-domain monolingual corpora to adapt the out-of-domain translation model for domain-specific translation task. In detail, we build an adapted translation model in the following steps:

- Build a topic-specific translation model to quantify the effect of the topic information on the translation probability estimation.
- Estimate the topic posterior distributions of phrases in the in-domain monolingual corpora.
- Score the phrase pairs according to the predefined topic-specific translation model and the topic posterior distribution of phrases.

Formally, we incorporate monolingual topic information into translation probability estimation, and decompose the phrase probability $\phi(\tilde{e}|\tilde{f})$[1] as follows:

$$\begin{aligned}\phi(\tilde{e}|\tilde{f}) &= \sum_{t_f} \phi(\tilde{e}, t_f|\tilde{f}) \\ &= \sum_{t_f} \phi(\tilde{e}|\tilde{f}, t_f) \cdot P(t_f|\tilde{f}) \qquad (4)\end{aligned}$$

where $\phi(\tilde{e}|\tilde{f}, t_f)$ indicates the probability of translating $\tilde{f}$ into $\tilde{e}$ given the source-side topic t_f, $P(t_f|\tilde{f})$ denotes the phrase-topic distribution of $\tilde{f}$.

To compute $\phi(\tilde{e}|\tilde{f})$, we first apply HTMM to respectively train two monolingual topic models with the following corpora: one is the source part of the out-of-domain bilingual corpus C_{f_out}, the other is the in-domain monolingual corpus C_{f_in} in the source language. Then, we respectively estimate $\phi(\tilde{e}|\tilde{f}, t_f)$ and $P(t_f|\tilde{f})$ from these two corpora. To avoid confusion, we further refine $\phi(\tilde{e}|\tilde{f}, t_f)$ and $P(t_f|\tilde{f})$ with $\phi(\tilde{e}|\tilde{f}, t_{f\_out})$ and $P(t_{f\_in}|\tilde{f})$, respectively. Here, t_{f_out} is the topic clustered from the corpus C_{f_out}, and t_{f_in} represents the topic derived from the corpus C_{f_in}.

However, the two above-mentioned probabilities can not be directly multiplied in formula (4) because they are related to different topic spaces from

[1] Due to the limit of space, we omit the description of the calculation method of the phrase probability $\phi(\tilde{f}|\tilde{e})$, which can be adjusted in a similar way to $\phi(\tilde{e}|\tilde{f})$ with the help of in-domain monolingual corpus in the target language.

different corpora. Besides, their topic dimensions are not assured to be the same. To solve this problem, we introduce the topic mapping probability $P(t_{f\_out}|t_{f\_in})$ to map the in-domain phrase-topic distribution into the one in the out-domain topic space. To be specific, we obtain the out-of-domain phrase-topic distribution $P(t_{f\_out}|\tilde{f})$ as follows:

$$P(t_{f\_out}|\tilde{f}) = \sum_{t_{f\_in}} P(t_{f\_out}|t_{f\_in}) \cdot P(t_{f\_in}|\tilde{f}) \quad (5)$$

Thus formula (4) can be further refined as the following formula:

$$\begin{aligned} \phi(\tilde{e}|\tilde{f}) &= \sum_{t_{f\_out}} \sum_{t_{f\_in}} \phi(\tilde{e}|\tilde{f}, t_{f\_out}) \\ &\quad \cdot P(t_{f\_out}|t_{f\_in}) \cdot P(t_{f\_in}|\tilde{f}) \quad (6) \end{aligned}$$

Next we will give detailed descriptions of the calculation methods for the three probability distributions mentioned in formula (6).

3.2.1 Topic-Specific Phrase Translation Probability $\phi(\tilde{e}|\tilde{f}, t_{f\_out})$

We follow the common practice (Koehn et al., 2003) to calculate the topic-specific phrase translation probability, and the only difference is that our method takes the topical context information into account when collecting the fractional counts of phrase pairs. With the sentence-topic distribution $P(t_{f\_out}|\mathbf{f})$ from the relevant topic model of C_{f_out}, the conditional probability $\phi(\tilde{e}|\tilde{f}, t_{f\_out})$ can be easily obtained by MLE method:

$$\begin{aligned} &\phi(\tilde{e}|\tilde{f}, t_{f\_out}) \\ &= \frac{\sum_{\langle \mathbf{f},\mathbf{e}\rangle \in C_{out}} count_{\langle \mathbf{f},\mathbf{e}\rangle}(\tilde{f}, \tilde{e}) \cdot P(t_{f\_out}|\mathbf{f})}{\sum_{\tilde{e}'} \sum_{\langle \mathbf{f},\mathbf{e}\rangle \in C_{out}} count_{\langle \mathbf{f},\mathbf{e}\rangle}(\tilde{f}, \tilde{e}') \cdot P(t_{f\_out}|\mathbf{f})} \quad (7) \end{aligned}$$

where C_{out} is the out-of-domain bilingual training corpus, and $count_{\langle \mathbf{f},\mathbf{e}\rangle}(\tilde{f}, \tilde{e})$ denotes the number of the phrase pair $(\tilde{f}, \tilde{e})$ in sentence pair $\langle \mathbf{f}, \mathbf{e}\rangle$.

3.2.2 Topic Mapping Probability $P(t_{f\_out}|t_{f\_in})$

Based on the two monolingual topic models respectively trained from C_{f_in} and C_{f_out}, we compute the topic mapping probability by using source word f as the pivot variable. Noticing that there are some words occurring in one corpus only, we use the words belonging to both corpora during the mapping procedure. Specifically, we decompose $P(t_{f\_out}|t_{f\_in})$ as follows:

$$\begin{aligned} &P(t_{f\_out}|t_{f\_in}) \\ &= \sum_{f \in C_{f\_out} \bigcap C_{f\_in}} P(t_{f\_out}|f) \cdot P(f|t_{f\_in}) \quad (8) \end{aligned}$$

Here we first get $P(f|t_{f\_in})$ directly from the topic model related to C_{f_in}. Then, considering the sentence-topic distribution $P(t_{f\_out}|\mathbf{f})$ from the relevant topic model of C_{f_out}, we define the word-topic distribution $P(t_{f\_out}|f)$ as:

$$\begin{aligned} &P(t_{f\_out}|f) \\ &= \frac{\sum_{\mathbf{f} \in C_{f\_out}} count_{\mathbf{f}}(f) \cdot P(t_{f\_out}|\mathbf{f})}{\sum_{t_{f\_out}} \sum_{\mathbf{f} \in C_{f\_out}} count_{\mathbf{f}}(f) \cdot P(t_{f\_out}|\mathbf{f})} \quad (9) \end{aligned}$$

where $count_{\mathbf{f}}(f)$ denotes the number of the word f in sentence $\mathbf{f}$.

3.2.3 Phrase-Topic Distribution $P(t_{f\_in}|\tilde{f})$

A simple way to compute the phrase-topic distribution is to take the fractional counts from C_{f_in} and then adopt MLE to obtain relative probability. However, it is infeasible in our model because some phrases occur in C_{f_out} while being absent in C_{f_in}. To solve this problem, we further compute this posterior distribution by the interpolation of two models:

$$\begin{aligned} P(t_{f\_in}|\tilde{f}) &= \theta \cdot P_{mle}(t_{f\_in}|\tilde{f}) + \\ &\quad (1-\theta) \cdot P_{word}(t_{f\_in}|\tilde{f}) \quad (10) \end{aligned}$$

where $P_{mle}(t_{f\_in}|\tilde{f})$ indicates the phrase-topic distribution by MLE, $P_{word}(t_{f\_in}|\tilde{f})$ denotes the phrase-topic distribution which is decomposed into the topic posterior distribution at the word level, and θ is the interpolation weight that can be optimized over the development data.

Given the number of the phrase $\tilde{f}$ in sentence $\mathbf{f}$ denoted as $count_{\mathbf{f}}(\tilde{f})$, we compute the in-domain phrase-topic distribution in the following way:

$$\begin{aligned} &P_{mle}(t_{f\_in}|\tilde{f}) \\ &= \frac{\sum_{\mathbf{f} \in C_{f\_in}} count_{\mathbf{f}}(\tilde{f}) \cdot P(t_{f\_in}|\mathbf{f})}{\sum_{t_{f\_in}} \sum_{\mathbf{f} \in C_{f\_in}} count_{\mathbf{f}}(\tilde{f}) \cdot P(t_{f\_in}|\mathbf{f})} \quad (11) \end{aligned}$$

Under the assumption that the topics of all words in the same phrase are independent, we consider two methods to calculate $P_{word}(t_{f\_in}|\tilde{f})$. One is a "Noisy-OR" combination method (Zens and Ney, 2004) which has shown good performance in calculating similarities between bags-of-words in different languages. Using this method, $P_{word}(t_{f\_in}|\tilde{f})$ is defined as:

$$\begin{aligned} & P_{word}(t_{f\_in}|\tilde{f}) \\ = \; & 1 - P_{word}(\bar{t}_{f\_in}|\tilde{f}) \\ \approx \; & 1 - \prod_{f_j \in \tilde{f}} P(\bar{t}_{f\_in}|f_j) \\ = \; & 1 - \prod_{f_j \in \tilde{f}} (1 - P(t_{f\_in}|f_j)) \end{aligned} \quad (12)$$

where $P_{word}(\bar{t}_{f\_in}|\tilde{f})$ represents the probability that t_{f_in} is not the topic of the phrase $\tilde{f}$. Similarly, $P(\bar{t}_{f\_in}|f_j)$ indicates the probability that t_{f_in} is not the topic of the word f_j.

The other method is an "Averaging" combination one. With the assumption that t_{f_in} is the topic of $\tilde{f}$ if at least one of the words in $\tilde{f}$ belongs to this topic, we derive $P_{word}(t_{f\_in}|\tilde{f})$ as follows:

$$P_{word}(t_{f\_in}|\tilde{f}) \approx \sum_{f_j \in \tilde{f}} P(t_{f\_in}|f_j)/|\tilde{f}| \quad (13)$$

where $|\tilde{f}|$ denotes the number of words in phrase $\tilde{f}$.

3.3 Adapted Lexical Probability Estimation

Now we briefly describe how to estimate the adapted lexical weight for phrase pairs, which can be adjusted in a similar way to the phrase probability.

Specifically, adopting our method, each word is considered as one phrase consisting of only one word, so

$$\begin{aligned} w(e|f) \; = \; & \sum_{t_{f\_out}} \sum_{t_{f\_in}} w(e|f, t_{f\_out}) \\ & \cdot P(t_{f\_out}|t_{f\_in}) \cdot P(t_{f\_in}|f) \end{aligned} \quad (14)$$

Here we obtain $w(e|f, t_{f\_out})$ with a similar approach to $\phi(\tilde{e}|\tilde{f}, t_{f\_out})$, and calculate $P(t_{f\_out}|t_{f\_in})$ and $P(t_{f\_in}|f)$ by resorting to formulas (8) and (9).

With the adjusted lexical translation probability, we resort to formula (4) to update the lexical weight for the phrase pair $(\tilde{f}, \tilde{e})$.

4 Experiment

We evaluate our method on the Chinese-to-English translation task for the *weblog* text. After a brief description of the experimental setup, we investigate the effects of various factors on the translation system performance.

4.1 Experimental setup

In our experiments, the out-of-domain training corpus comes from the FBIS corpus and the Hansards part of LDC2004T07 corpus ($54.6K$ documents with $1M$ parallel sentences, $25.2M$ Chinese words and $29M$ English words). We use the Chinese Sohu weblog in 2009[1] and the English Blog Authorship corpus[2] (Schler et al., 2006) as the in-domain monolingual corpora in the source language and target language, respectively. To obtain more accurate topic information by HTMM, we firstly filter the noisy blog documents and the ones consisting of short sentences. After filtering, there are totally $85K$ Chinese blog documents with $2.1M$ sentences and $277K$ English blog documents with $4.3M$ sentences used in our experiments. Then, we sample equal numbers of documents from the in-domain monolingual corpora in the source language and the target language to respectively train two in-domain topic models. The web part of the 2006 NIST MT evaluation test data, consisting of 27 documents with 1048 sentences, is used as the development set, and the weblog part of the 2008 NIST MT test data, including 33 documents with 666 sentences, is our test set.

To obtain various topic distributions for the out-of-domain training corpus and the in-domain monolingual corpora in the source language and the target language respectively, we use HTMM tool developed by Gruber et al.(2007) to conduct topic model training. During this process, we empirically set the same parameter values for the HTMM training of different corpora: $topics = 50$, $\alpha = 1.5$, $\beta = 1.01$, $iters = 100$. See (Gruber et al., 2007) for the meanings of these parameters. Besides, we set the interpolation weight θ in formula (10) to 0.5 by observing the results on development set in the additional experiments.

We choose MOSES, a famous open-source

[1] http://blog.sohu.com/

[2] http://u.cs.biu.ac.il/ koppel/BlogCorpus.html

phrase-based machine translation system (Koehn et al., 2007), as the experimental decoder. GIZA++ (Och and Ney, 2003) and the heuristics "grow-diag-final-and" are used to generate a word-aligned corpus, from which we extract bilingual phrases with maximum length 7. We use SRILM Toolkits (Stolcke, 2002) to train two 4-gram language models on the filtered English Blog Authorship corpus and the Xinhua portion of Gigaword corpus, respectively. During decoding, we set the ttable-limit as 20, the stack-size as 100, and perform minimum-error-rate training (Och and Ney, 2003) to tune the feature weights for the log-linear model. The translation quality is evaluated by case-insensitive BLEU-4 metric (Papineni et al., 2002). Finally, we conduct paired bootstrap sampling (Koehn, 2004) to test the significance in BLEU score differences.

4.2 Result and Analysis

4.2.1 Effect of Different Smoothing Methods

Our first experiments investigate the effect of different smoothing methods for the in-domain phrase-topic distribution: "Noisy-OR" and "Averaging". We build adapted phrase tables with these two methods, and then respectively use them in place of the out-of-domain phrase table to test the system performance. For the purpose of studying the generality of our approach, we carry out comparative experiments on two sizes of in-domain monolingual corpora: $5K$ and $40K$.

Adaptation Method	(Dev) MT06 Web	(Tst) MT08 Weblog
Baseline	30.98	20.22
Noisy-OR ($5K$)	31.16	20.45
Averaging ($5K$)	31.51	20.54
Noisy-OR ($40K$)	31.87	20.76
Averaging ($40K$)	31.89	21.11

Table 1: Experimental results using different smoothing methods.

Table 1 reports the BLEU scores of the translation system under various conditions. Using the out-of-domain phrase table, the baseline system achieves a BLEU score of 20.22. In the experiments with the small-scale in-domain monolingual corpora, the BLEU scores acquired by two methods are 20.45 and 20.54, achieving absolute improvements of 0.23 and 0.32 on the test set, respectively. In the experiments with the large-scale monolingual in-domain corpora, similar results are obtained, with absolute improvements of 0.54 and 0.89 over the baseline system.

From the above experimental results, we know that both "Noisy-OR" and "Averaging" combination methods improve the performance over the baseline, and "Averaging" method seems to be slightly better. This finding fails to echo the promising results in the previous study (Zens and Ney, 2004). This is because the "Noisy-OR" method involves the multiplication of the word-topic distribution (shown in formula (12)), which leads to much sharper phrase-topic distribution than "Averaging" method, and is more likely to introduce bias to the translation probability estimation. Due to this reason, all the following experiments only consider the "Averaging"method.

4.2.2 Effect of Combining Two Phrase Tables

In the above experiments, we replace the out-of-domain phrase table with the adapted phrase table. Here we combine these two phrase tables in a log-linear framework to see if we could obtain further improvement. To offer a clear description, we represent the out-of-domain phrase table and the adapted phrase table with "OutBP" and "AdapBP", respectively.

Used Phrase Table	(Dev) MT06 Web	(Tst) MT08 Weblog
Baseline	30.98	20.22
AdapBp ($5K$)	31.51	20.54
+ OutBp	31.84	20.70
AdapBp ($40K$)	31.89	21.11
+ OutBp	32.05	21.20

Table 2: Experimental results using different phrase tables. OutBp: the out-of-domain phrase table. AdapBp: the adapted phrase table.

Table 2 shows the results of experiments using different phrase tables. Applying our adaptation approach, both "AdapBP" and "OutBP + AdapBP" consistently outperform the baseline, and the lat-

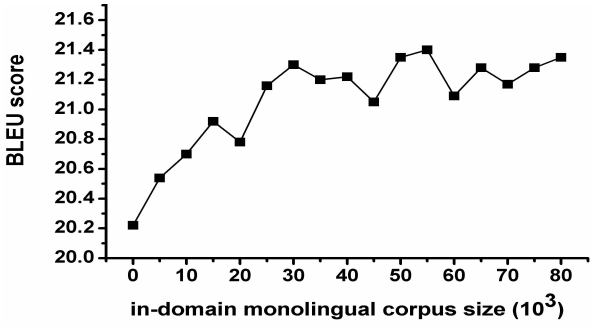

Figure 1: Effect of in-domain monolingual corpus size on translation quality.

ter produces further improvements over the former. Specifically, the BLEU scores of the "OutBP + AdapBP" method are 20.70 and 21.20, which obtain 0.48 and 0.98 points higher than the baseline method, and 0.16 and 0.09 points higher than the 'AdapBP" method. The underlying reason is that the probability distribution of each in-domain sentence often converges on some topics in the "AdapBP" method and some translation probabilities are overestimated, which leads to negative effects on the translation quality. By using two tables together, our approach reduces the bias introduced by "AdapBP", therefore further improving the translation quality.

4.2.3 Effect of In-domain Monolingual Corpus Size

Finally, we investigate the effect of in-domain monolingual corpus size on translation quality. In the experiment, we try different sizes of in-domain documents to train different monolingual topic models: from $5K$ to $80K$ with an increment of $5K$ each time. Note that here we only focus on the experiments using the "OutBP + AdapBP" method, because this method performs better in the previous experiments.

Figure 1 shows the BLEU scores of the translation system on the test set. It can be seen that the more data, the better translation quality when the corpus size is less than $30K$. The overall BLEU scores corresponding to the range of great N values are generally higher than the ones corresponding to the range of small N values. For example, the BLEU scores under the condition within the range $[25K, 80K]$ are all higher than the ones within the range $[5K, 20K]$. When N is set to $55K$, the BLEU score of our system is **21.40**, with **1.18** gains on the baseline system. This difference is statistically significant at $P < 0.01$ using the significance test tool developed by Zhang et al.(2004). For this experimental result, we speculate that with the increment of in-domain monolingual data, the corresponding topic models provide more accurate topic information to improve the translation system. However, this effect weakens when the monolingual corpora continue to increase.

5 Related work

Most previous researches about translation model adaptation focused on *parallel data collection*. For example, Hildebrand et al.(2005) employed information retrieval technology to gather the bilingual sentences, which are similar to the test set, from available in-domain and out-of-domain training data to build an adaptive translation model. With the same motivation, Munteanu and Marcu (2005) extracted in-domain bilingual sentence pairs from comparable corpora. Since large-scale monolingual corpus is easier to obtain than parallel corpus, there have been some studies on how to generate parallel sentences with monolingual sentences. In this respect, Ueffing et al. (2008) explored semi-supervised learning to obtain synthetic parallel sentences, and Wu et al. (2008) used an in-domain translation dictionary and monolingual corpora to adapt an out-of-domain translation model for the in-domain text.

Differing from the above-mentioned works on the acquirement of bilingual resource, several studies (Foster and Kuhn, 2007; Civera and Juan, 2007; Lv et al., 2007) adopted *mixture modeling* framework to exploit the full potential of the existing parallel corpus. Under this framework, the training corpus is first divided into different parts, each of which is used to train a sub translation model, then these sub models are used together with different weights during decoding. In addition, discriminative weighting methods were proposed to assign appropriate weights to the sentences from training corpus (Matsoukas et al., 2009) or the phrase pairs of phrase table (Foster et al., 2010). Final experimental results show that without using any additional resources, these approaches all improve SMT performance sig-

nificantly.

Our method deals with translation model adaptation by making use of the topical context, so let us take a look at the recent research development on the application of topic models in SMT. Assuming each bilingual sentence constitutes a mixture of hidden topics and each word pair follows a topic-specific bilingual translation model, Zhao and Xing (2006,2007) presented a bilingual topical admixture formalism to improve word alignment by capturing topic sharing at different levels of linguistic granularity. Tam et al.(2007) proposed a bilingual LSA, which enforces one-to-one topic correspondence and enables latent topic distributions to be efficiently transferred across languages, to cross-lingual language modeling and translation lexicon adaptation. Recently, Gong and Zhou (2010) also applied topic modeling into domain adaptation in SMT. Their method employed one additional feature function to capture the topic inherent in the source phrase and help the decoder dynamically choose related target phrases according to the specific topic of the source phrase.

Besides, our approach is also related to context-dependent translation. Recent studies have shown that SMT systems can benefit from the utilization of context information. For example, *trigger-based lexicon model* (Hasan et al., 2008; Mauser et al., 2009) and *context-dependent translation selection* (Chan et al., 2007; Carpuat and Wu, 2007; He et al., 2008; Liu et al., 2008). The former generated triplets to capture long-distance dependencies that go beyond the local context of phrases, and the latter built the classifiers which combine rich context information to better select translation during decoding. With the consideration of various local context features, these approaches all yielded stable improvements on different translation tasks.

As compared to the above-mentioned works, our work has the following differences.

- We focus on how to adapt a translation model for domain-specific translation task with the help of additional in-domain monolingual corpora, which are far from full exploitation in the *parallel data collection* and *mixture modeling* framework.
- In addition to the utilization of in-domain monolingual corpora, our method is different from the previous works (Zhao and Xing, 2006; Zhao and Xing, 2007; Tam et al., 2007; Gong and Zhou, 2010) in the following aspects: (1) we use a different topic model — HTMM which has different assumption from PLSA and LDA; (2) rather than modeling topic-dependent translation lexicons in the training process, we estimate topic-specific lexical probability by taking account of topical context when extracting word pairs, so our method can also be directly applied to topic-dependent phrase probability modeling. (3) Instead of rescoring phrase pairs online, our approach calculate the translation probabilities offline, which brings no additional burden to translation systems and is suitable to translate the texts without the topic distribution information.
- Different from *trigger-based lexicon model* and *context-dependent translation selection* both of which put emphasis on solving the translation ambiguity by the exploitation of the context information at the sentence level, we adopt the topical context information in our method for the following reasons: (1) the topic information captures the context information beyond the scope of sentence; (2) the topical context information is integrated into the posterior probability distribution, avoiding the sparseness of word or POS features; (3) the topical context information allows for more fine-grained distinction of different translations than the genre information of corpus.

6 Conclusion and future work

This paper presents a novel method for SMT system adaptation by making use of the monolingual corpora in new domains. Our approach first estimates the translation probabilities from the out-of-domain bilingual corpus given the topic information, and then rescores the phrase pairs via topic mapping and phrase-topic distribution probability estimation from in-domain monolingual corpora. Experimental results show that our method achieves better performance than the baseline system, without increasing the burden of the translation system.

In the future, we will verify our method on oth-

er language pairs, for example, Chinese to Japanese. Furthermore, since the in-domain phrase-topic distribution is currently estimated with simple smoothing interpolations, we expect that the translation system could benefit from other sophisticated smoothing methods. Finally, the reasonable estimation of topic number for better translation model adaptation will also become our study emphasis.

Acknowledgement

The authors were supported by 863 State Key Project (Grant No. 2011AA01A207), National Natural Science Foundation of China (Grant Nos. 61005052 and 61103101), Key Technologies R&D Program of China (Grant No. 2012BAH14F03). We thank the anonymous reviewers for their insightful comments. We are also grateful to Ruiyu Fang and Jinming Hu for their kind help in data processing.

References

Michiel Bacchiani and Brian Roark. 2003. Unsupervised Language Model Adaptation. In *Proc. of ICASSP 2003*, pages 224-227.

Michiel Bacchiani and Brian Roark. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, pages 477-504.

Nicola Bertoldi and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proc. of ACL Workshop 2009*, pages 182-189.

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

Ivan Bulyko, Spyros Matsoukas, Richard Schwartz, Long Nguyen and John Makhoul. 2007. Language Model Adaptation in Machine Translation from Speech. In *Proc. of ICASSP 2007*, pages 117-120.

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. 2006. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL 2007*, pages 33-40.

Boxing Chen, George Foster and Roland Kuhn. 2010. Bilingual Sense Similarity for Statistical Machine Translation. In *Proc. of ACL 2010*, pages 834-843.

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

David Chiang. 2010. Learning to Translate with Source and Target Syntax. In *Proc. of ACL 2010*, pages 1443-1452.

Jorge Civera and Alfons Juan. 2007. Domain Adaptation in Statistical Machine Translation with Mixture Modelling. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 177-180.

Matthias Eck, Stephan Vogel and Alex Waibel. 2004. Language Model Adaptation for Statistical Machine Translation Based on Information Retrieval. In *Proc. of Fourth International Conference on Language Resources and Evaluation*, pages 327-330.

Matthias Eck, Stephan Vogel and Alex Waibel. 2005. Low Cost Portability for Statistical Machine Translation Based on N-gram Coverage. In *Proc. of MT Summit 2005*, pages 227-234.

George Foster and Roland Kuhn. 2007. Mixture Model Adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128-135.

George Foster, Cyril Goutte and Roland Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proc. of EMNLP 2010*, pages 451-459.

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 MT SUMMIT 2010*, pages 24-28.

Amit Gruber, Michal Rosen-Zvi and Yair Weiss. 2007. Hidden Topic Markov Models. In *Journal of Machine Learning Research*, pages 163-170.

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.

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

Almut Silja Hildebrand. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proc. of EAMT 2005*, pages 133-142.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proc. of SIGIR 1999*, pages 50-57.

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

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

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

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL 2007, Demonstration Session*, pages 177-180.

Yang Liu, Qun Liu and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proc. of ACL 2006*, pages 609-616.

Yajuan Lv, Jin Huang and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proc. of EMNLP 2007*, pages 343-350.

Arne Mauser, Richard Zens and Evgeny Matusov, Saša Hasan and Hermann Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *Proc. of International Workshop on Spoken Language Translation*, pages 103-110.

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

Spyros Matsoukas, Antti-Veikko I. Rosti and Bing Zhang 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *Proc. of EMNLP 2009*, pages 708-717.

Nick Ruiz and Marcello Federico. 2011. Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models. In *Proc. of ACL Workshop 2011*, pages 294-302.

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

Jonathan Schler, Moshe Koppel, Shlomo Argamon and James Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proc. of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Holger Schwenk and Jean Senellart. 2009. Translation Model Adaptation for an Arabic/french News Translation System by Lightly-supervised Training. In *Proc. of MT Summit XII*.

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

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

Nicola Ueffing, Gholamreza Haffari and Anoop Sarkar. 2008. Semi-supervised Model Adaptation for Statistical Machine Translation. *Machine Translation*, pages 77-94.

Hua Wu, Haifeng Wang and Chengqing Zong. 2008. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora. In *Proc. of COLING 2008*, pages 993-1000.

Richard Zens and Hermann Ney. 2004. Improvments in phrase-based statistical machine translation. In *Proc. of NAACL 2004*, pages 257-264.

Ying Zhang, Almut Silja Hildebrand and Stephan Vogel. 2006. Distributed Language Modeling for N-best List Re-ranking. In *Proc. of EMNLP 2006*, pages 216-223.

Bing Zhao, Matthias Eck and Stephan Vogel. 2004. Language Model Adaptation for Statistical Machine Translation with Structured Query Models. In *Proc. of COLING 2004*, pages 411-417.

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.

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.