Employing Topic Modeling for Statistical Machine Translation

Gong Zhengxian
School of Computer Science and Technology
Soochow University
Suzhou, China
zhxgong@suda.edu.cn

Zhou Guodong
School of Computer Science and Technology
Soochow University
Suzhou, China
gdzhou@suda.edu.cn 20084227065099@suda.edu.c

Abstract—**The mixture modeling approaches have dominated the research of domain adaptation in Statistical Machine Translation (SMT). Such approaches construct a general model and several sub-models in advance and focus on the way of determining the relative importance of all the models. In this paper, we propose a simple yet effective approach for better domain adaptation in phrase-based SMT via topic modeling. Different from existing approaches, our topic modeling approach employs 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. Evaluation on a conversation corpus shows very encouraging results.**

Keywords-topic modeling;domain adaptation;statistical macine translation

I. INTRODUCTION

In the literature, the mixture modeling approaches have dominated the research of domain adaptation in SMT (Hastie et al. 2001; Zhao et al. 2004; Foster and Kuhn, 2007; Civera and Juan, 2007; Finch and Sumita, 2008; Bertoldi and Federico, 2009). Such approaches normally first partition the training data into different specific domains, then train a sub-model on each specific domain and finally weight each sub-model appropriately according to the matching between the specific domain and the domain of the test data. However, the main problem in mixture modeling is that training sub-models on small corpus will make these models unreliable.

This paper proposes a topic modeling approach to directly model the deviation between different specific domains (hereafter topics) by introducing one additional feature function to capture such topic information instead of constructing and combining multiple sub-models. As a baseline, a phrase-based SMT system is employed. The intuition behind is to give more probabilities to those related target phrases according to the specific topic of the source phrase. In this paper we employ a topic model to resolve this issue.

Here we assume the availability of a small-scale in-domain parallel corpus, which is used to build two topic models in the source language and the target language respectively. Here the source language topic model is mainly used to help detect the topic of the source input text while the target language topic model is employed to offer the word distribution of the target language for each topic derived from the target language part of the in-domain parallel corpuses. Since a phrase in phrase-based SMT can be an arbitrary word sequences, we can easily estimate the topic distribution of a target phrase using the target language topic model.

Evaluation on a conversation corpus shows that topic modeling can well address the domain adaptation problem in SMT by recasting specific domains as "topics" and greatly improve the performance of a phrase-based SMT system. The rest of this paper is organized as follows. Section 2 and 3 describes our corpora and baseline phrase-based SMT system respectively. Section 4 describes topic modeling in domain expression. Section 5 focuses on how to integrate topic modeling into our baseline phrase-based SMT system. Section 6 gives the experimental results. Finally, we conclude in Section 7.

II. CORPORA

Our experiments were done on Chinese to English SMT on a small sample of parallel in-domain corpus over five domains: Food (& Beverage), Transport, Travel, Business and Sports. Table 1 shows the statistics of our in-domain corpus over five domains

Table 1: Statistics of the small-scale in-domain parallel corpus over the five domains

In-Domain text	#Sentences
Food	11352
Business	9605
Transport	11869
Travel	11408
Sports	7993

This small sample of parallel in-domain corpus comes from a subset of CWMT2009, which developed by HARBIN institute of technology. The in-domain corpus contains 40,258 sentences in sum. And we adopt FBIS as a general (out-domain) corpus, which has about 239,413 sentences.

In particular, we split parallel in-domain corpus into monolingual corpus and trained two monolingual topic models in the source (Chinese) language and the target (English) language on the five specific domains. Section 4 describes this procedure in detail.

Table 2 shows the statistics of the bilingual development data and test data. Our development corpus and the whole evaluation data both come from the dialog

978-1-4244-8728-8/11/$26.00 ©2011 IEEE

Table 2: Statistics of bilingual development data and test data

#Dev sentences:437		
863 2003 test set	mix-domain	437
	Transportation	180
#Test sentences: 855		
863 2004 & 2005 test set	Food	131
	Transportation	161
	Travel	221
	Business	159
	Sports	183

part of 863 Machine Translation Evaluation Data (863 for short).

III. BASELINE: A PHRASE-BASED SMT SYSTEM

The translation process of SMT can be modeled as obtaining the translation e_{best} of the source sentence f by maximizing the following posterior probability (Brown et al., 1993).

$$e_{best} = \arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e) P_{lm}(e) \quad (1)$$

Where P(e|f) is a translation model and P_{lm} is a language model.

As a baseline, our phrase-based SMT follows Koehn et al. (2003) and adopt the following six popular feature functions: 1) two phrase translation probabilities (two directions); 2) two word translation probabilities (two directions); 3) one language model (target language); 4) distance-based model; (5) one phrase penalty (target language); (6) one word penalty (target language). In addition, the log-linear model as described in (Och and Ney, 2003) is employed to linearly interpolate these features for obtaining the best translation according to the formula (2).

These features are combined in the log-linear model (Och and Ney, 2003) to obtain the best translation according to the formula (2).

$$e_{best} = \arg\max \{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \} \quad (2)$$

Where $h_m(e , f)$ indicates a feature function, and λm its weight optimized using a discriminative training method on a held-out development data.

In phrase-based SMT, one source phrase may correspond to multiple target phrases with different probabilities. For example, the source (Chinese) phrase "一张" can correspond to the target phrase " a ticket" or " a table" depending on its context. Obviously, if the source phrase "一张" occurs in the specific domain of transportation, it may correspond to the target phrase "a ticket" more likely.

A typical approach to achieve the goal is to employ a mixture modeling approach, which gives more probabilities to those target phrases by assigning a biased weight to the specific sub-model according to the specific domain of the source phrase before combining it with a general model. The specific domain info in this paper is expressed using topic modeling.

IV. TOPIC MODELING IN DOMAIN EXPRESSION

A. Topic Modeling

Among various topic models, LDA has drawn more and more attention in the NLP community and has been applied successfully in topic detection. In principle, LDA is a generative three-level hierarchical Bayesian probabilistic model for analyzing the content of documents and the meaning of words. LDA is essentially the Bayesian version of PLSI (Hoffman, 1999). Since Bayesian formulation tends to avoid over-fitting and thus performs better on small datasets, we employ LDA in domain adaptation for SMT.

Basically, we can view LDA as a soft clustering tool because it derives inherent topics in the documents and clusters the words in the documents into the derived topics with a probability distribution. That is, LDA can give the probability distribution of a topic over the words, i.e. the word-topic distribution $p(word_j|topic_i)$ during training.

In addition, LDA can be viewed as a classifier essentially due to its inference ability. Given various topic models established during training, i.e. the word-topic distributions for various topics in both source and target languages, we can easily derive the topic distribution for a new text, $p(topic_i|doc_{new})$. That is, LDA can tell us which topic a new text belongs to. Therefore, using the inference ability of LDA, we can easily detect the specific domain (i.e. topic) of a source text. Similarly, we can get the topic distribution for a target phrase by averaging the word-topic distributions over all the words in the target phrase.

B. Domain Expression

For domain expression in both source and target languages, we first split the in-domain parallel corpus into two in-domain monolingual corpora in the source language and the target language respectively. Then, we train a source language topic model on the in-domain source language corpus over all the specific domains. Similarly, we can get a target language topic model. Table 3 shows the 5 top-ranked words in the "word-topic" distribution of each topic in target language. From Table 3, it is easy to see that topics 1-5 are related to the specific domains of Transport, Sports, Business, Food and Travel respectively. We will use this result to compute the relevance between a target phrase and a topic.

In addition, we can use LDA to infer the topic of a source input text from the source language topic model as indicated in Section 4.1. Table 4 presents the results produced by LDA in which documents 1-5 are picked from the specific domains of Food &Beverage, Transport, Travel, Business and Sports respectively. It shows that LDA can correctly infer the topic of each document.

C. Mapping between the source and target topic models

Because the source language topic model is built separately from the target one, the correspondence between them needs to be derived automatically for the particular topic distribution to be transferred from the source side to the target side. For this purpose, a mapping between the source and target topic models should be. established. In this paper, such mapping is determined automatically as follows:

(1)Perform word alignment on the special bilingual

Table 3 words-topics(domains) distribution in target language

Topic1		Topic2		Topic 3		Topic 4		Topic 5	
word	P(w|t)	word	P(w|t)	word	P(w|t)	word	P(w|t)	word	P(w|t)
car	0.019103	game	0.016787	Mr.	0.01196	please	0.02252	time	0.01034
time	0.018187	ball	0.016157	company	0.00637	have	0.019200	look	0.01006
train	0.008172	Olympic	0.010374	business	0.00545	drink	0.01603	help	0.00825
bus	0.007897	team	0.008462	price	0.00532	dinner	0.015736	luggage	0.00825
flight	0.007087	Match	0.008349	office	0.00532	food	0.010437	like	0.00714
(Transport)		(Sports)		(Business)		(Food&Beverage)		(Travel)	

Table 4 examples of monolingual documents-topics(domains) distribution

Domain Doc	Transport	Sports	Business	Food & Beverage	Travel
1	0.040299	0.124648	0.064667	**0.692596**	0.077788
2	**0.736808**	0.089637	0.045136	0.004450	0.123966
3	0.108320	0.098901	0.048665	0.017268	0.726844
4	0.090681	6.253908E-4	**0.573483**	0.239524	0.095684
5	0.045122	**0.813493**	0.029651	0.009024	0.102707

corpus, as described in Section 4.2, using GIZA++(Och and Ney,2000) in two directions, augmented to improve recall using the grow-diagonal-final heuristic.

(2)Choose the top-n (n is fine-tuned to 200 in this paper) word-topic distribution of each topic in both languages.

(3)With the help of lexical mapping (obtained from Step 1), pairwise comparison is performed based on Step 2. We count the mapping words between two topics in both languages and sum their distribution value to determine the mapping.

In this way, we can obtain the one-to-one topic correspondence between the source and target LDA-style topic models so that the particular topic distribution can be transferred from the source language to the target language

V. TOPIC MODELING IN PHRASE-BASED SMT

A. Computing the Relevance between a Target Phrase and a Topic

Given a parallel corpus, we can easily train a translation model, e.g. using MERT(Och,2003), and get a phrase table covering all topics. Given a specific topic, we should give more probabilities to those related target phrases, which widely occurs in the specific topic, and punish those unrelated target phrases, which scarcely occurs in this specific domain.

Assume that the topic of the source input text, denoted as *Ts*, has been determined, we can easily find the corresponding topic *Tg* in the target language topic model by looking up the source-to-target topic mapping table as constructed in Section 4.3. Besides, since a phrase in a phrase-based SMT system can be regarded as a word sequence, we can easily compute the topic distribution of a target phrase by averaging the word-topic distributions of all the words in the target phrase. For each target phrase PS_t in topic T_g, let the words of PS_t be $\{W_1, W_2, \ldots W_N\}$, the topic relevance for PS_t can be calculated using formula (3).

$$\mathrm{Re}\,l(PSt, Tg) = (\sum_{j=1}^{N} p(W_j \mid Tg)) / N * p(Tg) \quad (3)$$

Here, $p(w_j|t_g)$ is a "word-topic" distribution in the target language. Obviously, we can select the related target phrases with a high topic relevance value. The reason of introducing $p(T_g)$ into formula (3) is that we can easily back off to the baseline if we set $p(T_g)$ as zero. In addition, $p(T_g)$ measures the prior probability of a topic. Although T_g is normally unknown for the target language, we can estimate it from the source input text instead

B. Determining the Topic of a Source In-put Text

Foster and Kuhn (2007) proposed several measures, such as TF/IDF, LSA(Latent Semantic Analysis), and the perplexity of the language model, to capture the relationship between the source input text and a topic.. They found that the difference between different measures is rather small. In this paper, we focus on a topic-driven measure.

With the help of various topic models established during training, including the topic-document distributions and word-topic distributions, LDA can tell us which topic a new document belongs to (shown in Table 4). Now if we treat the source input text as a document, it would be easy to determine the topic of the source input text using following two schemes:

1) One topic per-document. Since one specific domain is regarded as a topic in this paper, it is natural to assume that a document only contains one topic.

2) Multiple topics per-document. When we evaluate the performance of a SMT system, we often tend to put a lot of sentences together, which perhaps come from multiple domains and contained multiple topics. In this scene, we can first divide such document into individual parts and let each part conform to the style of "One topic per-document".

Our test data belongs to the latter case. In order to regulate automatically, we need to use "word-topic" distribution in source language. This time we need compute the relevance between a source sentence and a topic. That is, we need to score topic relevance for a sentence over all the specific topics. After that, we label a sentence with the topic according to the highest topic relevance score. Finally we can group sentences into several documents according to different topic.

After we inferring the topic for a test document, we can obtain the probability $P(t_i|doc_{new})$ and we determine the topic's prior probability as follows:

$p(T_g)=\max(p(t_i|doc_{new}))$,

i=1…H, H is the number of topics.

C. Integrating Topic Modeling into Phrase-based SMT via an Additional Feature

Similar to Koehn et al. (2003), our decoder implements

Table 5: Weights of various features obtained by MERT training. Note: The last one indicates the topic relevance feature.

System	LM(e)	$P_{phr}(e\|f)$	$P_w(e\|f)$	$P_{phr}(f\|e)$	$P_w(f\|e)$	PP(f)	WP(e)	Rel(e)
Baseline	0.2816	0.0591	0.0154	0.0951	0.1149	0.1073	-0.3501	--
TopicModel	0.2305	0.0472	0.0179	0.0872	0.0175	0.1106	-0.2641	0.2401

Table 6: BLEU and NIST scores

Corpus	System	1-gram	2-gram	3-gram	4-gram	BLEU	NIST
general corpus	**Baseline**	56.52	23.67	10.76	4.89	16.29	3.633
	TopicModel	56.54	23.68	10.81	4.91	16.33	3.638
general corpus + Special corpus	**Baseline**	57.07	24.77	11.56	5.54	17.35	3.877
	TopicModel	57.61	26.07	12.48	5.61	18.00	3.931

a beam search. In particular, the decoder decides whether using a specified translation by informing the decoder the topic knowledge as a soft preference by adding the following additional feature:

$$\mathrm{Rel}(\mathrm{e},\mathrm{Tg}) = (\sum_{d=1}^{D} \mathrm{Re}\, l(PS_{t\_d}, Tg)) / D \quad (4)$$

where $\mathrm{Rel}(PS_{t\_d},T_g)$ is obtained by formula (3), D is the number of target phrases, and the weight of this feature, λ_r, can be obtained using MERT. In this way, our translation model becomes

$$\mathrm{P}(\mathrm{e}\,|\,\mathrm{f},\mathrm{Tg}) = \sum_{m=1}^{M} \lambda_m h_m(e,f) + \lambda_r * \log(\mathrm{Rel}(\mathrm{e},\mathrm{Tg})) \quad (5)$$

In this paper, when $p(T_g) < \varepsilon$ (fine-tuned to 0.5, $P(T_g)$ is the part of Rel(.) showed in Section 5.1), we let $p(T_g)$ equal to 0. This means the additional topic relevance feature will be considered invalid when $p(T_i) < \varepsilon$. Under this case, we are not sure which topic the source input text should belong to and the translation model should back off to the baseline model.

VI. Experimental Results

Our experiments were done on Chinese to English translation. Here, we used the SRI Language Modeling Toolkit to train a trigram general language model on English newswire text, mostly from the Xinhua portion of the Gigaword corpus(2007) and performed word alignment on the training parallel corpus(described in Section 5.1) using GIZA++(Och and Ney,2000) in two directions. For evaluation, the NIST BLEU script(version 11b) with the default setting is used to calculate the Bleu score(Papineni et al.,2002), which measures case-insensitive matching of n-grams with n up to 4.

We choose 180 sentences from the transportation topic in the development corpus to fine-tune the weight for our additional topic relevance feature (Rel(e)). Table 5 shows the weights of various features for the baseline and topic modeling systems. We notice that the weights for Rel(e) is close to LM(e). As shown in Table 3, word-topic distributions is similar to the unigram language model, so it is reasonable to have similar weights for LM and Rel(e). We also notice that the weight of WP(e) is much changed in our TopicModel, much due to the fact that Rel(e) tends to effect phrase selection in the target language as WP(e) does. In this paper, we employ the same weights for all the other four topics.

Before using our system, we first group the test text into 6 topics(additional topic is miscellaneous whose $p(T_i)=0$). Table 6 shows the contribution of topic modeling with or without the small-scale special parallel corpus. It shows that topic modeling with the help of the small-scale special parallel corpus much improves the performance, e.g. by 0.65 in BLEU. It also shows that the contribution of topic modeling without the help of the small-scale special parallel corpus can be ignored. This suggests the usefulness of a small-scale special (in-domain) parallel corpus in domain-adaptation.

VII. Related Work

Earlier domain adaptation research in SMT had been limited to language modeling (LM). For example, Zhao et al. (2004) converted the machine translation output into queries and extracted similar sentences from a large monolingual text collection. Specific language models were then built from the retrieved data and interpolated with a generic language model. Hasan and Ney (2005) proposed a class-dependent language model based on regular expressions for domain adaptation in SMT.

Recently, several studies (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Civera and Juan, 2007) investigated mixture modeling (Hastie et al., 2001) for both the translation model and language model in SMT. For example, Foster and Kuhn (2007) successfully employed a mixture modeling approach in SMT by first partitioning the training data into various bins during training and then decoding these bins independently using different models specific to the predicted domain of a source sentence. Koehn and Schroeder (Koehn and Schroeder ,2007) also showed significant results obtained from mixture modeling.

In particular for topic modeling in SMT, various topic modeling methods have been largely used in LM adaptation, such as latent semantic analysis (LSA)(Bellegarda, 2000), probabilistic latent semantic analysis (PLSA) (Gildea and Hofmann,1999) and LDA (Blei et al., 2003). For example, Tam and Schultz (2005) successfully implemented unsupervised LM adaptation by interpolating the generic LM with a dynamic unigram LM estimated from the LDA model while Hsu and Glass (2006) and Mrva and Woodland (2006) investigated using LDA to allow for both topic and style adaptation. However, there are only a few studies on employing topic modeling for the translation model. Zhao et al. (2006) proposed a Bilingual Topic AdMixture Model (BiTam) to improve word alignment. They further extended this idea to HMM-BiTam, which displays topic patterns for bilingual corpora in LDA-Style and infer optimal translation using the document context (Zhao et al., 2007). Tam et al. (2007) proposed a bilingual latent semantic analysis approach (bilingual LSA) to cross-lingual language modeling and translation lexicon adaptation for SMT by better deriving word alignment.

Our topic modeling approach differs from previous ones in that we integrate topic modeling into phrase-based SMT via one single topic relevance feature. In details,

(1) Instead of improving translation indirectly by improving the word alignment quality (BiTam and HMM-BiTam), we model the topic information via one additional topic relevance feature to directly improve the SMT performance.

(2) Instead of performing language model adaptation across languages (bilingual LSA), we employ topic modeling to adjust the translation model.

(3) Although bilingual LSA can achieve lexicon adaptation, it re-computes phrase scores offline, i.e. in a static way, instead of the online nature of our proposed approach.

VIII. Conclusion and Feature Work

Proper topic modeling for SMT shows that despite ignoring the document structure, the translation of a document from one topic (actually, domain) can be aided by the topic knowledge due to the low variability of lexical choices in the specific topic. Furthermore, it shows that the topic knowledge can significantly improve the performance of a phrase-based SMT system. In particular, we pave a way to integrate a topic model into SMT, which not only enhances the ability for domain adaptation but also avoids the linear growth of parameter numbers for interpolation of sub-models. Although, this paper adopts a widely-used phrase-based SMT system as a baseline, our approach can be applied to other kinds of SMT systems, such as Tree-to-String systems (Liu et al. 2006).

In this paper, we only apply topic modeling into the translation model on a small-scale corpus. In the future work, we will explore it more on a large-scale corpus and in the language model.

Acknowledgment

This work was partially supported by a grant from the National Natural Science Foundation of China No. 61003155

References

[1] Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with Monolingual resources. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 182-189.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning Research, pages 993–1022.

[3] PF Brown, SA Della Pietra, VJ Della Pietra,RL Mercer.1992.The Mathematics of Statistical Machine Translation:Parameter Estimation. Computational Linguistics. 19(2):263-309.

[4] Jorge Civera and Alfons Juan.2007. Domain Adaptation in Statistical Machine Translation with Mixture Modelling. Proceedings of the Second Workshop on Statistical Machine Translation,pages 177-180.

[5] Andrew FINCH and Eiichiro SUMITA.2008.Dynamic Model Interpolation for Statistical Machine Translation.Proceedings of the Third Workshop on Statistical Machine Translation,pages 208-215.

[6] George Foster and Roland Kuhn.2009. Mixture-model Adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation,pages 128–135.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman.2001. The Elements of Statistical Learning. Springer.

[8] Thomas Hofmann.1999.Probabilistic Latent Semantic Indexing. In Proceedings of 22nd annual international ACM SIGIR99,pages 50-57.

[9] Philipp Koehn,Franz .Josef Och ,and DanielMarcu.2003. Statistical Phrase-Based Translation .Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,pages 48-54.

[10] Yang Liu , Qun Liu , and Shouxun Lin. 2006.Tree-to-String Alignment Template for Statistical Machine Translation. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616.

[11] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. 2003.In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167.

[12] Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proc. of ACL00, pages 440–447.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. BLEU:A method for Automatic Evaluation of Machine Translation.In Proc. of ACL02, pages 311–318.

[14] Bing Zhao, Matthias Eck, and Stephan Vogel. 2004.Language Model Adaptation for Statistical Machine Translation with Structured Query Models. In COLING 2004, Geneva, August.