Bilingual-LSA Based LM Adaptation for Spoken Language Translation

Yik-Cheung Tam and Ian Lane and Tanja Schultz
InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{yct,ian.lane,tanja}@cs.cmu.edu

Abstract

We propose a novel approach to crosslingual language model (LM) adaptation based on bilingual Latent Semantic Analysis (bLSA). A bLSA model is introduced which enables latent topic distributions to be efficiently transferred across languages by enforcing a one-to-one topic correspondence during training. Using the proposed bLSA framework crosslingual LM adaptation can be performed by, first, inferring the topic posterior distribution of the source text and then applying the inferred distribution to the target language N-gram LM via marginal adaptation. The proposed framework also enables rapid bootstrapping of LSA models for new languages based on a source LSA model from another language. On Chinese to English speech and text translation the proposed bLSA framework successfully reduced word perplexity of the English LM by over 27% for a unigram LM and up to 13.6% for a 4-gram LM. Furthermore, the proposed approach consistently improved machine translation quality on both speech and text based adaptation.

1 Introduction

Language model adaptation is crucial to numerous speech and translation tasks as it enables higher-level contextual information to be effectively incorporated into a background LM improving recognition or translation performance. One approach is to employ Latent Semantic Analysis (LSA) to capture in-domain word unigram distributions which are then integrated into the background N-gram LM. This approach has been successfully applied in automatic speech recognition (ASR) (Tam and Schultz, 2006) using the Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The LDA model can be viewed as a Bayesian topic mixture model with the topic mixture weights drawn from a Dirichlet distribution. For LM adaptation, the topic mixture weights are estimated based on in-domain adaptation text (e.g. ASR hypotheses). The adapted mixture weights are then used to interpolate a topic-dependent unigram LM, which is finally integrated into the background N-gram LM using marginal adaptation (Kneser et al., 1997)

In this paper, we propose a framework to perform LM adaptation across languages, enabling the adaptation of a LM from one language based on the adaptation text of another language. In statistical machine translation (SMT), one approach is to apply LM adaptation on the target language based on an initial translation of input references (Kim and Khudanpur, 2003; Paulik et al., 2005). This scheme is limited by the coverage of the translation model, and overall by the quality of translation. Since this approach only allows to apply LM adaptation *after* translation, available knowledge cannot be applied to extend the coverage. We propose a bilingual LSA model (bLSA) for crosslingual LM adaptation that can be applied *before* translation. The bLSA model consists of two LSA models: one for each side of the language trained on parallel document corpora. The key property of the bLSA model is that

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 520–527,
Prague, Czech Republic, June 2007.

the latent topic of the source and target LSA models can be assumed to be a one-to-one correspondence and thus share a common latent topic space since the training corpora consist of bilingual parallel data. For instance, say topic 10 of the Chinese LSA model is about politics. Then topic 10 of the English LSA model is set to also correspond to politics and so forth. During LM adaptation, we first infer the topic mixture weights from the source text using the source LSA model. Then we transfer the inferred mixture weights to the target LSA model and thus obtain the target LSA marginals. The challenge is to enforce the one-to-one topic correspondence. Our proposal is to share common variational Dirichlet posteriors over the topic mixture weights of a document pair in the LDA-style model. The beauty of the bLSA framework is that the model searches for a common latent topic space in an unsupervised fashion, rather than to require manual interaction. Since the topic space is language independent, our approach supports topic transfer in multiple language pairs in O(N) where N is the number of languages.

Related work includes the Bilingual Topic Admixture Model (BiTAM) for word alignment proposed by (Zhao and Xing, 2006). Basically, the BiTAM model consists of topic-dependent translation lexicons modeling $Pr(c|e,k)$ where c, e and k denotes the source Chinese word, target English word and the topic index respectively. On the other hand, the bLSA framework models $Pr(c|k)$ and $Pr(e|k)$ which is different from the BiTAM model. By their different modeling nature, the bLSA model usually supports more topics than the BiTAM model. Another work by (Kim and Khudanpur, 2004) employed crosslingual LSA using singular value decomposition which concatenates bilingual documents into a single input supervector before projection.

We organize the paper as follows: In Section 2, we introduce the bLSA framework including Latent Dirichlet-Tree Allocation (LDTA) (Tam and Schultz, 2007) as a correlated LSA model, bLSA training and crosslingual LM adaptation. In Section 3, we present the effect of LM adaptation on word perplexity, followed by SMT experiments reported in BLEU on both speech and text input in Section 3.3. Section 4 describes conclusions and future works.

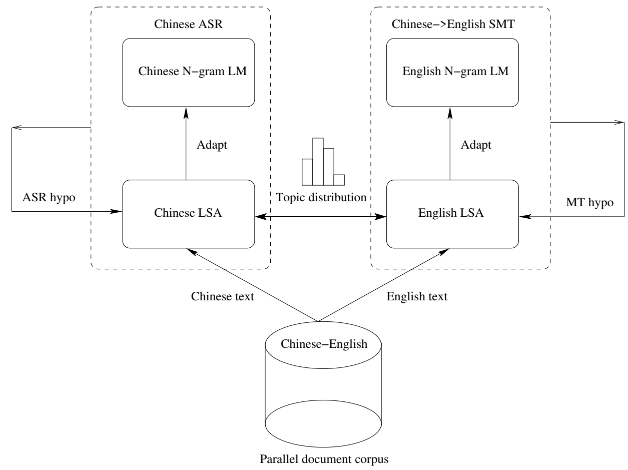

Figure 1: Topic transfer in bilingual LSA model.

2 Bilingual Latent Semantic Analysis

The goal of a bLSA model is to enforce a one-to-one topic correspondence between monolingual LSA models, each of which can be modeled using an LDA-style model. The role of the bLSA model is to transfer the inferred latent topic distribution from the source language to the target language assuming that the topic distributions on both sides are identical. The assumption is reasonable for parallel document pairs which are faithful translations. Figure 1 illustrates the idea of topic transfer between monolingual LSA models followed by LM adaptation. One observation is that the topic transfer can be bi-directional meaning that the "flow" of topic can be from ASR to SMT or vice versa. In this paper, we only focus on ASR-to-SMT direction. Our target is to minimize the word perplexity on the target language through LM adaptation. Before we introduce the heuristic of enforcing a one-to-one topic correspondence, we describe the Latent Dirichlet-Tree Allocation (LDTA) for LSA.

2.1 Latent Dirichlet-Tree Allocation

The LDTA model extends the LDA model in which correlation among latent topics are captured using a Dirichlet-Tree prior. Figure 2 illustrates a depth-two Dirichlet-Tree. A tree of depth one simply falls back to the LDA model. The LDTA model is a generative model with the following generative process:

1. Sample a vector of branch probabilities $b_j \sim$

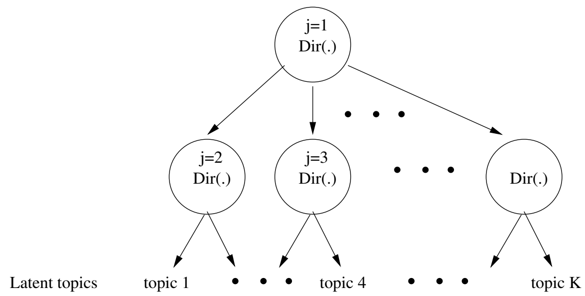

Figure 2: Dirichlet-Tree prior of depth two.

$Dir(\alpha_j)$ for each node $j = 1...J$ where α_j denotes the parameter (aka the pseudo-counts of its outgoing branches) of the Dirichlet distribution at node j.

2. Compute the topic proportions as:

$$\theta_k = \prod_{jc} b_{jc}^{\delta_{jc}(k)} \tag{1}$$

where $\delta_{jc}(k)$ is an indicator function which sets to unity when the c-th branch of the j-th node leads to the leaf node of topic k and zero otherwise. The k-th topic proportion θ_k is computed as the product of branch probabilities from the root node to the leaf node of topic k.

3. Generate a document using the topic multinomial for each word w_i:

$$\begin{aligned} z_i &\sim Mult(\theta) \\ w_i &\sim Mult(\beta_{.z_i}) \end{aligned}$$

where $\beta_{.z_i}$ denotes the topic-dependent unigram LM indexed by z_i.

The joint distribution of the latent variables (topic sequence z_1^n and the Dirichlet nodes over child branches b_j) and an observed document w_1^n can be written as follows:

$$p(w_1^n, z_1^n, b_1^J) = p(b_1^J|\{\alpha_j\}) \prod_i^n \beta_{w_i z_i} \cdot \theta_{z_i}$$

$$\text{where} \quad p(b_1^J|\{\alpha_j\}) = \prod_j^J Dir(b_j; \alpha_j) \propto \prod_{jc} b_{jc}^{\alpha_{jc}-1}$$

Similar to LDA training, we apply the variational Bayes approach by optimizing the lower bound of the marginalized document likelihood:

$$\begin{aligned} L(w_1^n; \Lambda, \Gamma) &= E_q[\log \frac{p(w_1^n, z_1^n, b_1^J; \Lambda)}{q(z_1^n, b_1^J; \Gamma)}] \\ &= E_q[\log p(w_1^n|z_1^n)] + E_q[\log \frac{p(z_1^n|b_1^J)}{q(z_1^n)}] \\ &\quad + E_q[\log \frac{p(b_1^J; \{\alpha_j\})}{q(b_1^J; \{\gamma_j\})}] \end{aligned}$$

where $q(z_1^n, b_1^J; \Gamma) = \prod_i^n q(z_i) \cdot \prod_j^J q(b_j)$ is a factorizable variational posterior distribution over the latent variables parameterized by Γ which are determined in the E-step. Λ is the model parameters for a Dirichlet-Tree $\{\alpha_j\}$ and the topic-dependent unigram LM $\{\beta_{wk}\}$. The LDTA model has an E-step similar to the LDA model:

E-Step:

$$\gamma_{jc} = \alpha_{jc} + \sum_i^n \sum_k^K q_{ik} \cdot \delta_{jc}(k) \tag{2}$$

$$q_{ik} \propto \beta_{w_i k} \cdot e^{E_q[\log \theta_k]} \tag{3}$$

where

$$\begin{aligned} E_q[\log \theta_k] &= \sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}] \\ &= \sum_{jc} \delta_{jc}(k) \left(\Psi(\gamma_{jc}) - \Psi(\sum_c \gamma_{jc}) \right) \end{aligned}$$

where q_{ik} denotes $q(z_i = k)$ meaning the variational topic posterior of word w_i. Eqn 2 and Eqn 3 are executed iteratively until convergence is reached.

M-Step:

$$\beta_{wk} \propto \sum_i^n q_{ik} \cdot \delta(w_i, w) \tag{4}$$

where $\delta(w_i, w)$ is a Kronecker Delta function. The alpha parameters can be estimated with iterative methods such as Newton-Raphson or simple gradient ascent procedure.

2.2 Bilingual LSA training

For the following explanations, we assume that our source and target languages are Chinese and English respectively. The bLSA model training is a two-stage procedure. At the first stage, we train a Chinese LSA model using the Chinese documents in parallel corpora. We applied the variational EM algorithm (Eqn 2–4) to train a Chinese LSA model. Then we used the model to compute the term $e^{E_q[\log \theta_k]}$ needed in Eqn 3 for each Chinese document in parallel corpora. At the second stage, we apply the same $e^{E_q[\log \theta_k]}$ to *bootstrap* an English LSA model, which is the key to enforce a one-to-one topic correspondence. Now the hyper-parameters of the variational Dirichlet posteriors of each node in the Dirichlet-Tree are shared among the Chinese and English model. Precisely, we apply only Eqn 3 with fixed $e^{E_q[\log \theta_k]}$ in the E-step and Eqn 4 in the M-step on $\{\beta_{wk}\}$ to bootstrap an English LSA model. Notice that the E-step is non-iterative resulting in rapid LSA training. In short, given a monolingual LSA model, we can rapidly bootstrap LSA models of new languages using parallel document corpora. Notice that the English and Chinese vocabulary sizes do not need to be similar. In our setup, the Chinese vocabulary comes from the ASR system while the English vocabulary comes from the English part of the parallel corpora. Since the topic transfer can be bi-directional, we can perform the bLSA training in a reverse manner, i.e. training an English LSA model followed by bootstrapping a Chinese LSA model.

2.3 Crosslingual LM adaptation

Given a source text, we apply the E-step to estimate variational Dirichlet posterior of each node in the Dirichlet-Tree. We estimate the topic weights on the source language using the following equation:

$$\hat{\theta}_k^{(CH)} \propto \prod_{jc} \left(\frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \tag{5}$$

Then we apply the topic weights into the target LSA model to obtain an in-domain LSA marginals:

$$Pr_{EN}(w) = \sum_{k=1}^{K} \beta_{wk}^{(EN)} \cdot \hat{\theta}_k^{(CH)} \tag{6}$$

We integrate the LSA marginal into the target background LM using marginal adaptation (Kneser et al., 1997) which minimizes the Kullback-Leibler divergence between the adapted LM and the background LM:

$$Pr_a(w|h) \propto \left(\frac{Pr_{ldta}(w)}{Pr_{bg}(w)} \right)^{\beta} \cdot Pr_{bg}(w|h) \tag{7}$$

Likewise, LM adaptation can take place on the source language as well due to the bi-directional nature of the bLSA framework when target-side adaptation text is available. In this paper, we focus on LM adaptation on the target language for SMT.

3 Experimental Setup

We evaluated our bLSA model using the Chinese–English parallel document corpora consisting of the Xinhua news, Hong Kong news and Sina news. The combined corpora contains 67k parallel documents with 35M Chinese (CH) words and 43M English (EN) words. Our spoken language translation system translates from Chinese to English. The Chinese vocabulary comes from the ASR decoder while the English vocabulary is derived from the English portion of the parallel training corpora. The vocabulary sizes for Chinese and English are 108k and 69k respectively. Our background English LM is a 4-gram LM trained with the modified Kneser-Ney smoothing scheme using the SRILM toolkit on the same training text. We explore the bLSA training in both directions: EN→CH and CH→EN meaning that an English LSA model is trained first and a Chinese LSA model is bootstrapped or vice versa. Experiments explore which bootstrapping direction yield best results measured in terms of English word perplexity. The number of latent topics is set to 200 and a balanced binary Dirichlet-Tree prior is used.

With an increasing interest in the ASR-SMT coupling for spoken language translation, we also evaluated our approach with Chinese ASR hypotheses and compared with Chinese manual transcriptions. We are interested to see the impact due to recognition errors on the ASR hypotheses compared to the manual transcriptions. We employed the CMU-InterACT ASR system developed for the GALE 2006 evaluation. We trained acoustic models with over 500 hours of quickly transcribed speech data released by the GALE program and the LM with over 800M-word Chinese corpora. The character error rates on the CCTV, RFA and NTDTV shows in the RT04 test set are 7.4%, 25.5% and 13.1% respectively.

Topic index	Top words
“CH-40” “EN-40”	flying, submarine, aircraft, air, pilot, land, mission, brand-new air, sea, submarine, aircraft, flight, flying, ship, test
“CH-41” “EN-41”	satellite, han-tian, launch, space, china, technology, astronomy space, satellite, china, technology, satellites, science
“CH-42” “EN-42”	fire, airport, services, marine, accident, air fire, airport, services, department, marine, air, service

Table 1: Parallel topics extracted by the bLSA model. Top words on the Chinese side are translated into English for illustration purpose.

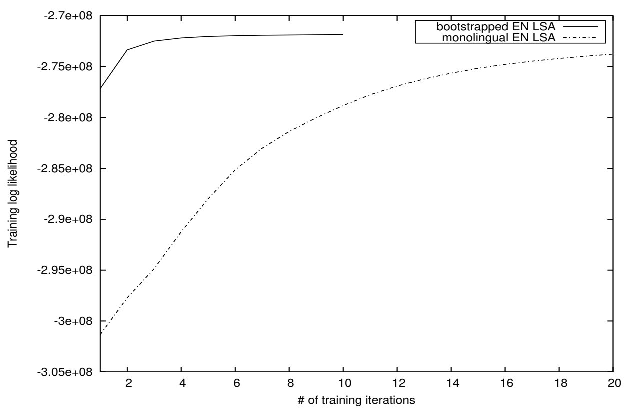

Figure 3: Comparison of training log likelihood of English LSA models bootstrapped from a Chinese LSA and from a flat monolingual English LSA.

3.1 Analysis of the bLSA model

By examining the top-words of the extracted parallel topics, we verify the validity of the heuristic described in Section 2.2 which enforces a one-to-one topic correspondence in the bLSA model. Table 1 shows the latent topics extracted by the CH→EN bLSA model. We can see that the Chinese-English topic words have strong correlations. Many of them are actually translation pairs with similar word rankings. From this viewpoint, we can interpret bLSA as a crosslingual word trigger model. The result indicates that our heuristic is effective to extract parallel latent topics. As a sanity check, we also examine the likelihood of the training data when an English LSA model is bootstrapped. We can see from Figure 3 that the likelihood increases monotonically with the number of training iterations. The figure also shows that by sharing the variational Dirichlet posteriors from the Chinese LSA model, we can bootstrap an English LSA model *rapidly* compared to monolingual English LSA training with both training procedures started from the same flat model.

LM (43M)	CCTV	RFA	NTDTV
BG EN unigram	1065	1220	1549
+CH→EN (CH ref)	755	880	1113
+EN→CH (CH ref)	762	896	1111
+CH→EN (CH hypo)	757	885	1126
+EN→CH (CH hypo)	766	896	1129
+CH→EN (EN ref)	731	838	1075
+EN→CH (EN ref)	747	848	1087

Table 2: English word perplexity (PPL) on the RT04 test set using a unigram LM.

3.2 LM adaptation results

We trained the bLSA models on both CH→EN and EN→CH directions and compared their LM adaptation performance using the Chinese ASR hypotheses (hypo) and the manual transcriptions (ref) as input. We adapted the English background LM using the LSA marginals described in Section 2.3 for each show on the test set.

We first evaluated the English word perplexity using the EN unigram LM generated by the bLSA model. Table 2 shows that the bLSA-based LM adaptation reduces the word perplexity by over 27% relative compared to an unadapted EN unigram LM. The results indicate that the bLSA model successfully leverages the text from the source language and improves the word perplexity on the target language. We observe that there is almost no performance difference when either the ASR hypotheses or the manual transcriptions are used for adaptation. The result is encouraging since the bLSA model may be insensitive to moderate recognition errors through the projection of the input adaptation text into the latent topic space. We also apply an English translation reference for adaptation to show an oracle performance. The results using the Chinese hypotheses are not too far off from the oracle performance. Another observation is that the CH→EN bLSA model seems to give better performance than the EN→CH bLSA model. However, their differences are not significant. The result may imply that the direction of the bLSA training is not important since the latent topic space captured by either language is similar when parallel training corpora are used. Table 3 shows the word perplexity when the background 4-gram English LM is adapted with the tuning parameter β set

LM (43M, $\beta = 0.7$)	CCTV	RFA	NTDTV
BG EN 4-gram	118	212	203
+CH→EN (CH ref)	102	191	179
+EN→CH (CH ref)	102	198	179
+CH→EN (CH hypo)	102	193	180
+EN→CH (CH hypo)	103	198	180
+CH→EN (EN ref)	100	186	176
+EN→CH (EN ref)	101	190	176

Table 3: English word perplexity (PPL) on the RT04 test set using a 4-gram LM.

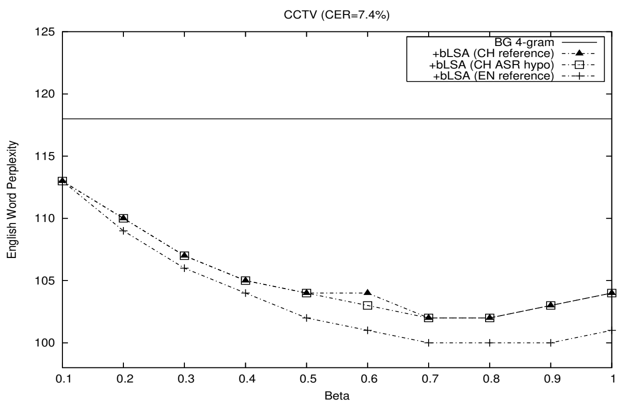

Figure 4: Word perplexity with different β using manual reference or ASR hypotheses on CCTV.

to 0.7. Figure 4 shows the change of perplexity with different β. We see that the adaptation performance using the ASR hypotheses or the manual transcriptions are almost identical on different β with an optimal value at around 0.7. The results show that the proposed approach successfully reduces the perplexity in the range of 9–13.6% relative compared to an unadapted baseline on different shows when ASR hypotheses are used. Moreover, we observe similar performance using ASR hypotheses or manual Chinese transcriptions which is consistent with the results on Table 2. On the other hand, it is interesting to see that the performance gap from the oracle adaptation is somewhat related to the degree of mismatch between the test show and the training condition. The gap looks wider on the RFA and NTDTV shows compared to the CCTV show.

3.3 Incorporating bLSA into Spoken Language Translation

To investigate the effectiveness of bLSA LM adaptation for spoken language translation, we incorporated the proposed approach into our state-of-the-art phrase-based SMT system. Translation performance was evaluated on the RT04 broadcast news evaluation set when applied to both the manual transcriptions and 1-best ASR hypotheses. During evaluation two performance metrics, BLEU (Papineni et al., 2002) and NIST, were computed. In both cases, a single English reference was used during scoring. In the transcription case the original English references were used. For the ASR case, as utterance segmentation was performed automatically, the number of sentences generated by ASR and SMT differed from the number of English references. In this case, Levenshtein alignment was used to align the translation output to the English references before scoring.

3.4 Baseline SMT Setup

The baseline SMT system consisted of a non adaptive system trained using the same Chinese-English parallel document corpora used in the previous experiments (Sections 3.1 and 3.2). For phrase extraction a cleaned subset of these corpora, consisting of 1M Chinese-English sentence pairs, was used. SMT decoding parameters were optimized using manual transcriptions and translations of 272 utterances from the RT04 development set (LDC2006E10).

SMT translation was performed in two stages using an approach similar to that in (Vogel, 2003). First, a translation lattice was constructed by matching all possible bilingual phrase-pairs, extracted from the training corpora, to the input sentence. Phrase extraction was performed using the "*PESA*" (Phrase Pair Extraction as Sentence Splitting) approach described in (Vogel, 2005). Next, a search was performed to find the best path through the lattice, i.e. that with maximum *translation-score*. During search reordering was allowed on the target language side. The final translation result was that hypothesis with maximum *translation-score*, which is a log-linear combination of 10 scores consisting of Target LM probability, Distortion Penalty, Word-Count Penalty, Phrase-Count and six Phrase-Alignment scores. Weights for each component score were optimized to maximize BLEU-score on the development set using MER optimization as described in (Venugopal et al., 2005).

SMT Target LM	Translation Quality - BLEU (NIST) CCTV	RFA	NTDTV	ALL
	Manual Transcription			
Baseline LM:	0.162 (5.212)	0.087 (3.854)	0.140 (4.859)	0.132 (5.146)
bLSA (bLSA-Adapted LM):	0.164 (5.212)	0.087 (3.897)	0.143 (4.864)	0.134 (5.162)
	1-best ASR Output			
CER (%)	7.4	25.5	13.1	14.9
Baseline LM:	0.129 (4.15)	0.051 (2.77)	0.086 (3.50)	0.095 (3.90)
bLSA (bLSA-Adapted LM):	0.132 (4.16)	0.050 (2.79)	0.089 (3.53)	0.096 (3.91)

Table 4: Translation performance of baseline and bLSA-Adapted Chinese-English SMT systems on manual transcriptions and 1-best ASR hypotheses

3.5 Performance of Baseline SMT System

First, the baseline system performance was evaluated by applying the system described above to the reference transcriptions and 1-best ASR hypotheses generated by our Mandarin speech recognition system. The translation accuracy in terms of BLEU and NIST for each individual show ("*CCTV*", "*RFA*", and "*NTDTV*"), and for the complete test-set, are shown in Table 4 (**Baseline LM**). When applied to the reference transcriptions an overall BLEU score of 0.132 was obtained. BLEU-scores ranged between 0.087 and 0.162 for the "*RFA*", "*NTDTV*" and "*CCTV*" shows, respectively. As the "*RFA*" show contained a large segment of conversational speech, translation quality was considerably lower for this show due to genre mismatch with the training corpora of newspaper text.

For the 1-best ASR hypotheses, an overall BLEU score of 0.095 was achieved. For the ASR case, the relative reduction in BLEU scores for the RFA and NTDTV shows is large, due to the significantly lower recognition accuracies for these shows. BLEU score is also degraded due to poor alignment of references during scoring.

3.6 Incorporation of bLSA Adaptation

Next, the effectiveness of bLSA based LM adaptation was evaluated. For each show the target English LM was adapted using bLSA-adaptation, as described in Section 2.3. SMT was then applied using an identical setup to that used in the baseline experiments.

The translation accuracy when bLSA adaptation was incorporated is shown in Table 4. When applied to the manual transcriptions, bLSA adaptation improved the overall BLEU-score by 1.7% relative (from 0.132 to 0.134). For all three shows bLSA adaptation gained higher BLEU and NIST metrics. A similar trend was also observed when the proposed approach was applied to the 1-best ASR output. On the evaluation set a relative improvement in BLEU score of 1.0% was gained.

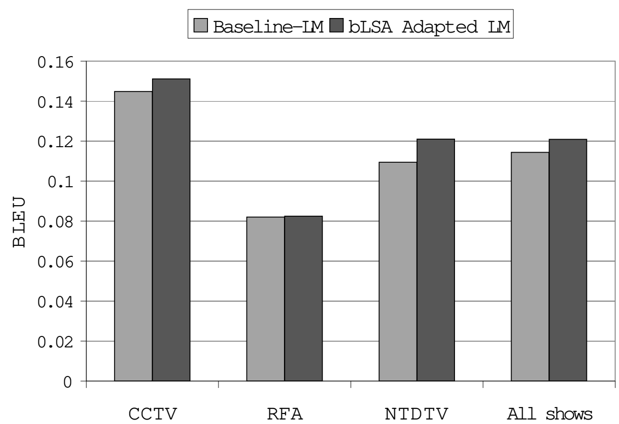

Figure 5: BLEU score for those 25% utterances which resulted in different translations after bLSA adaptation (manual transcriptions)

The semantic interpretation of the majority of utterances in broadcast news are not affected by topic context. In the experimental evaluation it was observed that only 25% of utterances produced different translation output when bLSA adaptation was performed compared to the topic-independent baseline. Although the improvement in translation quality (BLEU) was small when evaluated over the entire test set, the improvement in BLEU score for these 25% utterances was significant. The translation quality for the baseline and bLSA-adaptive system when evaluated only on these utterances is shown in Figure 5 for the manual transcription case. On this subset of utterances an overall improvement in BLEU of 0.007 (5.7% relative) was gained, with a gain of 0.012 (10.6% relative) points for the "*NTDTV*" show. A similar trend was observed when applied to the 1-best ASR output. In this case a relative improvement in BLEU of 12.6% was gained for "*NTDTV*", and for "All shows" 0.007 (3.7%) was gained. Current evaluation metrics for translation, such as "BLEU", do not consider the relative importance of specific words or phrases during translation and thus are unable to highlight the true effectiveness of the proposed approach. In future work, we intend to investigate other evaluation metrics which consider the relative informational content of words.

4 Conclusions

We proposed a bilingual latent semantic model for crosslingual LM adaptation in spoken language translation. The bLSA model consists of a set of monolingual LSA models in which a one-to-one topic correspondence is enforced between the LSA models through the sharing of variational Dirichlet posteriors. Bootstrapping a LSA model for a new language can be performed rapidly with topic transfer from a well-trained LSA model of another language. We transfer the inferred topic distribution from the input source text to the target language effectively to obtain an in-domain target LSA marginals for LM adaptation. Results showed that our approach significantly reduces the word perplexity on the target language in both cases using ASR hypotheses and manual transcripts. Interestingly, the adaptation performance is not much affected when ASR hypotheses were used. We evaluated the adapted LM on SMT and found that the evaluation metrics are crucial to reflect the actual improvement in performance. Future directions include the exploration of story-dependent LM adaptation with automatic story segmentation instead of show-dependent adaptation due to the possibility of multiple stories within a show. We will investigate the incorporation of monolingual documents for potentially better bilingual LSA modeling.

Acknowledgment

This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-2-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, pages 1107–1135.

W. Kim and S. Khudanpur. 2003. LM adaptation using cross-lingual information. In *Proc. of Eurospeech*.

W. Kim and S. Khudanpur. 2004. Cross-lingual latent semantic analysis for LM. In *Proc. of ICASSP*.

R. Kneser, J. Peters, and D. Klakow. 1997. Language model adaptation using dynamic marginals. In *Proc. of Eurospeech*, pages 1971–1974.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*.

M. Paulik, C. Fügen, T. Schaaf, T. Schultz, S. Stüker, and A. Waibel. 2005. Document driven machine translation enhanced automatic speech recognition. In *Proc. of Interspeech*.

Y. C. Tam and T. Schultz. 2006. Unsupervised language model adaptation using latent semantic marginals. In *Proc. of Interspeech*.

Y. C. Tam and T. Schultz. 2007. Correlated latent semantic model for unsupervised language model adaptation. In *Proc. of ICASSP*.

A. Venugopal, A. Zollmann, and A. Waibel. 2005. Training and evaluation error minimization rules for statistical machine translation. In *Proc. of ACL*.

S. Vogel. 2003. SMT decoder dissected: Word reordering. In *Proc. of ICNLPKE*.

S. Vogel. 2005. PESA: Phrase pair extraction as sentence splitting. In *Proc. of the Machine Translation Summit*.

B. Zhao and E. P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proc. of ACL*.