Bilingual LSA-based Translation Lexicon Adaptation for Spoken Language Translation

Yik-Cheung Tam and Tanja Schultz

InterACT, Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213
{yct,tanja}@cs.cmu.edu

Abstract

We present a bilingual LSA (bLSA) framework for translation lexicon adaptation. The idea is to apply marginal adaptation on a translation lexicon so that the lexicon marginals match to in-domain marginals. In the framework of speech translation, the bLSA method transfers topic distributions from the source to the target side, such that the translation lexicon can be adapted before translation based on the source document. We evaluated the proposed approach on our Mandarin RT04 spoken language translation system. Results showed that the conditional likelihood on the test sentence pairs is improved significantly using an adapted translation lexicon compared to an unadapted baseline. The proposed approach showed improvement on BLEU-score in SMT. When both the target-side LM and the translation lexicon were adapted and applied simultaneously for SMT decoding, the gain on BLEU-score was more than additive compared to the scenarios when the adapted models were individually applied.

Index Terms: Bilingual LSA, Marginal adaptation, Translation lexicon, LM

1. Introduction

One challenge in large-domain statistical machine translation (SMT) is to adapt the SMT models (e.g. translation lexicon, language model) to the topic of the test set. One research direction is to extract the topic information of a source text and then transfer the information to the target side for adaptation. This approach has two advantages: firstly, it can be applied before translation, and thus has immediate impact on the translation output. Second, it does not rely on translation output, and therefore does not suffer from any translation errors. Recently, we have proposed a bilingual latent semantic analysis (bLSA [1]) framework for language model (LM) adaptation on the target side before translation. The idea is to train bilingual LSA models, one for the source, another for the target side. The training is constrained by enforcing a one-to-one topic correspondence between the source and target LSA model. For instance, say topic 10 of the source Chinese LSA model is about politics. Then topic 10 of the target English LSA model is also set to to politics. Before translation, we infer a topic distribution of the source text using source LSA model. Then we transfer the inferred distribution to target LSA model and thus obtain the target LSA marginals. The target LM is then adapted using marginal adaptation [2]. We achieved significant reduction in word perplexity on the target side compared to an unadapted baseline, and also showed improvements in SMT performance [1].

We extend the bLSA framework to the adaptation of translation lexicon. Our goal is to minimize the Kullback-Leibler divergence between an adapted lexicon $p_a(c|e)$ and a background lexicon $p_{bg}(c|e)$ subject to constraints that the lexicon marginals $p_{lex}(c)$ are matched to in-domain marginals $p_a(c)$ where c and e denotes a source Chinese word and a target English word respectively. Lexicon marginals $p_{lex}(c)$ can be computed by marginalization: $p_{lex}(c) = \sum_e p_a(c|e)p_a(e)$. With this problem formulation, the adapted lexicon can be estimated using generalized iterative scaling which is commonly applied in maximum entropy modeling. In this paper, we report results using the manual source transcription for adaptation.

Related work includes the Bilingual Topic Admixture Model (BiTAM) for word alignment [3]. Basically, the BiTAM model consists of topic-dependent translation lexicon modeling $p(c|e,k)$ 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 $p(c|k)$ and $p(e|k)$ which is different from the BiTAM model. By their different modeling nature, the bLSA model usually supports more topics than the BiTAM model.

We organize the paper as follows: In Section 2, we review the bLSA framework including the Latent Dirichlet-Tree Allocation (LDTA [4]) as a correlated LSA model and bLSA training. In Section 3, we present the lexicon adaptation approach. In Section 4, we report adaptation experiments followed by conclusions and future works in Section 5.

2. Bilingual Latent Semantic Analysis

bLSA can be viewed as a "meta" model which consists of two monolingually-trained LSA models, each of which can be modeled using a LDA-style model [5]. The goal of bLSA is to enforce a one-to-one topic correspondence between monolingual LSA models via sharing a language-independent latent topic space. Transferring the inferred latent topic distribution from the source side to the target side could be performed under the assumption that the topic distributions on both sides are identical. In the translation framework this is a very reasonable assumption since the topic distributions on both sides of a parallel bilingual corpus are identical by definition. Figure 1 illustrates the idea of topic transfer between monolingual LSA models followed by LM adaptation and translation lexicon adaptation. Our

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.

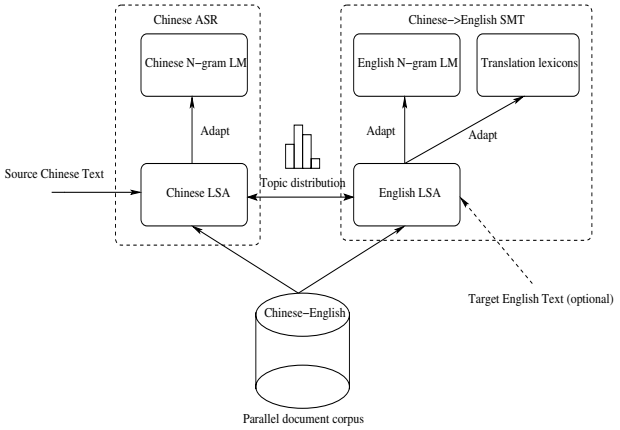

Figure 1: Bilingual LSA framework for spoken language translation via topic transfer.

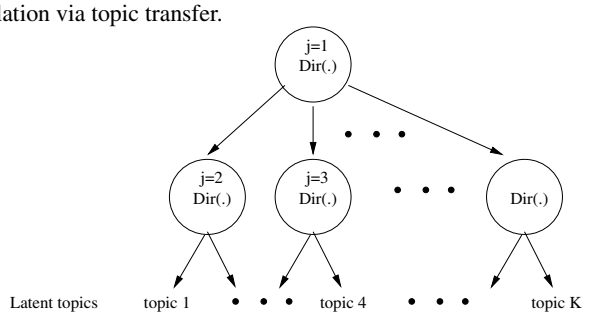

Figure 2: Dirichlet-Tree prior of depth two.

target is to increase the conditional likelihood $p(C|E)$ of a test sentence pair $< C = c_1^I, E = e_1^J >$ where C and E denote a Chinese and English sentence respectively after the lexicon is adapted. In the following, we first review the Latent Dirichlet-Tree Allocation (LDTA [4]) for correlated LSA and describe the bLSA training.

2.1. Review of 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 $b_j \sim Dir(\alpha_j)$ 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:

$$\theta_k = \prod_{jm} b_{jm}^{\delta_{jm}(k)} \quad (1)$$

where $\delta_{jm}(k)$ is an indicator function which sets to unity when the m-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:

$$z_i \sim Mult(\theta)$$
$$w_i \sim Mult(\beta_{.z_i})$$

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 observed document w_1^n can be written as follows:

$$p(w_1^n, z_1^n, b_1^J) = p(b_1^J|\{\alpha_j\}) \prod_i^n \beta_{w_i z_i} \cdot \theta_{z_i}$$
$$\text{where} \quad p(b_1^J|\{\alpha_j\}) = \prod_j^J Dir(b_j; \alpha_j)$$

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

$$L(w_1^n; \Lambda, \Gamma) = E_q[\log \frac{p(w_1^n, z_1^n, b_1^J; \Lambda)}{q(z_1^n, b_1^J; \Gamma)}]$$

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:

$$\gamma_{jm} = \alpha_{jm} + \sum_i^n \sum_k^K q_{ik} \cdot \delta_{jm}(k) \quad (2)$$
$$q_{ik} \propto \beta_{w_i k} \cdot e^{E_q[\log \theta_k]} \quad (3)$$

where

$$E_q[\log \theta_k] = \sum_{jm} \delta_{jm}(k) E_q[\log b_{jm}]$$
$$= \sum_{jm} \delta_{jm}(k) \left(\Psi(\gamma_{jm}) - \Psi(\sum_m \gamma_{jm}) \right)$$

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

M-Step:

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

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 explanation convenience, we assume that our source and target languages are Chinese (CH) and English (EN) respectively. The bLSA model training is a two-stage procedure. In the first stage, we train a Chinese LSA model using the Chinese documents in parallel corpora. We apply the variational EM algorithm (Eqn 2–4) to train a Chinese LSA model. Then we use the model to compute the term $e^{E_q[\log \theta_k]}$ needed in Eqn 3 for each Chinese document in parallel corpora. In 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 index	Top words
"CH-40"	flying, submarine, aircraft, air, pilot, land, mission, brand-new
"EN-40"	air, sea, submarine, aircraft, flight, flying, ship, test
"CH-41"	satellite, han-tian, launch, space, china, technology, astronomy
"EN-41"	space, satellite, china, technology, satellites, science
"CH-42"	fire, airport, services, marine, accident, air
"EN-42"	fire, airport, services, department, marine, air, service

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

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. Table 1 shows sample correlated topics extracted by bLSA using a parallel document corpora.

2.3. Estimating LSA marginals

Given a source text d_{ch}, 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:

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

Then we apply the topic weights into the source and target LSA models to obtain an adapted LSA marginals:

$$p_{lsa}(e|d_{ch}) \quad = \quad \sum_{k=1}^{K} \beta_{ek}^{(EN)} \cdot \hat{\theta}_k^{(CH)} \tag{6}$$

For crosslingual LM adaptation [1], we apply marginal adaptation which minimizes the Kullback-Leibler divergence between the adapted LM $p_a(e|h)$ and the background LM $p_{bg}(e|h)$ on the target language:

$$p_a(e|h) \quad \propto \quad \left(\frac{p_{lsa}(e|d_{ch})}{p_{bg}(e)} \right)^{\beta} \cdot p_{bg}(e|h) \tag{7}$$

where $p_{bg}(e)$ denotes the background unigram marginal on the target language and the tuning factor β is set to 0.7.

3. Translation lexicon adaptation

We adapt the translation lexicon by marginal adaptation. The goal is to search for an adapted translation lexicon $p_a(c|e)$ such that its KL divergence from the background lexicon $p_{bg}(c|e)$ is minimized such that the model marginals is equal to in-domain marginals:

$$p_a(.|.) \quad = \quad argmin_{p(.|.)} \sum_e p_a(e) \cdot KL(p(.|e)||p_{bg}(.|e))$$

such that

$$\forall c : \sum_e p_a(e) \cdot p_a(c|e) = p_a(c|d_{ch})$$

By using the Lagrange multipliers and taking the derivative of the cost function with respect to $p_a(c = j|e)$, we get:

$$p_a(c = j|e) \quad \propto \quad p_{bg}(c = j|e) \cdot \exp \lambda_j$$

It is known that the optimization problem is similar to the maximum entropy settings except that we start with a non-uniform background distribution $p_{bg}(c|e)$. Therefore, we can solve λ_j using generalized iterative scaling (GIS):

$$\lambda_j^{(new)} \quad = \quad \lambda_j^{(old)} + \log \frac{p_a(c = j|d_{ch})}{\sum_e p_a(e) \cdot p_a^{(old)}(c = j|e)} \tag{8}$$

where $p_a(e)$ is approximated by an English LSA marginal $p_{lsa}(e|d_{ch})$ from Eqn 6. Notice that the range of e is limited by the entries in the translation lexicon, which means that computing the denominator is usually very fast without evaluating all possible e. We estimate $p_a(c = j|d_{ch})$ using the smoothed relative word frequency on the source document with Good-Turing discounting scheme.

Since the translation lexicon is optimized such that the lexicon marginals and the source-side marginals are matched, we can check that their KL divergence $KL(p_{lex}(.)||p_{lsa}(.))$ is expected to decrease monotonically for each GIS iteration. If an accurate in-domain marginal is provided, the conditional likelihood $p(C|E)$ is expected to improve compared to an unadapted lexicon where $C = c_1^I$ and $E = e_1^J$ denote a pair of Chinese and English sentences respectively. We compute the likelihood using the IBM model 1 as follows:

$$p(C|E) \quad = \quad \sum_A p(C|A, E) \cdot p(A|E) \quad \propto \prod_{i=1}^{I} \sum_{j=0}^{J} p(c_i|e_j)$$

where e_0 denotes a NULL word on an English sentence. We ignore computing the term $p(A|E)$ for model comparison since it is a constant with respect to adapted/baseline lexicons. We can also use a metric similar to "word perplexity" to measure the averaged number of fanouts of a target word:

$$PPL(C|E) \quad = \quad \exp(-\frac{1}{I} \log p(C|E))$$

4. 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 of 1M sentence pairs with 57M Chinese (CH) characters and 43M English (EN) words. Our spoken language translation system translates from Chinese to English. The sizes of Chinese and English vocabulary are 88k and 121k respectively which were derived from the baseline translation word lexicons. 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. The number of latent topics is set to 200 and a balanced binary Dirichlet-Tree prior is used.

The baseline SMT system was trained using the same parallel document corpora. For phrase extraction a cleaned subset of these corpora 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 [6]. 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 [7]. Next, a search was performed to find the best path through the lattice, i.e. that with maximum *translation-score*.

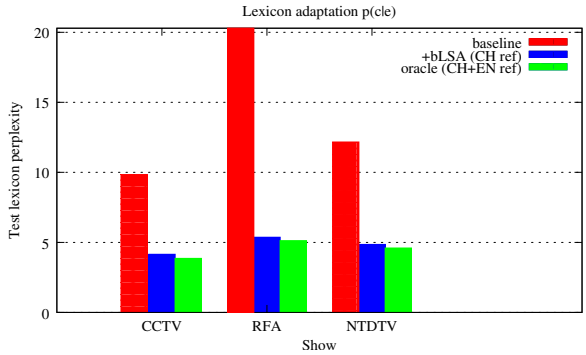

Figure 3: Lexicon "perplexity" $PPL(C|E)$ on RT04-eval set.

BLEU	CCTV	RFA	NTDTV	Overall
baseline	0.162	0.087	0.140	0.132
+lex	0.164	**0.092**	0.139	0.133
+lm	0.164	0.087	**0.144**	0.134
+lex+lm	**0.168**	0.088	0.143	**0.136**

Table 2: Translation performance (BLEU score) of the baseline and bLSA-adapted models on manual source transcriptions. "lex"/"lm" means only the lexicon/LM is adapted.

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 (λ_{LM} (Target LM), λ_{DM} (Distortion Penalty), λ_{WC} (Word Count Penalty), λ_{PC} (Phrase Count) , $\lambda_{PA_1}, \ldots, \lambda_{PA_6}$ (Phrase-Alignment Scores)) were optimized to maximize BLEU-score on the development set using MER optimization as described in [8]. No model adaptation was performed during MER training and the same MER-trained weights were applied for the adapted models.

4.1. Lexicon perplexity results

We first adapt the translation lexicon $p(e|c)$ using the source reference for each show with one GIS iteration. Then we compute the lexicon "perplexity" on the source reference C given the target reference E. Figure 3 shows that the adapted lexicon reduces the lexicon "perplexity" significantly in the range of 58–74% relative compared to an unadapted baseline. We also compute the oracle adaptation using the target English reference to estimate $p_a(e)$ in Eqn 8. Interestingly, the lexicon "perplexity" of the oracle and bLSA are very similar, indicating that the bLSA-derived English marginal is a reasonable estimate compared to the oracle marginal estimated by the target English reference.

4.2. Translation results

We evaluated the effectiveness of the bLSA adaptation on the target English LM and the translation lexicon in different stages. We adapted the LM and the translation lexicon and applied the adapted models separately and simultaneously for decoding. Table 2 shows the translation performance based on BLEU-score using the manual source reference as input. Results show that adapting the translation lexicon improves the BLEU-score by 0.2% and 0.5% on CCTV and RFA respectively compared to the baseline. However, we observed 0.1% degradation on NTDTV. On the other hand, adapting the target English LM improves the BLEU-score by 0.2% and 0.4% on CCTV and NTDTV respectively, but no improvement on RFA. When we apply both adapted models together, we achieved an extra improvement on BLEU-score by 0.4% on CCTV compared to each individually-adapted model. However, extra improvement does not hold on RFA and NTDTV which may be explained by the fact that only one of the adapted models helps but not both on both shows. Overall, we gained absolute 0.4% improvement on BLEU-score compared to the baseline, and the combined gain from each adapted model was more than additive on CCTV and the overall case.

5. Conclusions

We extended the bilingual latent semantic model for translation lexicon adaptation in SMT. bLSA 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. Given a source document, bLSA estimates the target-side LSA marginals which can be used for LM and translation lexicon adaptation via marginal adaptation before translation. Results showed that the proposed approach reduces the lexicon "perplexity" significantly compared to an unadapted baseline, and improves the SMT performance. When the adapted target-side LM and translation lexicon were applied simultaneously for SMT decoding, we achieved additional improvement in overall BLEU-score compared to scenarios in which adapted models were individually applied. The gain was more than additive which may imply a certain level of synergy among the adapted LM and translation lexicon. Future work includes the incorporation of monolingual documents and evaluation using ASR hypotheses as input.

6. Acknowledgement

We are very grateful to Ian Lane for building a SMT baseline system and his insightful discussion.

7. References

[1] Y. C. Tam, I. Lane, and T. Schultz, "Bilingual LSA-based LM adaptation for spoken language translation," in *Proc. of ACL*, 2007.

[2] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. of Eurospeech*, 1997, pp. 1971–1974.

[3] B. Zhao and E. P. Xing, "BiTAM: Bilingual topic admixture models for word alignment," in *Proc. of ACL*, 2006.

[4] Y. C. Tam and T. Schultz, "Correlated latent semantic model for unsupervised language model adaptation," in *Proc. of ICASSP*, 2007.

[5] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," in *Journal of Machine Learning Research*, 2003, pp. 1107–1135.

[6] S. Vogel, "SMT decoder dissected: Word reordering," in *Proc. of ICNLPKE*, 2003.

[7] S. Vogel, "PESA: Phrase pair extraction as sentence splitting," in *Proc. of the Machine Translation Summit*, 2005.

[8] A. Venugopal, A. Zollman, and A. Waibel, "Training and evaluation error minimization rules for statistical machine translation," in *Proc. of ACL*, 2005.