Mach Translat (2007) 21:187–207
DOI 10.1007/s10590-008-9045-2

Bilingual LSA-based adaptation for statistical machine translation

Yik-Cheung Tam · Ian Lane · Tanja Schultz

Received: 27 March 2008 / Accepted: 31 October 2008 / Published online: 19 November 2008

Abstract We propose a novel approach to cross-lingual language model and translation lexicon adaptation for statistical machine translation (SMT) based on bilingual latent semantic analysis. Bilingual LSA enables latent topic distributions to be efficiently transferred across languages by enforcing a one-to-one topic correspondence during training. Using the proposed bilingual LSA framework, model adaptation can be performed by, first, inferring the topic posterior distribution of the source text and then applying the inferred distribution to an n-gram language model of the target language and translation lexicon via marginal adaptation. The background phrase table is enhanced with the additional phrase scores computed using the adapted translation lexicon. The proposed framework also features rapid bootstrapping of LSA models for new languages based on a source LSA model of another language. Our approach is evaluated on the Chinese–English MT06 test set using the medium-scale SMT system and the GALE SMT system measured in BLEU and NIST scores. Improvement in both scores is observed on both systems when the adapted language model and the adapted translation lexicon are applied individually. When the adapted language model and the adapted translation lexicon are applied simultaneously, the gain is additive. At the 95% confidence interval of the unadapted baseline system, the gain in both scores is statistically significant using the medium-scale SMT system, while the gain in the NIST score is statistically significant using the GALE SMT system.

Y.-C. Tam (✉) · I. Lane · T. Schultz
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
e-mail: yct@cs.cmu.edu

I. Lane
e-mail: ian.lane@cs.cmu.edu

T. Schultz
e-mail: tanja@cs.cmu.edu

Keywords Bilingual latent semantic analysis · Latent Dirichlet-tree allocation · Cross-lingual language model adaptation · Lexicon adaptation · Topic distribution transfer · Statistical machine translation

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 language model improving recognition or translation performance. Exploiting topical context for monolingual language model adaptation is well studied in automatic speech recognition. Previous approaches include topic mixtures (Iyer and Ostendorf 1996), latent semantic analysis (LSA) using singular value decomposition (SVD) (Bellegarda 2000), probabilistic LSA using EM (Hofmann 1999), latent Dirichlet allocation (LDA) using variational Bayes (Blei et al. 2003) and HMM-LDA (Griffiths et al. 2004; Hsu and Glass 2006). Language model adaptation using LDA has been shown to be effective in reducing speech recognition errors (Tam and Schultz 2005, 2006; Mrva and Woodland 2006). This success has motivated applying the same monolingual language model adaptation approach on the target language in statistical machine translation (SMT) based on an initial translation of the input text (Kim and Khudanpur 2003; Paulik et al. 2005). However, this scheme may depend on the quality of the initial translation and the sensitivity of the language model adaptation approach towards translation errors. Also, this requires a two-pass decoding procedure.

We propose a novel bilingual LSA framework to perform language model adaptation across languages, enabling adaptation of a language model from one language based on the adaptation text of another language in a single decoding pass. Bilingual LSA consists of two LDA-style models: one for each language trained on parallel document corpora. The key feature of bilingual LSA is a one-to-one topic correspondence between the source and target LSA models. For instance, say topic 10 of the source LSA model is about politics. Then topic 10 of the target LSA model also corresponds to politics and so forth. During language model adaptation, we first infer the topic mixture weights of the source text using the source LSA model. We then transfer the inferred mixture weights into the target LSA model to estimate the target-side LSA marginals for marginal language model adaptation (Kneser et al. 1997). Since bilingual LSA adapts the target language model *before* translation, it does not require the adaptation text to be pre-translated as in monolingual adaptation. The challenge in bilingual LSA is to enforce a one-to-one topic correspondence. Our proposal is to share variational Dirichlet posteriors over topics between the source and target LDA-style models so that a common latent topic space is captured in an unsupervised fashion. Since the topic space is language independent, our approach supports topic transfer in multiple language pairs in O(G) where G is the number of languages. The bilingual LSA framework can also be extended to adapt the translation lexicon via marginal adaptation so that the likelihood of bilingual phrases is sensitive to the topics of the input source text.

Related work on bilingual adaptation includes the Bilingual Topic Admixture Model (BiTAM) for word alignment (Zhao and Xing 2006) and its recent extension to

HMM-BiTAM (Zhao and Xing 2007). BiTAM consists of topic-dependent translation lexicons modeling $p(c|e, k)$ where c, e and k denote the source Chinese word, target English word and the topic index respectively. In contrast, bilingual LSA models only $p(c|k)$ and $p(e|k)$. Bilingual LSA also enables topic-sensitive translation lexicon via marginal adaptation without training topic-dependent lexicons explicitly. Bilingual LSA based on SVD (Kim and Khudanpur 2004) is another approach which attempts to perform language model adaptation across languages. In their approach, bilingual documents are concatenated into a single supervector before SVD. Since SVD does not incorporate prior topic knowledge, bilingual LSA is attractive due to the natural integration of a probabilistic topic prior.

The paper is organized as follows: Sect. 2 gives a brief review of LSA approaches including LDA and latent Dirichlet-tree allocation for monolingual language model adaptation. Section 3 introduces the bilingual LSA approach for cross-lingual language model adaptation, translation lexicon adaptation and phrase table adaptation for SMT. Section 4 presents bilingual LSA adaptation experiments on the medium-scale and the large-scale GALE SMT systems. Section 5 concludes our work with future extensions.

2 Latent semantic analysis

The goal of LSA is to find a set of latent word patterns to describe the document corpora in an unsupervised fashion. The idea is similar to dimensionality reduction techniques such as principal component analysis (Jolliffe 2002) which finds a subset of eigenvectors to represent the data. Latent semantic indexing (LSI) (Deerwester et al. 1990) is an early approach for indexing documents based on the latent semantics/topics behind a document for information retrieval. A term–document matrix W_{VM} is used to represent a training corpus with M documents and V vocabulary items where each column vector consists of the term frequency of a document. SVD is then applied to decompose the matrix into three matrices: $W_{VM} = U_{VK} \cdot S_{KK} \cdot V^T_{MK}$. The matrix U contains K column vectors spanning the latent semantic space. S is a diagonal matrix containing the eigenvalues corresponding to each of the base vectors. Each column of V^T represents the coordinates of a training document in the latent semantic space. As a vector-space model, it is difficult to incorporate prior knowledge into LSI. Therefore, probabilistic modeling for LSA is attractive so that prior knowledge can be incorporated into the model via a Bayesian framework. The probability of a word is usually represented as a linear interpolation of topic-dependent unigram language models with the topic weights modeled with a prior distribution.

One useful application of LSA is unsupervised language model adaptation for automatic speech recognition (ASR) which has shown to be effective in reducing speech recognition errors (Tam and Schultz 2005). The LSA-based adaptation approach is robust because only the K topic weights need to be adapted during the test time according to the current context. We have shown that the estimation of topic weights is robust against speech recognition errors and they can be reliably estimated with a small amount of text from the ASR decoder. In this paper, we highlight that the same adaptation approach can be applied to SMT in a cross-lingual manner.

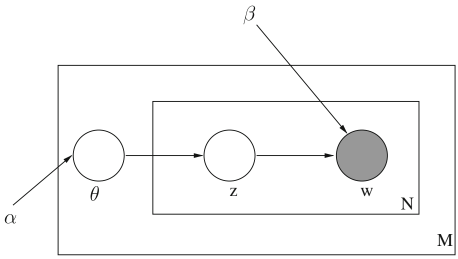

Fig. 1 Graphical representation of the document generation procedure using an LDA-style model. N and M denote the number of words in a document and the number of documents in a training corpus respectively

For background details, we first review LDA (Blei et al. 2003) and describe our extension, latent Dirichlet-tree allocation (LDTA), which extends LDA for modeling topic correlations. We present the marginal language model adaptation approach motivated by information theory.

2.1 Latent Dirichlet allocation

One distinctive feature in LDA over other LSA approaches such as SVD-based LSA and probabilistic LSA is putting a Dirichlet prior over the topic distribution θ. Figure 1 illustrates the graphical model representation of LDA with the left node being the Dirichlet prior over θ.

The document generation procedure of LDA is described as follows:

1. Sample θ from Dirichlet $(\cdot; \{\alpha_k\})$
2. For each word w_i in a document w_1^N,
   - sample a latent topic index z_i from Multinomial (θ)
   - sample w_i from $p(w|z_i)$ (or β_{wz_i})
3. Repeat step 1 to process the next document

where the model parameters Λ are $\{\alpha_k\}$ and $\{\beta_{wk}\}$ for the Dirichlet prior and the topic-dependent unigram distributions respectively with k and w denoting the topic and vocabulary indices respectively. Intuitively, α_k can be interpreted as the prior pseudocount of topic k.

2.2 Latent Dirichlet-tree allocation

One assumption in LDA is the use of a Dirichlet distribution to model the prior topic distribution of a document, implying that LDA inherently assumes topic independence. In other words, knowing the proportion of one topic does not provide any information about the proportion of another topic. In reality, the assumption may not be true since topics may be correlated. For instance, news articles on a newspaper website are usually organized into the main-topic and subtopic hierarchy. From a human point of view, it would be advantageous to model the topic correlation which motivates extending LDA into LDTA (Tam and Schultz 2007). LDTA captures the topic correlation via a structural Dirichlet-tree prior. In fact, a Dirichlet prior is a special case of a Dirichlet-tree prior since a Dirichlet distribution can be visualized as a flat tree

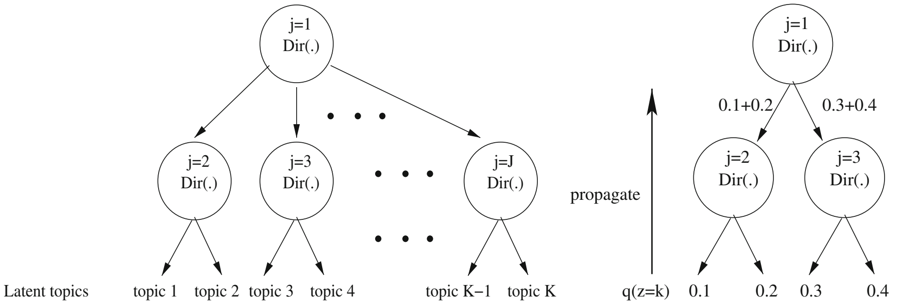

Fig. 2 Left: Dirichlet-tree prior of depth 2. Right: Variational E-step as bottom-up propagation and summation of fractional topic counts

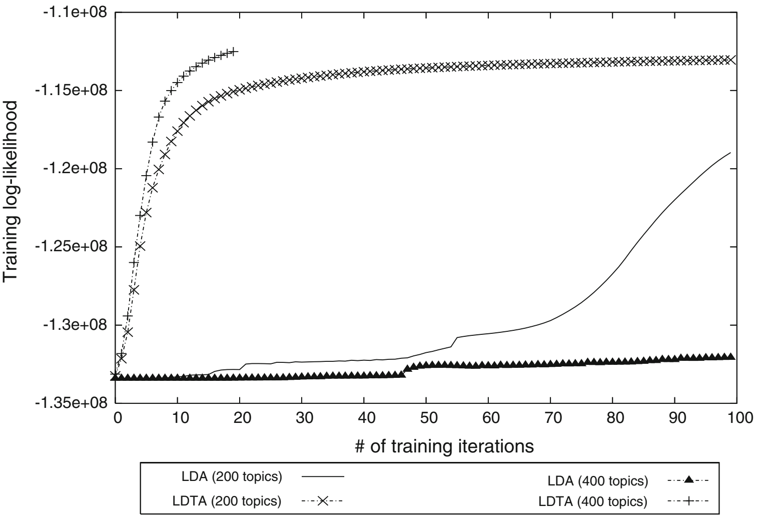

Fig. 3 Log-likelihood comparison of the LDA and LDTA models on the *Xinhua News* 2002 training corpus

with depth 1. Sampling a Dirichlet distribution becomes assigning probability to the branches under a node with probability summing to unity. In general, a Dirichlet tree can have different depth and structure. Figure 2 (Left) illustrates a depth-2 Dirichlet tree where the root node is a Dirichlet distribution with more than two branches while the Dirichlet nodes at the bottom allow only binary branches. Figure 3 shows that the proposed structural prior in LDTA has a better training convergence compared to the unstructured Dirichlet prior in LDA when both models are initialized with a uniform model. Using the structural Dirichlet-tree prior becomes quite significant when the number of topics is increased from 200 to 400 since the convergence curve of LDA becomes flat while the convergence curve of LDTA remains steep. Another advantage of using the Dirichlet-tree prior is the improved perplexity and ASR performance.

Despite the slight complexity, LDTA has the same graphical model representation as LDA and thus shares a similar document generation procedure. The major difference lies in the sampling procedure for the topic distribution θ:

1. Sample a vector of branch probabilities $b_j \sim$ Dirichlet $(\cdot; \{\alpha_{jc}\})$ for each node $j = 1 \ldots J$ where $\{\alpha_{jc}\}$ denotes the parameter of the Dirichlet distribution at node j, that is, the pseudocounts of the outgoing branch c at node j.
2. Compute the topic distribution as in (1),

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

   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 weight θ_k is computed as the product of sampled branch probabilities from the root node to the leaf node corresponding to topic k.
3. For each word w_i in a document w_1^N,
   – sample a latent topic index z_i from Multinomial(θ)
   – sample w_i from $p(w|z_i)$.

The joint distribution of the latent variables (that is, the 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 (2),

$$p(w_1^N, z_1^N, b_1^J) = p(b_1^J|\{\alpha_{jc}\}) \prod_{i=1}^{N} \beta_{w_i z_i} \cdot \theta_{z_i} \tag{2}$$

where

$$p(b_1^J|\{\alpha_{jc}\}) = \prod_{j=1}^{J} \text{Dirichlet}\,(b_j; \{\alpha_{jc}\}) \propto \prod_{jc} b_{jc}^{\alpha_{jc}-1} \tag{3}$$

Similar to LDA training, we apply the variational Bayes approach to optimize the lower bound of the marginalized document likelihood using the Jensen's inequality (4),

$$\begin{aligned} \log p(w_1^N; \Lambda) &= \log \int_{b_1^J} \sum_{z_1^N} q(z_1^N, b_1^J; \Gamma) \cdot \frac{p(w_1^N, z_1^N, b_1^J; \Lambda)}{q(z_1^N, b_1^J; \Gamma)} \\ &\geq \int_{b_1^J} \sum_{z_1^N} q(z_1^N, b_1^J; \Gamma) \cdot \log \frac{p(w_1^N, z_1^N, b_1^J; \Lambda)}{q(z_1^N, b_1^J; \Gamma)} \\ &= L(w_1^N; \Lambda, \Gamma) \end{aligned} \tag{4}$$

where

$$\begin{aligned} L(w_1^N; \Lambda, \Gamma) &= E_q\left[\log \frac{p(w_1^N, z_1^N, b_1^J; \Lambda)}{q(z_1^N, b_1^J; \Gamma)}\right] \\ &= E_q\left[\log p(w_1^N | z_1^N)\right] + E_q\left[\log \frac{p(z_1^N | b_1^J)}{q(z_1^N)}\right] + E_q\left[\log \frac{p(b_1^J; \{\alpha_j\})}{q(b_1^J; \{\gamma_j\})}\right] \end{aligned} \tag{5}$$

$q(z_1^N, b_1^J; \Gamma) = \prod_{i=1}^N q(z_i) \cdot \prod_{j=1}^J q(b_j)$ is a factorizable variational posterior distribution over the latent variables parametrized by Γ which are determined in the E-step. Λ are the model parameters for the Dirichlet tree $\{\alpha_{jc}\}$ and the topic-dependent unigram language model $\{\beta_{wk}\}$. The Dirichlet-tree posterior has the same form as the Dirichlet-tree prior given the topic sequence z_1^N since

$$\begin{aligned} p(b_1^J | z_1^N) &\propto p(z_1^N | b_1^J) \cdot p(b_1^J; \{\alpha_{jc}\}) \\ &\propto \left(\prod_{i=1}^N \prod_{jc} b_{jc}^{\delta_{jc}(z_i)}\right) \cdot \prod_{jc} b_{jc}^{\alpha_{jc}-1} \\ &= \prod_{jc} b_{jc}^{\left(\alpha_{jc} + \sum_{i=1}^N \delta_{jc}(z_i)\right)-1} \\ &= \prod_{j=1}^J Dirichlet(b_j; \{\gamma'_{jc}\}) \end{aligned} \tag{6}$$

Therefore, the conjugate property suggests that the posterior branch count γ_{jc} can be computed by accumulating the expected branch counts from the current observations. Due to the same graphical structure as LDA, the E-steps of LDTA are similar to LDA:
E-Steps:

$$\begin{aligned} \gamma_{jc} &= \alpha_{jc} + \sum_{i=1}^N E_q[\delta_{jc}(z_i)] \\ &= \alpha_{jc} + \sum_{i=1}^N \sum_{k=1}^K \mathbf{q_{ik}} \cdot \delta_{jc}(k) \end{aligned} \tag{7}$$

$$q_{ik} \propto \beta_{w_i k} \cdot e^{E_q[\log \theta_k; \{\gamma_{jc}\}]} \tag{8}$$

where

$$\begin{aligned} E_q[\log \theta_k] &= \sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}] \\ &= \sum_{jc} \delta_{jc}(k) \left(\Psi(\gamma_{jc}) - \Psi\left(\sum_c \gamma_{jc}\right)\right) \end{aligned} \tag{9}$$

where q_{ik} denotes $q(z_i = k | w_1^N)$ meaning the variational topic posterior of word w_i. Equations (7) and (8) are executed iteratively until convergence is reached. Equation

(7) can be implemented as propagation and summation of fractional topic counts q_{ik} from the leaf nodes to the root node in a bottom-up fashion as shown in Fig. 2 (Right).
M-Step:

$$\hat{\beta}_{wk} \propto \sum_{i=1}^{N} q_{ik} \cdot \delta(w_i, w) \tag{10}$$

The re-estimation formula for $\{\beta_{wk}\}$ is the weighted relative word frequency in (10) where $\delta(w_i, w)$ denotes a Kronecker Delta function. The $\{\alpha_{jc}\}$ parameters can be re-estimated with iterative methods such as Newton-Raphson or simple gradient ascent procedure.

2.3 Marginal language model adaptation

Given an adaptation text, we apply the E-steps (7)–(8) to estimate the variational Dirichlet posterior of each node in the Dirichlet tree. We estimate the topic weights of the adaptation text using (11).

$$\hat{\theta}_k \propto \prod_{jc} \left(\frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \tag{11}$$

Then we apply the topic weights into the LSA model to obtain the in-domain LSA marginals (12).

$$p_{ldta}(w) = \sum_{k=1}^{K} \beta_{wk} \cdot \hat{\theta}_k \tag{12}$$

We integrate the LSA marginals into the background language model using marginal adaptation (Kneser et al. 1997) which minimizes the Kullback-Leibler divergence between the adapted language model and the background language model in (13),

$$p_a(w|h) = \frac{\alpha(w) \cdot p_{bg}(w|h)}{\sum_v \alpha(v) \cdot p_{bg}(v|h)} \tag{13}$$

where

$$\alpha(w) = \left(\frac{p_{ldta}(w)}{p_{bg}(w)} \right)^{\epsilon} \tag{14}$$

and ϵ is a tuning parameter between 0 and 1. Intuitively, the likelihood ratio $\alpha(w)$ controls when to boost/de-emphasize the probability of on-topic/off-topic n-gram entry.

Computing the normalization factor in (13) involves summing over all possible words v which is prohibitive on large vocabulary. We employed the strategy proposed in Kneser et al. (1997) to compute the normalization factor efficiently. The idea is to further impose a constraint that the total probability mass of the observed word transition $(h, \cdot)$ in the background training corpus is conserved after language model adaptation (15),

$$\sum_{v:(h,v)\in T} p_a(v|h) = \sum_{v:(h,v)\in T} p_{bg}(v|h) = \text{Mass } (h) \tag{15}$$

where the summation is taken *only* on the observed history and word pair (h, v) in a training set T. Given that our background language model has a standard backoff structure plus the above constraint, the adapted language model has the recursive backoff formula in (16),

$$p_a(w|h) = \begin{cases} \frac{\alpha(w)\cdot p_{bg}(w|h)}{z_0(h)} & \text{if } (h, w) \in T \\ bo(h) \cdot p_a(w|\hat{h}) & \text{otherwise} \end{cases} \tag{16}$$

where

$$\frac{1}{z_0(h)} = \frac{\text{Mass } (h)}{\sum_{v:(h,v)\in T} \alpha(v) \cdot p_{bg}(v|h)} \tag{17}$$

and

$$bo(h) = \frac{1 - \text{Mass } (h)}{1 - \text{Mass } (\hat{h})} \tag{18}$$

$bo(h)$ denotes the backoff weight of word history h to ensure that $p_a(w|h)$ sums to unity. The backoff weights need to be updated accordingly after all the n-gram probability entries in the backoff language model are adapted. $\hat{h}$ denotes the reduced word history of h. The intuition behind the factor $z_0(h)$ is to perform "normalization" similar to (13), but the summation involves only a subset of words observed in T with the same word history h.

When w is a stopword such as auxiliary verb, article, conjunction, sentence boundary marker or punctuation mark, we do not adapt its n-gram probabilities because predicting stopwords mostly relies on the syntactic context but not the topical context. We can easily model this effect by inserting a new branch in (16) to indicate that $p_a(w|h) = p_{bg}(w|h)$ when w is a stopword. Hence, the computation of Mass (h) and $z_0(h)$ needs to be modified with stopwords being excluded from summation in (15) and (17) respectively.

3 Bilingual latent semantic analysis

The goal of bilingual LSA is to enforce a one-to-one topic correspondence between the source and the target LDA-style models so that the inferred latent topic distribution can be transferred 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 4 illustrates the idea of topic transfer between monolingual LSA models followed by language model and

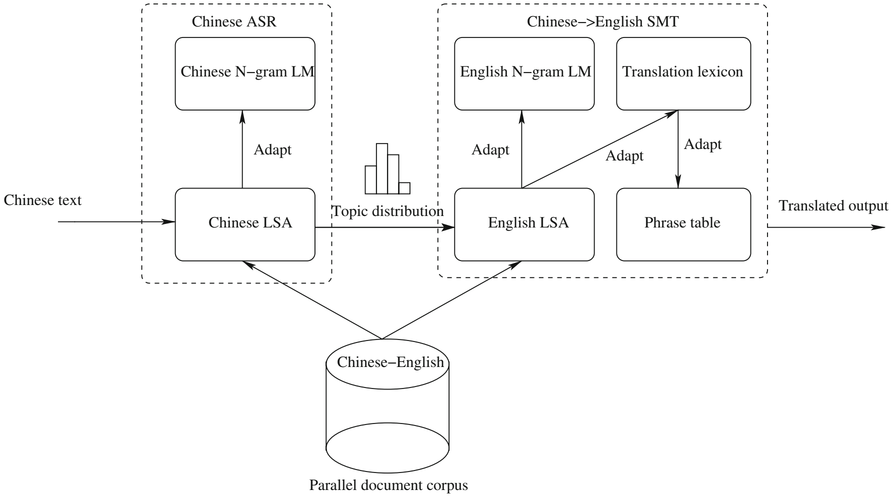

Fig. 4 Bilingual LSA-based adaptation via transfer of topic distribution from the source language to the target language for SMT

translation lexicon adaptation. The novelty of our work lies in exploiting the source text to adapt the language model on the target language and the translation lexicon for SMT. Since our approach adapts the models before translation using the source text, the extra decoding pass to obtain an initial translated output for adaptation can be saved. For the same reason, propagation of translation errors can be avoided by using the source text for adaptation.

In the following sections, we present an approach for bilingual LSA training. We transfer the language model adaptation technique from a monolingual to a bilingual setting and extend marginal adaptation for adapting the translation lexicon.

3.1 Bilingual LSA training

Bilingual LSA training is a two-stage procedure. In the first stage, we perform monolingual LSA training using the variational EM algorithm (7)–(10) on the source documents in the parallel corpora. We use the model to compute the term $e^{E_q[\log \theta_k]}$ in (8) for each source document. In the second stage, we apply the same term $e^{E_q[\log \theta_k]}$ to *bootstrap* a target LSA model, which is the key to enforcing a one-to-one topic correspondence. The hyperparameters of the variational Dirichlet posteriors of each node in the Dirichlet tree are now shared among the source and target models. Precisely, we apply only (8) with fixed $e^{E_q[\log \theta_k]}$ in the E-step and (10) in the M-step to estimate $\{\beta_{wk}\}$ of the target LSA model. Figure 5 illustrates the idea of enforcing one-to-one topic correspondence of parallel document pairs during bootstrapping a target LSA model from a source LSA model denoted as bLSA$_{(src,tgt)}$.

Notice that the E-step is noniterative resulting in rapid LSA training. In short, given a monolingual LSA model, we can rapidly bootstrap LSA models of new languages using parallel corpora. Since the topic transfer can be bidirectional, we can perform

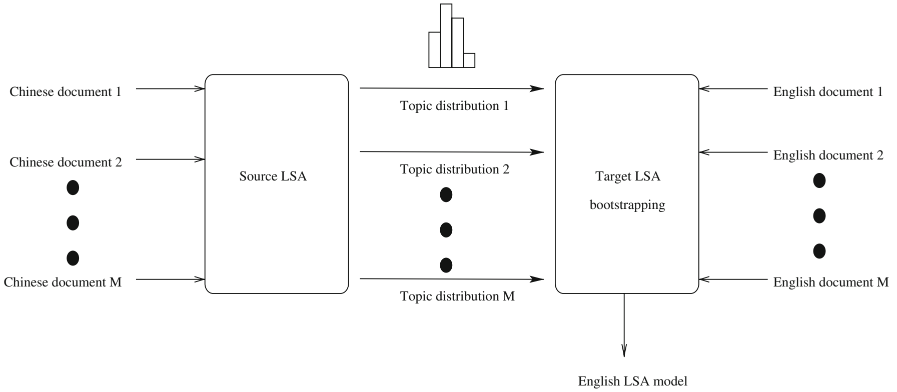

Fig. 5 LSA bootstrapping via sharing of variational Dirichlet posteriors for parallel documents

the bilingual LSA training in a reverse manner, that is, training a target LSA model followed by bootstrapping a source LSA model denoted as bLSA$_{(tgt,src)}$.

3.2 Cross-lingual language model adaptation

Marginal language model adaptation in cross-lingual settings can be performed in almost the same manner as in monolingual settings as described in Sect. 2.3 except that the source text is used for adaptation in the cross-lingual case. First, we estimate the topic weights of LDTA on the source language using (19).

$$\hat{\theta}_k^{(CH)} \propto \prod_{jc} \left(\frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \tag{19}$$

Then we apply the source topic weights into the target LSA model to obtain the in-domain LSA marginals as in (20).

$$p_{EN}(w) = \sum_{k=1}^{K} \beta_{wk}^{(EN)} \cdot \hat{\theta}_k^{(CH)} \tag{20}$$

Finally, we follow the same marginal language model adaptation procedure described in Sect. 2.3.

3.3 Translation lexicon adaptation

The bilingual LSA approach not only applies for language model adaptation, but also for translation lexicon adaptation. Thus, the adapted translation lexicon can be used to score phrase pairs depending on the topical context. Motivated by information theory, we formulate the problem as marginal adaptation under the bilingual LSA framework.

The goal is to minimize the Kullback-Leibler divergence between the adapted lexicon $p_a(c|e)$ and the background lexicon $p_{bg}(c|e)$ such that the lexical marginals computed from the adapted lexicon are equal to the in-domain marginals $p(c|d_{ch})$ which are given a priori. Thus the objective function to minimize is as in (21),

$$\begin{aligned} &\text{Minimize} \sum_e p_a(e) \cdot KL\left(p_a(.|e)||p_{bg}(.|e)\right) \\ &\text{such that } \forall c : \sum_e p_a(e) \cdot p_a(c|e) = p(c|d_{ch}) \\ &\forall e : \sum_c p_a(c|e) = 1 \end{aligned} \tag{21}$$

where d_{ch} is the source text to estimate the unigram distribution. We write the Lagrangian of the objective function, take the derivative with respect to $p_a(c|e)$ and set it to zero (22)–(23).

$$\begin{aligned} D(p_a(.|.)) = &\sum_e p_a(e) \cdot \sum_c p_a(c|e) \cdot \log \frac{p_a(c|e)}{p_{bg}(c|e)} \\ &- \sum_c \lambda_c \left(\sum_e p_a(e) \cdot p_a(c|e) - p(c|d_{ch}) \right) \\ &- \sum_e \mu_e \left(\sum_c p_a(c|e) - 1 \right) \end{aligned} \tag{22}$$

$$\begin{aligned} \frac{\partial D(.)}{\partial p_a(c|e)} &= p_a(e) \cdot (1 + \log \frac{p_a(c|e)}{p_{bg}(c|e)}) - \lambda_c \cdot p_a(e) - \mu_e = 0 \\ \Rightarrow p_a(c|e) &\propto p_{bg}(c|e) \cdot e^{\lambda_c} \propto p_{bg}(c|e) \cdot e^{\sum_j \lambda_j \cdot f_j(c,e)} \end{aligned} \tag{23}$$

where

$$f_j(c, e) = \begin{cases} 1 & \text{if } c = j \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

$f_j(c, e)$ is a unigram feature function independent of e. Since the solution of the adapted lexicon is in exponential form, the optimization problem is similar to the maximum entropy settings. Therefore, we solve λ_j using the generalized iterative scaling (GIS) (Darroch and Ratcliff 1972) as in (25)–(29),

$$\forall j : \quad \lambda_j^{(t+1)} = \lambda_j^{(t)} + \log \frac{\tilde{E}[f_j(c, e)]}{E[f_j(c, e)]} \tag{25}$$

$$= \lambda_j^{(t)} + \log \frac{\sum_{c,e} \tilde{p}(c, e|d_{ch}) \cdot f_j(c, e)}{\sum_{c,e} p_a^{(t)}(c|e) \cdot p_a(e) \cdot f_j(c, e)} \tag{26}$$

$$= \lambda_j^{(t)} + \log \frac{\sum_e \tilde{p}(c = j, e|d_{ch})}{\sum_e p_a^{(t)}(c = j|e) \cdot p_a(e)} \tag{27}$$

$$= \lambda_j^{(t)} + \log \frac{p(c = j|d_{ch})}{\sum_e p_a^{(t)}(c = j|e) \cdot p_a(e)} \tag{28}$$

$$\approx \lambda_j^{(t)} + \log \frac{p(c = j|d_{ch})}{\sum_e p_a^{(t)}(c = j|e) \cdot \mathbf{p_{blsa}(e|d_{ch})}} \tag{29}$$

where t denotes the GIS iteration index with $p_a^{(0)}(c|e) = p_{bg}(c|e)$ and $\lambda_j^{(0)} = 0$. $p_a(e)$ is approximated by the English LSA marginals $p_{blsa}(e|d_{ch})$ from the bilingual LSA. Since the range of e in (28) is limited to the number of possible translation word pairs (c, e) in the lexicon, computing the denominator term is efficient without evaluating all possible e. We estimate $p(c|d_{ch})$ using the smoothed relative word frequency of the source text with the Good-Turing discounting scheme. Since the optimization is convex, a global optimal solution of the adapted lexicon is guaranteed. Since the source marginals $p(c|d_{ch})$ are accurately estimated using the source text, the adapted lexicon is expected to outperform the background lexicon in terms of the conditional likelihood $p(C|E)$ where $C = c_1^I$ and $E = e_1^J$ denote the translation pair of Chinese and English sentences respectively.

3.4 Phrase table adaptation

Ideally, the adapted translation lexicon can be applied directly during phrase extraction. But this involves extra implementation work in the phrase extraction algorithm. An alternative approach is to take the background phrase table and assume that good phrase pairs are already captured in the table. With the adapted translation word lexicons, we can score each phrase pair (c_1^I, e_1^J) in a background phrase table similar to IBM model 1 (Brown et al. 1994) (30),

$$p_a(c_1^I|e_1^J) = \prod_{i=1}^{I} \frac{1}{J_i} \cdot \sum_{j}^{J_i} p_a(c_i|e_j) \tag{30}$$

where $0 < J_i \leq J$ denotes the effective number of target words e_j aligned to a source word c_i after pruning the unlikely lexical entries with probability less than 10^{-4} in the adapted translation word lexicon. The motivation is to have a “sharper” average of word probability and thus make the phrase score more discriminative. The NULL model $p(c|\text{NULL})$ or the minimum of $p(c|\text{NULL})$ is used as a backoff model to avoid zero probability of unseen translation. $p_a(e_1^J|c_1^I)$ can be defined in the same manner. For phrase table adaptation, these two bilingual LSA-adapted phrase scores are simply added to the background phrase table for subsequent minimum error rate training (MERT) and SMT decoding.

4 Experimental setup

The proposed adaptation approach was evaluated on the medium-scale SMT development system and the large-scale GALE Phase 2.5 SMT system translating from Chinese to English. Phrase extraction was performed using parallel corpora as shown in Table 1. The table also shows that 4-gram and 5-gram language models were trained using 500M and 2.7G English words for the medium-scale system and the GALE system respectively.

The SRILM toolkit (Stolcke 2002) was used for language model training with the modified Kneser-Ney smoothing. SMT was performed by constructing a translation lattice which contained all possible matched bilingual phrase pairs of an input source sentence. In the GALE system, part-of-speech based word reordering (Rottmann and Vogel 2007) was performed on an input sentence to produce an input source lattice before building the translation lattice. Search was then performed on this lattice using our STTK beam-search decoder (Vogel et al. 2003). The word reordering window was set to 3 for the medium-scale system while monotonic decoding was applied for the GALE system since word reordering was already applied in the source lattices. An optimal path was returned with the maximum translation score consisting of a log-linear combination of feature functions such as language model probability, distortion penalty, word-count penalty, phrase count, phrase-alignment scores and the bilingual LSA-adapted phrase scores. The corresponding feature weights were optimized in BLEU (Papineni et al. 2002) using MERT (Och 2003). Both systems were tuned on MT03 and evaluated on MT06 with four English references as shown in Table 2.

The bilingual LSA model was trained using the Chinese–English parallel document corpora consisting of the FBIS corpus, *Xinhua News*, *Hong Kong News*, *Donga News* and *Sinorama* articles. The combined corpora contain 96k parallel documents with 41M Chinese words and 50M English words as shown in Table 3.

Our bilingual LSA training did not take advantage of the larger parallel corpora used in phrase extraction due to the loss of document boundary information. However, encouraging results were still achieved. The number of latent topics K in LSA was set to 200 based on our best knowledge of language model adaptation for ASR. A balanced binary Dirichlet-tree prior was used. The source and target vocabulary

Table 1 Size of the language model training corpora and the parallel training corpora for phrase extraction in terms of number of words

System	Language model	Training corpus	
		Chinese	English
Development system	500M	59M	67M
GALE Phase 2.5 system	2.7G	232M	260M

Table 2 Statistics of the development set (MT03) and the test set (MT06)

Set	Sentences	Documents	References
MT03 (dev)	919	100	4
MT06 (test)	1664	79	4

Table 3 Size of the parallel training corpora for bilingual LSA training

Language	Words	Documents
Chinese	41M	96k
English	50M	96k

in bilingual LSA were limited to words occurring in the phrase table. The Stanford Chinese word segmenter (Tseng et al. 2005) was applied to segment the Chinese side of the parallel corpora. Monolingual LSA training was first applied on the Chinese side followed by LSA bootstrapping on the English side. Prior empirical results indicated that the reverse bootstrapping direction resulted in similar performance.

Each source test document was used for model adaptation before SMT decoding. The marginal adaptation approach described in Sect. 3.2 was applied to the English background language model. Words on the stopword list[1] plus punctuation were filtered out from the language model adaptation. Translation word lexicons were also adapted via marginal adaptation as described in Sect. 3.3. Two bilingual LSA-adapted phrase scores on both translation directions were added as new feature functions in an offline fashion before MERT and SMT decoding. The performance of language model adaptation was measured using word perplexity while the MT performance was reported using BLEU and NIST (Doddington 2002) scores. Significance testing was performed using a bootstrapping approach described in Zhang and Vogel (2004). Human evaluation was carried out to compare the translation performance of the bilingual LSA-adapted GALE SMT system with the unadapted baseline. For comparison purposes, only the test sentences which had different translations from the SMT systems were considered. Due to limited resources, only a random subset of test sentences was used. The test sentences were randomly divided into the core set and the remaining set. Each grader worked on the same core set while the remaining set was subdivided into nonoverlapping sets for each grader. The core set and the grader-specific set contained 30 and 131 sentences respectively. Each grader assigned two scores to each sentence from two different systems based on fluency and adequacy with respect to the English references ranging from 1 (worst) to 5 (best). Four graders were involved in the human evaluation.

4.1 Language model adaptation results

The proposed bilingual LSA training approach enforces a one-to-one topic correspondence successfully and extracts parallel topics as shown in Table 4. The Chinese and English topical words in the table are strongly correlated and many of them are translation pairs, indicating that bilingual LSA works as a cross-lingual word trigger model. Figure 6 demonstrates that our proposed approach leads to rapid training convergence due to sharing of variational Dirichlet posteriors with the Chinese LSA model compared to monolingual English LSA training starting with the same flat model. On

[1] See http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words, last consulted 22 October 2008

Table 4 Parallel topics extracted by $bLSA_{(CH,EN)}$

Topic	Top words sorted by $p(w\|k)$
CH-40	飞 *fei* ‘flying’, 潜 艇 *qianting* ‘submarine’, 飞 机 *feiji* ‘aircraft’, 空 中 *kongzhong* ‘in the air’, 飞行员 *feixingyuan* ‘pilot’, 任务 *renwu* ‘mission’
EN-40	air, sea, submarine, aircraft, flight, flying, ship, test
CH-41	卫星 *weixing* ‘satellite’, 航天 *hangtian* ‘space travel’, 发射 *fashe* ‘launch’, 太空 *taikong* ‘space’, 中国 *zhongguo* ‘china’, 技术 *jishu* ‘technology’
EN-41	space, satellite, china, technology, satellites, science
CH-42	消防 *xiaofang* ‘fire control’, 机场 *jichang* ‘airport’, 服务 *fuwu* ‘services’, 火警 *huojing* ‘fire accident’, 船只 *chuanzhi* ‘ship’, 乘客 *chengke* ‘passengers’
EN-42	fire, airport, services, department, marine, air, service, passengers

Top words on the Chinese side are translated into English for illustration purposes

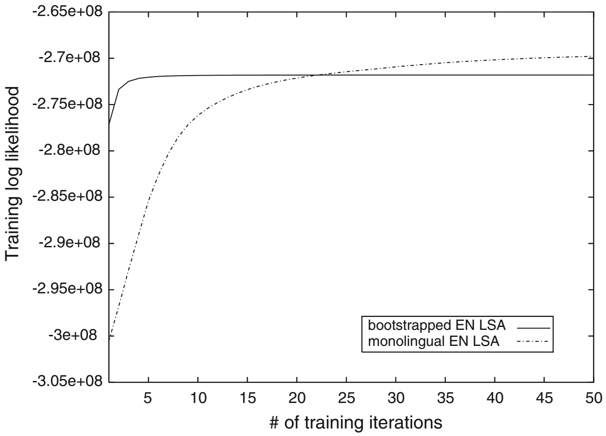

Fig. 6 Training log-likelihood of the bootstrapped English LSA from the Chinese LSA compared to the flat monolingual English LSA

the other hand, monolingual LSA training has better training likelihood with more training iterations which is reasonable since the bootstrapping approach constrains the parameter space so that a one-to-one topic correspondence is satisfied while the parameter space of monolingual LSA training is unconstrained.

4.1.1 Perplexity results

Table 5 shows that the proposed approach effectively reduces the English word perplexity by 17.5% and 10.9% relative for the 4-gram and 5-gram language models used in the medium-scale system and the GALE system respectively compared to the unadapted language model. Bilingual LSA adaptation still helps even on a huge 5-gram language model trained on a large amount of text.

Table 5 MT06 evaluation results on target word perplexity, BLEU and NIST using 67M-word (260M-word) parallel corpora for phrase extraction and 500M-word (2.7G-word) English corpora for language model training

Language model	Perplexity	BLEU (%)	NIST
Baseline EN 4-gram (500M)	154	28.06	8.71
bilingual LSA-adapted	127	28.62	8.80
bilingual LSA-adapted lexicon	–	28.59	8.92
bilingual LSA-adapted + lexicon	–	28.91	8.97
mono LSA-adapted	125	28.41	8.81
mono LSA-adapted lexicon	–	28.72	8.96
mono LSA-adapted + lexicon	–	28.97	9.00
Baseline EN 5-gram (2.7G)	147	31.26	9.15
bilingual LSA-adapted	131	31.43	9.20
bilingual LSA-adapted lexicon	–	31.69	9.26
bilingual LSA-adapted + lexicon	–	31.85	9.31

Vocabulary size of the target language model is 1.3M (4.1M). Four English references are used for scoring

4.2 Lexicon adaptation results

Marginal adaptation results in a sharper translation lexicon in which the uncertainty of word-to-word translation is reduced. For instance, the probability of translating the English word *Korea* into the related (correct) Chinese translation 韩国 *hanguo* was boosted from 0.32 to 0.57 while the probability of unrelated (incorrect) translation 访问 *fangwen* ‘visit’ was greatly de-emphasized from 1.8×10^{-4} to 8.7×10^{-7} after bilingual LSA adaptation. Redistribution of probability mass from the unrelated words to the related words occurs during translation lexicon adaptation according to the topical context of the source text.

4.3 Translation results

The upper section of Table 5 shows the translation performance in BLEU and NIST scores on MT06 using the medium-scale SMT system. 2% relative improvement in BLEU is achieved compared to the unadapted baseline after applying bilingual LSA-based language model adaptation and translation lexicon adaptation separately. When both techniques are applied simultaneously, the gain is additive giving 3% relative improvement in BLEU compared to the unadapted baseline. The improvement is statistically significant at the 95% confidence interval [27.29%,28.84%] with respect to the unadapted baseline. The same performance trend in NIST is also observed with 3% relative improvement compared to the unadapted baseline. The improvement is statistically significant at the 95% confidence interval [8.61,8.85] with respect to the unadapted baseline.

The middle section of Table 5 shows that monolingual LSA adaptation using the first-pass translated hypotheses achieves similar performance compared to bilingual

Table 6 Examples demonstrating some degree of semantic paraphrasing with bilingual LSA

Sample output 1	
Baseline	To achieve the extensive support from the international community to save this problem, the **government of Denmark, and Denmark** is very important.
Bilingual LSA	To achieve the extensive support from the international community to save this problem, the **Danish government and people of Denmark** is very important.
Reference	It is extremely important to the **Danish government and the Danish people** to obtain the broad support of the international community to pass through this difficulty.
Sample output 2	
Baseline	In an interview **Hoffman** CBS news magazine "60 minutes" . . .
Bilingual LSA	**Hoffman told** the CBS news magazine "60 minutes" . . .
Reference	**Hoffman sighed** when doing an interview with America's CBS news magazine "60 minutes"

LSA adaptation using the source text. In other words, the source text and the initial MT hypotheses are equally effective for LSA adaptation since LSA is robust against translation errors in the adaptation text. We conjecture that the quality of translation of topical unigrams should be acceptable in the initial translation. But in terms of computation, bilingual LSA is more elegant and requires only a single decoding pass compared to monolingual LSA.

Table 6 shows some sample sentences demonstrating some degree of semantic paraphrasing with bilingual LSA, such as *people of Denmark* versus *Danish people*, and *told* versus *sighed*.

The lower section of Table 5 shows the translation performance using the GALE Phase 2.5 SMT system. The performance trend is similar to the medium-scale system. Improvement in BLEU and NIST are observed after applying bilingual LSA-based language model adaptation and translation lexicon adaptation. Additive gain is observed after applying both techniques together, yielding 1.9% relative improvement in BLEU compared to the unadapted baseline. The gain in BLEU is reduced compared to the results on the medium-scale setting, and the gain is not statistically significant. This may be explained by having a stronger baseline language model and a better word reordering strategy in the GALE system. The improvement in NIST follows a similar trend with 1.7% relative improvement compared to the unadapted baseline. However, the gain is statistically significant at the 95% confidence interval [9.058,9.305] with respect to the unadapted baseline.

4.4 Human evaluation results

Table 7 shows the human evaluation results in sentence fluency and adequacy. Consistent improvement in fluency but slight degradation in adequacy were observed across most graders on the bilingual LSA-adapted sentences. Overall, bilingual LSA achieves a better average score than the unadapted baseline although the gain is not statistically significant.

Table 8 shows an example in which bilingual LSA gives a better fluency than the unadapted baseline.

Table 7 Human evaluation results on sentence fluency and adequacy on MT06 using the GALE Phase 2.5 SMT system compared with the bilingual LSA (bLSA)

Grader ID	Fluency		Adequacy		Average	
	Baseline	bLSA	Baseline	bLSA	Baseline	bLSA
1	3.15	**3.29**	**3.76**	3.70	3.46	**3.50**
2	3.34	**3.38**	**3.28**	3.26	3.31	**3.32**
3	2.88	**3.03**	**2.97**	2.95	2.93	**2.99**
4	3.96	**4.00**	**3.92**	3.79	**3.94**	3.90

Worst score is 1 and the best score is 5

Table 8 Example where bilingual LSA gives a better fluency than the unadapted baseline

Sample output 3	
Baseline	It is necessary to cultivate the sense of innovation in the whole society, vigorously promote **innovative spirit, courage competition,** strive to create a good atmosphere of talent.
Bilingual LSA	It is necessary to cultivate the sense of innovation in the whole society, vigorously promote **the spirit of innovation, and be bold enough to compete and** strive to create a good atmosphere of talent.
Reference	Anhui must foster innovative knowledge among the entire society, greatly promote **a spirit of willingness to innovate and compete, and** exert itself to build an excellent atmosphere where human resources come forth in large numbers.

5 Conclusion

We proposed a bilingual LSA for cross-lingual language model adaptation and translation lexicon adaptation for SMT. The bilingual LSA model consists of source and target monolingual LSA models in which a one-to-one topic correspondence is enforced between the models via sharing of variational Dirichlet posteriors. The inferred topic distribution of an input source text can be transferred to the target language to estimate the in-domain LSA marginals for language model adaptation and translation lexicon adaptation on the target side. We also show that rapid bootstrapping of an LSA model for a new language can be performed from a well-trained LSA model of another language. Results show that our approach significantly reduces the word perplexity of the target language model. When the adapted language model or lexicon is applied separately, improvement in BLEU and NIST scores is observed. When both models are applied simultaneously, the gain is additive. On the medium-scale SMT system, the improvement is statistically significant at the 95% confidence interval with respect to the unadapted baseline. Effective language model adaptation improves word reordering while a better translation word lexicon leads to better phrase table. Our approach works well on the large-scale evaluation and produces consistent improvement using the GALE SMT system. The improvement in NIST is statistically significant at the 95% confidence interval.

Future extensions include evaluating the approach on the Arabic–English language pair and the exploration of automatic story segmentation so that the bilingual LSA training can benefit from a parallel sentence corpus without story boundaries.

Acknowledgements We would like to thank Silja Hildebrand for the SMT setup, Matthias Paulik for his code for language model reloading, Ying Zhang for scripts to perform significance tests, and Stephan Vogel for discussion and scripts to subsample phrase tables. We would like to thank our colleagues for their generous participation in human evaluation. 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

Bellegarda JR (2000) Large vocabulary speech recognition with multispan statistical language models. IEEE Trans Speech Audio Process 8:76–84

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach Learn Res 3:1107–1135

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1994) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19:263–311

Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. Ann Math Stat 43:1470–1480

Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407

Doddington G (2002) Automatic evaluation of MT quality using n-gram co-occurrence statistics. In: Proceedings of human language technology conference 2002, San Diego, CA, pp 138–145

Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB (2004) Integrating topics and syntax. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems 17, Proceedings of the 2004 conference, MIT Press, Cambridge MA, pp 537–544

Hofmann T (1999) Probabilistic latent semantic indexing. In: UAI '99, proceedings of the fifteenth conference on uncertainty in artificial intelligence, Stockholm, Sweden, pp 289–296

Hsu B-J(P), Glass J (2006) Style & topic language model adaptation using HMM-LDA. In: EMNLP 2006, 2006 conference on empirical methods in natural language processing, Sydney, Australia, pp 373–381

Iyer R, Ostendorf M (1996) Modeling long distance dependence in language: topic mixtures vs. dynamic cache models. In: ICSLP 96, fourth international conference on spoken language processing, Philadelphia, PA, pp 236–239

Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York

Kim W, Khudanpur S (2003) LM adaptation using cross-lingual information. In: 8th European conference on speech communication and technology (Eurospeech 2003 – Interspeech 2003), Geneva, Switzerland, pp 3129–3132

Kim W, Khudanpur S (2004) Cross-lingual latent semantic analysis for LM. In: 2004 IEEE international conference on acoustics, speech, and signal processing, vol 1. Montreal, Quebec, Canada, pp 257–260

Kneser R, Peters J, Klakow D (1997) Language model adaptation using dynamic marginals. In: Proceedings of Eurospeech '97, 5th European conference on speech communication and technology, Rhodes, Greece, pp 1971–1974

Mrva D, Woodland PC (2006) Unsupervised language model adaptation for Mandarin broadcast conversation transcription. In: Interspeech 2006 – ICSLP, ninth international conference on spoken language processing, Pittsburgh, Pennsylvania, paper 1549-Thu1A2O.3

Och FJ (2003) Minimum error rate training in statistical machine translation. In: ACL-03, 41st annual meeting of the Association for Computational Linguistics, Sapporo, Japan, pp 160–167

Papineni K, Roukos S, Ward T, Zhu W (2002) BLEU: a method for automatic evaluation of machine translation. In: 40th annual meeting of the Association of Computational Linguistics, Philadelphia, Pennsylvania, pp 311–318

Paulik M, Fügen C, Schaaf T, Schultz T, Stüker S, Waibel A (2005) Document driven machine translation enhanced automatic speech recognition. In: Proceedings of Interspeech'2005 – Eurospeech, 9th European conference on speech communication and technology, Lisbon, Portugal, pp 2261–2264

Rottmann K, Vogel S (2007) Word reordering in statistical machine translation with a POS-based distortion model. In: TMI 2007, proceedings of the 11th international conference on theoretical and methodological issues in machine translation, Skövde, pp 171–180

Stolcke A (2002) SRILM – an extensible language modeling toolkit. In: Proceedings of the 7th international conference on spoken language processing ICSLP/Interspeech, Denver, Colorado, pp 901–904

Tam YC, Schultz T (2005) Language model adaptation using variational Bayes inference. In: Proceedings of Interspeech'2005 – Eurospeech, 9th European conference on speech communication and technology, Lisbon, Portugal, pp 5–8

Tam YC, Schultz T (2006) Unsupervised language model adaptation using latent semantic marginals. In: Interspeech 2006 – ICSLP, ninth international conference on spoken language processing, Pittsburgh, Pennsylvania, paper 1705-Thu1A2O.2

Tam YC, Schultz T (2007) Correlated latent semantic model for unsupervised language model adaptation. In: Proceedings of ICASSP 2007, international conference on acoustics, speech, and signal processing, vol IV. Honolulu, Hawaii, pp 41–44

Tseng H, Chang P, Andrew G, Jurafsky D, Manning C (2005) A conditional random field word segmenter. In: IJCNLP-05, fourth SIGHAN workshop on Chinese language processing, Jeju Island, Korea, pp 168–171

Vogel S, Zhang Y, Huang F, Tribble A, Venugopal A, Zhao B, Waibel A (2003) The CMU statistical translation system. In: MT summit IX, proceedings of the ninth machine translation summit, New Orleans, pp 402–409

Zhang Y, Vogel S (2004) Measuring confidence intervals for the machine translation evaluation metrics. In: Proceedings of the tenth conference on theoretical and methodological issues in machine translation TMI-04, Baltimore, Maryland, pp 85–94

Zhao B, Xing EP (2006) BiTAM: Bilingual topic admixture models for word alignment. In: Coling · ACL 2006, 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics, proceedings of the main conference poster sessions, Sydney, Australia, pp 969–976

Zhao B, Xing EP (2007) HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In: Twenty-second annual conference on neural information processing systems, Vancouver BC, Canada