Rapid Unsupervised Topic Adaptation – a Latent Semantic Approach

Yik-Cheung Tam

CMU-LTI-09-016

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Tanja Schultz (Chair)
Alex Waibel
Stephan Vogel
Sanjeev P. Khudanpur, Johns Hopkins University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

To Cindy, and my family,

Abstract

In open-domain language exploitation applications, a wide variety of topics with swift topic shifts has to be captured. Consequently, it is crucial to rapidly adapt all language components of a spoken language system. This thesis addresses unsupervised topic adaptation in both monolingual and crosslingual settings. For automatic speech recognition we rapidly adapt a language model on a source language. For statistical machine translation, we adapt a language model of a target language, a translation lexicon and a phrase table using a source text.

For monolingual adaptation, we propose latent Dirichlet-Tree allocation for Bayesian latent semantic analysis. Our model enables rapid incremental language model adaptation via caching the fractional topic counts of word hypotheses decoded from previous speech utterances. Latent Dirichlet-Tree allocation models topic correlation in a tree-based hierarchy and thus addresses the model initialization issue. To address the "bag-of-word" assumption in latent semantic analysis, we extend our approach to N-gram latent Dirichlet-Tree allocation. We investigate a fractional Kneser-Ney smoothing approach to handle fractional counts for topic models. The algorithm produces a more compact model compared to the Witten-Bell smoothing. Using multi-stage language model adaptation via N-gram latent Dirichlet-Tree allocation, we achieve significant reduction in speech recognition errors using our large-scale GALE systems on two different languages: Mandarin and Arabic. For end-to-end translation on speech inputs, applying topic adaptation on automatic speech recognition is beneficial to translation performance.

For crosslingual adaptation, we propose bilingual latent semantic analysis for statistical machine translation. A key feature of bilingual latent semantic analysis is a one-to-one topic correspondence between models of a source and a target language. Since topical information is language independent, our model enables transfer of a topic distribution inferred from a source text to a target language for crosslingual adaptation. Our approach has two advantages: first, it can be applied before translation, and thus has immediate impact on translation. Secondly, it does not rely on an translation output for adaptation, and

therefore does not suffer from translation errors. Together with N-gram latent Dirichlet-Tree allocation on a target language, we achieve significant improvement in translation performance using our large-scale GALE systems for text translation.

A limitation of bilingual latent semantic analysis is the requirement of parallel corpora that are relative expensive to collect. We propose a semi-supervised approach to incorporate non-parallel documents into model training. We achieve improvement in crosslingual language model adaptation performance, especially when bilingual resources are deficient.

Acknowledgments

First and foremost, thanks to my advisor, Tanja Schultz, for offering me a chance to learn and grow in the ISL/InterACT lab. I express my deep appreciation for her support, guidance and freedom to work on my thesis. Many thanks to Tanja Schultz, Stephan Vogel, Alex Waibel, and Sanjeev Khudanpur for reading my thesis. Their valuable comments have improved the quality of my thesis.

Special thanks to Thomas Schaaf for teaching me how to incorporate C++ modules into our IBIS speech decoder. I am very thankful for the opportunity to learn from Thomas especially during the RT04 and TCSTAR evaluation. Many thanks to Christian Fügen, Sebastian Stüker, Florian Metze, Hua Yu and Stan Jou for sharing their experience on the decoder. I would like to thank Stephan Vogel for his advice on the STTK decoder for statistical machine translation.

I am very grateful to Ian Lane for his insightful comments on my research and his contribution to our spoken language translation system for our ACL paper. My work has been benefited from his suggestion and encouragement.

Many thanks to my colleague for offering help and pointers over the past years: Mohamed Noamany, Mark Fuhs, Roger Hsiao, Udhyakumar Nallasamy and Qin Jin for my experiment on automatic speech recognition; Silja Hildebrand, Qin Gao, Nguyen Bach, Thuylinh Nguyen, Qin Gao, Matthias Eck, Matthias Paulik, Anthony D'Auria, Bing Zhao, Joy Zhang for my experiment on statistical machine translation. Many results in this thesis would not be possible without their help. Special thanks to Freya Fridy, Ian Lane, Isaac Harris, Lisa Mauti, Matthias Eck, Matthias Paulik and Roger Hsiao for evaluating the quality of my translation output. Thanks to Isaac Harris for his faithful service and effective maintenance of the computing environment in our lab. Thanks to Lisa Mauti and Kelly Widmaier for making a happier workplace filled with fun parties, and Jie Yang for organizing dumpling parties during the Chinese new year celebration.

I am deeply grateful to Ciprian Chelba and Milind Mahajan for mentoring me at the speech group of Microsoft Research in summer 2003. Their passion and guidance have

had significant impact on my research direction. I am indebted to Brian Mak, my professor in Hong Kong University of Science and Technology, for his guidance in the past.

Roger Hsiao has been a faithful friend and confidant. Thank you for your friendship and discussion on my research. I enjoyed the great accompany with Fei Huang and Wen Wu as my squash and jogging partners, and Jing-Hao Fei for sharing his interesting life experience with me. Thank you for the friendship of my LTI classmates including Antoine Raux, Banerjee Satanjeev, Betty Cheng, Brian Langner, Jason Zhang, Jean Oh, June Sisson, Robert Chen, Pam Suebvisai, Peter Suen and Yifen Huang.

I am grateful to the love and support from the brothers and sisters of Antioch cell group, including Steve Sheng and Chiapi Chao, Yunghui Li and Hwaning Lee, Roger Hsiao and Rui Zhao, Eugene Ooi and Meiru Ooi, Jason Tong and Ying Liu, Steve Sun, Chiachi Chuang, Jen Lee, Kai-Zhe Huang, Francis Hong, Mike Tang and Mary Kung. Special thanks to Yunghui for bringing me into the big family, and Steve Sheng for his friendship and mentorship.

Finally, I would like to thank my family for their love and encouragement. I am very grateful to my mom for her love and patience over the past years. Thank you with all my heart!

Abbreviations

AM	Acoustic Model
ASR	Automatic Speech Recognition
BC	Broadcast Conversation
BN	Broadcast News
CER	Character Error Rate
EM	Expectation Maximization
FSA	Feature Space Adaptation
GALE	Global Autonomous Language Exploitation
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LDA	Latent Dirichlet Allocation
LDC	Linguistic Data Consortium
LDTA	Latent Dirichlet-Tree Allocation
LM	Language Model
LSA	Latent Semantic Analysis
bLSA	Bilingual Latent Semantic Analysis
LSI	Latent Semantic Indexing
pLSA	Probabilistic Latent Semantic Analysis
MAP	Maximum a Posteriori
MERT	Minimum Error Rate Training
MFCC	Mel-Frequency Cepstral Coefficient
MLE	Maximum Likelihood Estimation
MLLR	Maximum Likelihood Linear Regression
fMLLR	Feature Maximum Likelihood Linear Regression
MMIE	Maximum Mutual Information Estimation
SAT	Speaker Adaptive Training
SMT	Statistical Machine Translation
SVD	Singular Value Decomposition
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Statistical language modeling (LM) is a crucial research area that has wide applications, including automatic speech recognition (ASR) and statistical machine translation (SMT). In speech translation, an input speech utterance X is first recognized into a text F of a source language. The text F is then translated into a text E of another language. These processes can be summarized by the following Bayes decision rules:

$$\hat{F} = \arg\max_{F} p(F|X) = \arg\max_{F} \underbrace{p(X|F)}_{\text{acoustic model}} \cdot \underbrace{p(F)}_{\text{source language model}} \quad (1.1)$$

$$\hat{E} = \arg\max_{E} p(E|F) = \arg\max_{E} \underbrace{p(F|E)}_{\text{translation model}} \cdot \underbrace{p(E)}_{\text{target language model}} \quad (1.2)$$

where equation 1.1 and equation 1.2 are the decision rules for automatic speech recognition and statistical machine translation respectively. Clearly, statistical language models $p(F)$ and $p(E)$ play an important role in guiding decoding processes via pruning unlikely word hypotheses.

An effective representation of a statistical language model $p(w_1 w_2 ... w_I)$ is an N-gram

language model that makes a Markov assumption due to data sparseness:

$$p(w_1^I) = \prod_{i=1}^{I} p(w_i|w_{i-1}...w_1) \approx \prod_{i=1}^{I} p(w_i|w_{i-1}...w_{i-N+1}) \quad (1.3)$$

Usually, a 4-gram language model and a 5-gram (sometimes up to 6-gram) language model are common for automatic speech recognition and statistical machine translation respectively depending on the amount of training text. While the N-gram language model captures a local context well, it cannot capture a long-distance context due to the Markov assumption in equation 1.3. However, a long-distance topical context is useful for word prediction. For instance, if an input utterance is about "sports", the probability of an on-topic term (e.g. "basketball") is increased while the probability of an off-topic term (e.g. "economy") is de-emphasized. In broadcast news, topics can change from one story to another story. Therefore, a dynamic language model is preferable that adapts to the current word context rapidly.

1.2 Proposed Research

In this thesis, we propose a unified unsupervised topic adaptation framework that can be applied in monolingual and crosslingual fashions, such as automatic speech recognition and statistical machine translation. Our framework features rapid topic adaptation in the sense of using few unsupervised adaptation data. Not only the framework adapts a language model in automatic speech recognition monolingually, but it also adapts a language model and a translation model in statistical machine translation crosslingually using an input text from another language. Both monolingual and crosslingual adaptation are performed via a combination of unigram and N-gram latent semantic analysis.

Figure 1.1 shows our unified topic adaptation framework for speech translation that consists of three main components from left to right: automatic speech recognition, bilingual latent semantic analysis, and statistical machine translation. First a language model is adapted monolingually in automatic speech recognition, producing a word hypothesis of a source language. Bilingual latent semantic analysis bridges the gap between automatic

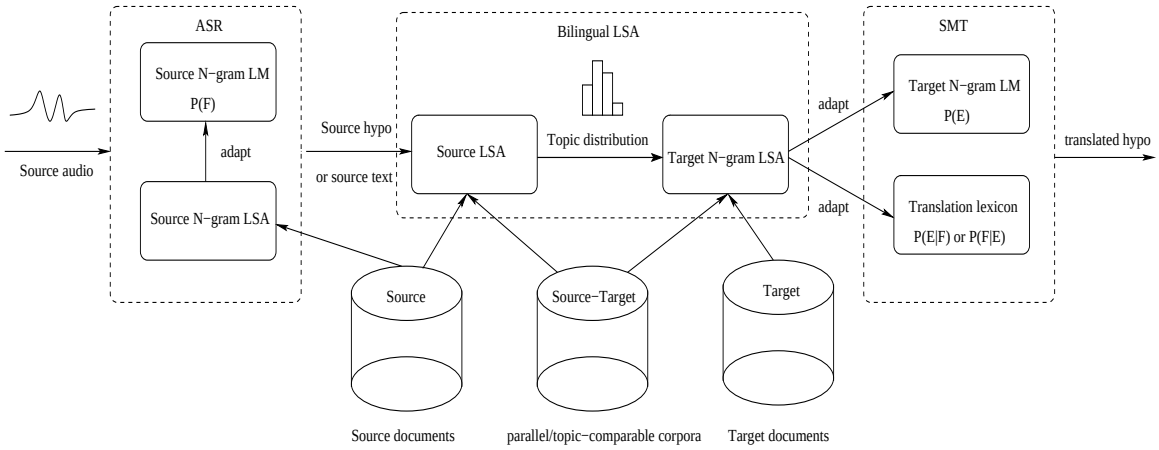

Figure 1.1: A unified topic adaptation framework for speech translation.

speech recognition and statistical machine translation via predicting and transferring the topic distribution of the word hypothesis from one language into another language so that a target language model and a translation model are adapted readily before translation.

1.2.1 Monolingual Adaptation

First, a speech recognizer decodes an input audio into its initial word hypothesis, which is used to adapt a language model. Then the adapted language model is applied to re-decode the input audio to produce a final word hypothesis. There has been a significant amount of work on unsupervised language model adaptation for automatic speech recognition. Cache-based language model (Kuhn and Mori, 1990; Clarkson and Robinson, 1997) takes advantage of a long-distance context to track the frequency of recently occurred words in an exponentially decaying N-gram cache. Word triggering approach uses words from a past context to trigger the probability of future words via the maximum entropy language modeling (Rosenfeld, 1994; Chen et al., 1998; Wu and Khudanpur, 2002). Information retrieval technique (Mahajan et al., 1999) retrieves relevant documents from background training corpora to build an in-domain N-gram language model.

Closely related research to this thesis include latent semantic analysis based language model adaptation, such as singular value decomposition (Deerwester et al., 1990; Bel-

legarda, 2000; Kim, 2004; Bellegarda, 2005), and probabilistic latent semantic analysis (Gildea and Hofmann, 1999; Mrva and Woodland, 2004; Akita and Kawahara, 2004). Our first language modeling work is based on latent Dirichlet allocation (Blei et al., 2003), which is a Bayesian approach for latent semantic analysis. First, we draw connections between latent Dirichlet allocation and the cache-based language model as topic caching (Tam and Schultz, 2005). Instead of caching the frequency of the recently occurred words, we cache the frequency of the recently occurred topics. Related works on latent Dirichlet allocation for language model adaptation have been investigated by other researchers based on hidden Markov model (Hsu and Glass, 2006), another topic inference algorithm (Heidel et al., 2007), linear transformation (Chien and Chueh, 2008) and name entity (Liu and Liu, 2008).

The in-domain unigram language model generated from latent semantic analysis can be integrated into a background language model via linear interpolation or marginal adaptation (Kneser et al., 1997). To make the marginal adaptation computationally less expensive, we investigate an incremental LSA-marginal adaptation for lattice rescoring that avoids manipulating a large background N-gram language model.

In the machine learning community, modeling topic correlation (Blei and Lafferty, 2005; Li and McCallum, 2006) has shown better performance on word perplexity compared to latent Dirichlet allocation that makes the topic independence assumption. We propose latent Dirichlet-Tree allocation (Tam and Schultz, 2007b), a tree-based latent semantic model to capture topic correlation. Our model generalizes latent Dirichlet allocation using a tree-based probabilistic prior with comparable complexity of the training algorithm driven by a variational Expectation-Maximization procedure.

Latent semantic analysis makes the "bag-of-word" assumption that ignores the word ordering in a document. To relax this assumption, we propose N-gram latent Dirichlet-Tree allocation to model the word ordering and the topical information simultaneously (Tam and Schultz, 2008). For rapid model training, we bootstrap N-gram latent semantic analysis using a well-trained latent semantic analysis that are based on unigrams. For better smoothing, we investigate a fractional Kneser-Ney smoothing algorithm that handles fractional counts and generalizes the original formulation based on integral counts (Kneser

and Ney, 1995). The sentence-level topic mixtures (Iyer and Ostendorf, 1999) is another approach that is based on a mixture of topic-dependent N-gram language models but with a different modeling assumption from N-gram latent semantic analysis.

1.2.2 Crosslingual Adaptation

The idea of crosslingual adaptation is to exploit information in one language to adapt models in another language. For instance, crosslingual information retrieval first uses a decoded word hypothesis as an input query to retrieve relevant documents on another language (Eck et al., 2004; Zhao et al., 2004). The foreign retrieved documents are transformed back to the original language to train an in-domain unigram language model. The transformation approaches includes statistical machine translation, crosslingual word triggers, and crosslingual latent semantic analysis using singular value decomposition (Kim, 2004). The in-domain unigram language model is applied to improve the performance of automatic speech recognition and statistical machine translation via language model adaptation for a resource-deficient language. In a co-operative multilingual speech translation where a human translator is involved, (Paulik et al., 2005b) employs statistical machine translation to translate an initial word hypothesis of a speech decoder into a target language to improve the recognition performance of the target language via language model adaptation.

The second component in Figure 1.1 is the proposed bilingual latent semantic analysis to facilitate crosslingual adaptation for statistical machine translation. Since topical information are language independent, the topic distribution of a parallel document pair is assumed identical. Therefore, bilingual latent semantic analysis is trained so that a one-to-one topic correspondence between a source and a target language is enforced using parallel document corpora. During adaptation, the topic distribution of an input document is inferred using latent semantic analysis of a source language. Then the inferred topic distribution is transferred to a target language so that an in-domain unigram/N-gram language model is generated via linearly interpolating topic-dependent unigram/N-gram language models of the target language. The proposed approach has two advantages: first,

it can be applied before translation, and thus has immediate impact on the translation output. Secondly, it does not rely on an initial translation output for adaptation, and therefore does not suffer from translation errors. In other words, statistical machine translation is not required for crosslingual adaptation.

It is motivated from an observation that a source word can be translated into different target words depending on a topical context. One popular example is the word "bank" that can be related to either a "financial bank" or a "river bank". Bilingual Topic Admixture Model for word alignment has been proposed to address this issue via explicit modeling of topic-dependent translation lexicons (Zhao and Xing, 2006). We address the same issue via an adaptation approach so that probabilities in a background translation word lexicon are adapted towards an input document without explicit modeling of the topic-dependent translation lexicons. The scores of phrase pairs in a phrase table are rescored using the adapted translation lexicons so that the scores are sensitive to the topical context of the input document.

After the target language model, translation lexicon and phrase table are adapted towards an input document, the sentences of the input document are translated.

One limitation of bilingual latent semantic analysis is the requirement of parallel documents for model training. Collecting parallel documents are relatively expensive compared to monolingual non-parallel documents. In addition, monolingual non-parallel documents have better topic and vocabulary coverage than parallel documents, especially in a resource deficient scenario where parallel resources are scarce. Therefore, it is attractive to incorporate non-parallel documents into bilingual latent semantic analysis. Previous research includes an extension of bilingual singular value decomposition where a monolingual document is treated as a pseudo-parallel document by filling zeros into the missing entries of a bilingual document vector (Kim, 2004). We employ a semi-supervised approach to incorporate monolingual non-parallel documents via a notion of parallel clusters formed from the non-parallel documents. The parallel clusters are served as constraints for optimization in bilingual latent semantic analysis.

1.3 Thesis Organization

In Chapter 2, we cover the background materials relevant to the development of the thesis starting with variational Bayes, a useful variational inference technique for graphical models including latent Dirichlet allocation. We review different approaches for latent semantic analysis from the traditional vector-space approach to the modern Bayesian approach that enables an integration of prior knowledge into the model. We introduce higher-order models for latent semantic analysis and topic modeling, techniques for language model smoothing, and unsupervised language model adaptation.

In Chapter 3, we describe our baseline Mandarin and Arabic transcription systems and our Chinese-English statistical machine translation systems of different implementation scales.

In Chapter 4, we present our topic adaptation framework for automatic speech recognition via N-gram latent semantic analysis. Our approaches include topic caching for decoding, incremental marginal adaptation for lattice rescoring, latent Dirichlet-Tree allocation for tree-based latent semantic analysis, and N-gram latent Dirichlet-Tree allocation to relax the "bag-of-word" assumption. We evaluate our approaches on large-scale GALE evaluations on two languages: Mandarin and Arabic.

In Chapter 5, we extend our topic adaptation framework to crosslingual adaptation for statistical machine translation via bilingual latent semantic analysis. We adapt a target language model, translation lexicon and a phrase table using an input source document. In addition, we apply N-gram latent semantic analysis on a target language. We evaluate our approaches on text translation and end-to-end speech translation using state-of-the-art statistical machine translation systems. To tackle the limitation of bilingual latent semantic analysis, we employ a semi-supervised approach to integrate monolingual non-parallel documents for model training. We evaluate this approach in a simulated scenario where parallel resources are deficient.

In Chapter 6, we summarize our contributions and propose possible future extensions.

In Appendix A, we describe an alternative approach for training latent Dirichlet allo-

cation and bigram topic model using the Gibbs sampling.

In Appendix B, we include a complete mathematical derivation for latent Dirichlet-Tree allocation.

Chapter 2

Background

This chapter covers the basics and related works of the thesis including variational Bayes, approaches for latent semantic analysis, language model smoothing and language model adaptation.

2.1 Variational Bayes

Variational Bayes (Jordan et al., 1999; Bishop et al., 2003) is a powerful technique for approximate inference in a directed graphical model. This approach has been applied to different applications such as latent Dirichlet allocation, which is a graphical model for Bayesian latent semantic analysis. The solution of a variational posterior distribution has a generic form for any directed graphical model that makes variational Bayes attractive and useful.

Given a joint probability distribution $p(X, Z; \Lambda)$ over latent variables $Z = \{z_j\}$ and observed variables $X = \{x_i\}$ parametrized by Λ, the Expectation-Maximization algorithm can be employed to maximize the lower-bound of the log likelihood $L(X; \Lambda)$ on the

observation using the Jensen's inequality as follows:

$$L(X;\Lambda) = \log \sum_{Z} p(Z|X;\Lambda^{(t-1)}) \cdot \frac{p(X,Z;\Lambda)}{p(Z|X;\Lambda^{(t-1)})} \quad (2.1)$$

$$\geq \sum_{Z} p(Z|X;\Lambda^{(t-1)}) \log \frac{p(X,Z;\Lambda)}{p(Z|X;\Lambda^{(t-1)})} \quad (2.2)$$

$$= E_p[\log \frac{p(X,Z;\Lambda)}{p(Z|X;\Lambda^{(t-1)})}] = Q(X,\Lambda;\Lambda^{(t-1)}) \quad (2.3)$$

where the expectation is taken using the exact posterior distribution computed from $\Lambda^{(t-1)}$ from the (t-1) iteration. Computing the exact posterior distribution over the latent variables can be non-trivial and sometimes intractable. Using the Bayes rule, the posterior distribution over the latent variables Z can be computed as follows:

$$p(Z|X;\Lambda^{(t-1)}) = \frac{p(X,Z;\Lambda^{(t-1)})}{\sum_{Z'} p(X,Z';\Lambda^{(t-1)})} \quad (2.4)$$

where the intractable part is the normalization term involving all possible assignments of the latent variables Z.

In variational Bayes, instead of computing the exact posterior distribution over Z directly, a variational posterior distribution $q(Z|X;\Gamma)$ is introduced to approximate the true posterior distribution. A convenient factorizable distribution is employed so that the latent variables are independent given an observation X:

$$q(Z|X;\Gamma) = \prod_{j} q(z_j|X;\Gamma_j) = \prod_{j} q(z_j) \quad (2.5)$$

The independence assumption is a trade off between simplicity and accuracy of the posterior inference over the latent variables. After replacing $p(Z|X;\lambda^{(t-1)})$ with $q(Z|X;\Gamma)$ in equation 2.3, the auxiliary function using variational Bayes has the following form:

$$Q_{vb}(X;\Lambda,\Gamma) = E_q[\log \frac{p(X,Z;\Lambda)}{q(Z|X;\Gamma)}] \quad (2.6)$$

$$= E_q[\log p(X,Z;\Lambda)] - \sum_{j} E_q[\log q(z_j)] \quad (2.7)$$

where the variational parameters Γ are determined via iterative E-steps over each latent variable. By computing the partial derivative of the auxiliary function with respect to each $q(z_j)$, the generic E-steps can be derived:

$$\frac{\partial Q_{vb}(X;\Lambda,\Gamma)}{\partial q(z_j)} = E_q[\log p(X,Z;\Lambda)]_{Z\setminus z_j} - \log q(z_j) + \text{constant} = 0 \quad (2.8)$$

$$\Longrightarrow q(z_j) \propto e^{E_q[\log p(X,Z;\Lambda)]_{Z\setminus z_j}} \quad (2.9)$$

where the expectation is taken over all other latent variables $\{z_i\}$ excluding the current variable z_j. Instead of considering the full joint distribution for expectation in equation 2.9, only a subset of conditional distributions involving z_j and its Markov blankets (i.e. parent nodes, child nodes and their co-parents) are needed while the other $q(z_i)$ are kept fixed. For a hidden Markov model with s_t and x_t being the hidden state and the observation at time t respectively, the variational E-steps for $q(s_t)$ can be derived easily using equation 2.9:

$$q(s_t) \propto e^{E_q[\log p(s_t|s_{t-1})]_{\setminus s_t} + E_q[\log p(x_t|s_t)]_{\setminus s_t} + E_q[\log p(s_{t+1}|s_t)]_{\setminus s_t}} \quad \forall t \quad (2.10)$$

$$\propto e^{\sum_i q(s_{t-1}=i)\log p(s_t|s_{t-1}=i) + \log p(x_t|s_t) + \sum_i q(s_{t+1}=i)\log p(s_{t+1}=i|s_t)} \quad \forall t \quad (2.11)$$

where the Markov blankets of s_t are s_{t-1} and s_{t+1}.

The variational E-steps actually minimize the Kullback-Leibler divergence $KL(q||p|X)$ between the variational posteriors and the true posteriors since the difference between the log likelihood and the auxiliary log likelihood is:

$$\log p(X) - E_q[\log \frac{p(X,Z)}{q(Z|X)}] = \sum_Z q(Z|X)\log p(X) - E_q[\log \frac{p(X,Z)}{q(Z|X)}] \quad (2.12)$$

$$= \sum_Z q(Z|X)\log\left(q(Z|X)\cdot\frac{p(X)}{p(X,Z)}\right) \quad (2.13)$$

$$= \sum_Z q(Z|X)\log\frac{q(Z|X)}{p(Z|X)} \quad (2.14)$$

$$= KL(q||p|X) \quad (2.15)$$

This implies that maximizing the auxiliary function using variational Bayes leads to minimizing the KL divergence. After the E-steps, the M-step is usually straightforward via a weighted maximum likelihood estimate of the model parameters using the variational posteriors to re-weight the observations.

2.2 Latent Semantic Analysis

Latent semantic analysis (LSA) is an unsupervised technique to find a set of patterns to describe a data set. Different approaches based on a vector space model or a probabilistic model have been developed and applied to various areas including image and text modeling.

2.2.1 Latent Semantic Indexing

Latent semantic indexing (LSI) (Deerwester et al., 1990) is a useful technique for document indexing for semantic information retrieval. As a vector space model, latent semantic indexing searches for a set of basis vectors for document representation. Each basis vector represents a "semantic" dimension. To determine the basis vectors, a term-document matrix W_{VM} is formed by packing each training document as a column vector in W_{VM}, where M is the number of training documents. Each component in a document vector corresponds to a term frequency or other variants such as the popular TF.IDF, which is the multiplication of a term frequency with an inverse document frequency. Singular value decomposition is applied to decompose W_{VM} into three matrices: $W_{VM} = U_{VK} \cdot S_{KK} \cdot V_{MK}^T$. The matrix U contains K basis vectors spanning the latent semantic space. S is a diagonal matrix containing the corresponding eigenvalue of each basis vector. Each column of V^T represents the "coordinate" of a training document in the K-dimensional latent semantic space.

Latent semantic indexing does not provide a natural way to compute the probability of a document. Moreover, there is a question about the validity of latent semantic indexing since singular value decomposition minimizes the least-square error of training documents assuming that the document vectors are normally distributed. However, each document vector has non-negative word counts violating the Gaussian assumption.

Latent semantic indexing can be applied for language model adaptation (Bellegarda, 2000). To obtain the probability of a word w given a "bag-of-word" history h, a unigram

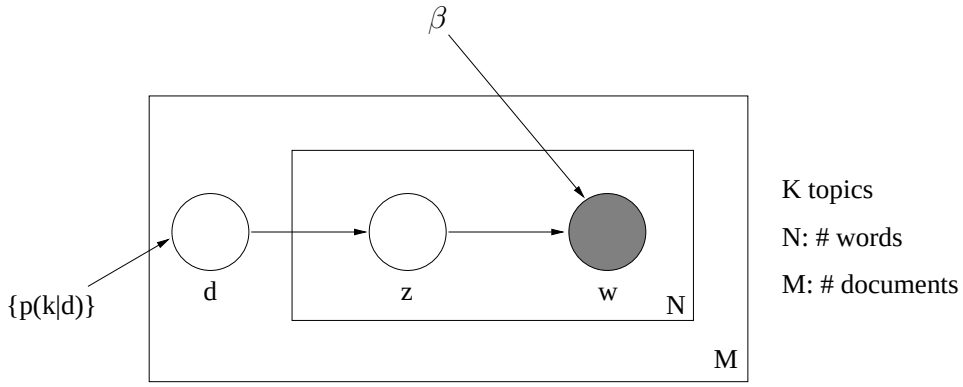

Figure 2.1: Graphical representation of probabilistic latent semantic analysis.

language model can be generated from latent semantic indexing as follows:

$$p_{lsi}(w|h) \quad = \quad \frac{sim(w,h)^{\gamma}}{\sum_{w'} sim(w',h)^{\gamma}} \tag{2.16}$$

where γ is a tuning factor $\gg 1$ and $sim(w,h)$ defines the cosine similarity between w and h in the latent semantic space.

2.2.2 Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (pLSA) (Gildea and Hofmann, 1999) is a significant step towards probabilistic models from the vector space model in latent semantic indexing. The graphical model structure is shown in Figure 2.1 with the unshaded circles denoting the latent variables: a document index d and topic labels z. Each training document d is associated with a document-level topic distribution $p(k|d)$. The document generation procedure is defined as follows:

1. Sample a document index d to retrieve a document-level topic distribution $p(k|d)$.
2. For each word w_i in a document $w_1^{N_d}$,
   - Sample a latent topic index z_i from $p(k|d)$.
   - Sample w_i from $p(w|z_i)$ (denoted as $\beta_{w_i z_i}$).

The generative procedure defines the joint distribution $p(d, w_1^{N_d}, z_1^{N_d})$ over the document index d, topic assignment $z_1^{N_d}$ and the document $w_1^{N_d}$. With M training documents, the marginal likelihood can be obtained by marginalizing out d and $z_1^{N_d}$:

$$p_{plsi}(\{w_1^{N_d}\}) = \sum_{d=1}^{M} p(d) \prod_{i=1}^{N_d} \sum_{k=1}^{K} p(w_i|z_i = k) \cdot p(z_i = k|d) \tag{2.17}$$

The model parameters include the document-specific topic distribution $p(k|d)$ and the topic-dependent unigram language model $p(w|k)$ denoted as β_{wk}. Since the number of model parameters for $p(k|d)$ grows proportional to the number of document, overfitting is reported and therefore an annealed E-step is required to prevent overfitting (Gildea and Hofmann, 1999). The E-steps and the M-step are given as follows:

E-step:

$$p^{(t+1)}(z_i = k|d) \propto p^{(t)}(w_i|z_i = k) \cdot p^{(t)}(k|d) \tag{2.18}$$

where the word-level topic posteriors $p(z_i = k|d)$ are re-estimated.

M-step:

$$p^{(t+1)}(k|d) \propto \sum_{i=1}^{N} p^{(t+1)}(z_i = k|d) \tag{2.19}$$

$$p^{(t+1)}(w|k) \propto \sum_{d=1}^{M} \sum_{i=1}^{N_d} p^{(t+1)}(z_i = k|d) \cdot \delta(w, w_i|d) \tag{2.20}$$

where the document-level topic posterior $p(k|d)$ and the topic-dependent unigram language models $p(w|k)$ are re-estimated. N_d denotes the number of word tokens in the document d.

The document generation procedure for a test document is not well defined since the document-level topic posterior of a test document is unknown since it is not indexed by the latent variable d. But in practice, a folding-in procedure can be performed to include the test document. In other words, the document-level topic posterior $p(k|d)$ of the test document can be re-estimated via the Expectation-Maximization algorithm. For

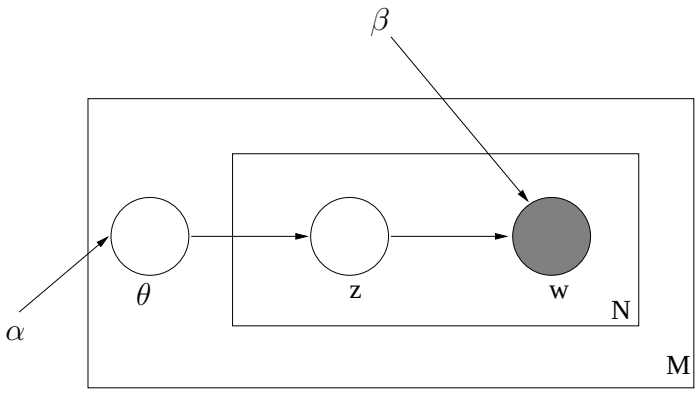

Figure 2.2: Graphical representation of latent Dirichlet allocation.

language model adaptation (Gildea and Hofmann, 1999; Mrva and Woodland, 2004; Akita and Kawahara, 2004), a unigram language model can be computed via linear interpolation:

$$p_{plsa}(w|h) = \sum_{k=1}^{K} p(w|k) \cdot p(k|h) \tag{2.21}$$

where h denotes a "bag-of-word" history that is treated as a "document". $p(k|h)$ is the result of the EM iterations until convergence is reached with $p(w|k)$ kept fixed. This approach avoids the expensive normalization step required in equation 2.16 in latent semantic indexing.

2.2.3 Latent Dirichlet Allocation

To remedy the deficiency of document generation in probabilistic latent semantic analysis, latent Dirichlet allocation (Blei et al., 2003) has been proposed. Instead of sampling a document index to retrieve the corresponding topic distribution $p(k|d)$ in Figure 2.1, a Dirichlet prior over a topic distribution θ is employed so that the topic distribution of a test

document can be sampled from a prior Dirichlet distribution:

$$\text{Dirichlet}(\theta;\{\alpha_k\}) = \frac{1}{A(\alpha)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} = e^{\sum_{k=1}^{K}(\alpha_k-1)\log\theta_k-\log A(\alpha)} \quad (2.22)$$

$$\text{where } A(\alpha) = \frac{\prod_{k'=1}^{K}\Gamma(\alpha_{k'})}{\Gamma(\sum_{k'=1}^{K}\alpha_{k'})} \quad (2.23)$$

$$\text{and } \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \quad (2.24)$$

In equation 2.22, the Dirichlet distribution is rewritten in an exponential family since computing the expectation of the sufficient statistics $\{\text{E}[\log\theta_k]\}$ is required in the variational E-steps later. Figure 2.2 illustrates the graphical representation of latent Dirichlet allocation with the topic distribution θ drawn from a Dirichlet prior parametrized by $\{\alpha_k\}$, which denotes the "pseudo-count" of topic k informally. Therefore, the model parameters Λ are $\{\alpha_k\}$ and $\{p(w|k)\}$ denoted as $\{\beta_{wk}\}$. The document generation of latent Dirichlet allocation is described as follows:

1. Sample θ from Dirichlet$(\theta;\alpha)$ *(different from pLSA)*
2. For each word w_i in a document w_1^N,
   - Sample a latent topic index z_i from θ.
   - Sample w_i from $p(w|z_i)$ (denoted as β_{wz_i}).

Variational EM (Blei et al., 2003) can be employed for model estimation. Appendix A provides an alternative approach based on collapsed Gibbs sampling for model estimation (Griffiths and Steyvers, 2004; Porteous et al., 2008). On the other hand, the E-step formulae can be seen directly using the generic form of variational EM introduced in Section 2.1. From the graphical representation, the Markov blanket of the latent variables θ and z_i are $\{z_1...z_N\}$ and $\{\theta, w_i\}$ respectively. Using the generic form of variational EM in equation 2.9, the variational E-steps involving the latent variables θ and z_i are shown as follows:

E-steps:

$$
\begin{aligned}
q(\theta) &\propto e^{\sum_{i=1}^{N} E_q[\log \theta_{z_i}] + E_q[\log p(\theta)]_{\setminus \theta}} && (2.25)\\
&\propto e^{\sum_{i=1}^{N}\sum_{k=1}^{K} q(z_i=k)\log\theta_k + \log p(\theta)} && (2.26)\\
&\propto p(\theta)\prod_{k=1}^{K}\prod_{i=1}^{N}\theta_k^{q(z_i=k)} && (2.27)\\
&\propto \prod_{k=1}^{K}\theta_k^{\alpha_k-1}\prod_{k=1}^{K}\prod_{i=1}^{N}\theta_k^{q(z_i=k)} && (2.28)\\
&\propto \prod_{k=1}^{K}\theta_k^{\alpha_k+\sum_{i=1}^{N} q(z_i=k)-1} && (2.29)\\
&= \text{Dirichlet}(\theta;\{\gamma_k\})\ , && (2.30)\\
\text{where } \gamma_k &= \alpha_k + \sum_{i=1}^{N} q(z_i=k)\ , && (2.31)
\end{aligned}
$$

$$
\begin{aligned}
q(z_i=k) &\propto e^{E_q[\log\theta_k]+\log p(w_i|z_i)}\ , && (2.32)\\
\text{where } E_q[\log\theta_k] &= \Psi(\gamma_k) - \Psi(\sum_{k'=1}^{K}\gamma_{k'})\ , && (2.33)\\
\text{and } \Psi(\gamma_k) &= \frac{\partial \log\Gamma(\gamma_k)}{\partial\gamma_k} && (2.34)
\end{aligned}
$$

Since the Dirichlet prior can be expressed as an exponential family, the expected sufficient statistics E[$\log\theta_k$] can be computed as the first derivative of the log partition function $\log A(\alpha)$ with respect to α_k giving the result in equation 2.33. Equation 2.31 and equation 2.32 are executed iteratively until convergence is reached by measuring the relative change of norm of $\{\gamma_k\}$ between successive iterations. Equation 2.31 can be understood intuitively by the fact that Dirichlet prior and multinomial distribution are a conjugate pair. Thus, the variational posterior distribution $q(\theta;\{\gamma_k\})$ is also a Dirichlet distribution where the posterior counts γ_k of topic k is re-estimated by accumulating the word-level topic posteriors $q(z_i=k)$ of a document plus the prior pseudo-count α_k.

M-step:

$$
p(w|k) \propto \sum_{d=1}^{M}\sum_{i=1}^{N_d} q(z_i=k|d)\cdot\delta(w_i,w|d) \qquad (2.35)
$$

Parameters of a Dirichlet prior $\{\alpha_k\}$ can be re-estimated using gradient ascent in the $\log$ space of $\{\alpha_k\}$ to ensure that the final values are larger than zero. We first perform parameter transformation using $\log(.)$: $\tilde{\alpha_k} = \log \alpha_k$. Then we rewrite the auxiliary function as $Q(\tilde{\alpha_k})$ and perform the gradient ascent as follows:

$$\tilde{\alpha}_k^{(t+1)} = \tilde{\alpha}_k^{(t)} + \rho^{(t)} \cdot \frac{\partial Q(\tilde{\alpha})}{\partial \tilde{\alpha_k}} \tag{2.36}$$

$$= \tilde{\alpha}_k^{(t)} + \rho^{(t)} \cdot \frac{\partial Q(\alpha)}{\partial \alpha_k} \cdot \frac{\partial \alpha_k}{\partial \tilde{\alpha_k}} \tag{2.37}$$

$$= \tilde{\alpha}_k^{(t)} + \rho^{(t)} \cdot \frac{\partial Q(\alpha)}{\partial \alpha_k} \cdot \alpha_k^{(t)} \tag{2.38}$$

where $\rho^{(t)}$ denotes the learning rate at iteration t. After the gradient ascent procedure finishes, we exponentiate $\tilde{\alpha_k}$ to obtain the final α_k. Latent Dirichlet allocation makes an independence assumption over the topics due to the Dirichlet prior. In Chapter 4, we will describe latent Dirichlet-Tree allocation so that topic correlation is captured.

2.2.4 Correlated Topic Model

Correlated topic model (Blei and Lafferty, 2005) is an extension of the latent Dirichlet allocation to model topic correlation. Their approach replaces a Dirichlet prior with a logistic-normal prior. The document generation procedure is shown as follows:

1. Sample η from a multivariate Normal distribution $N(\mu, \Sigma)$ of dimension K.
2. For each word w_i in a document w_1^n,
   - Sample a latent topic index z_i from a Multinomial(f(η)).
   - Sample w_i from $p(w|z_i)$.

where f(η) is a logistic normal distribution normalized between 0 and 1:

$$f(\eta_k) = \frac{e^{\eta_k}}{\sum_{k'=1}^{K} e^{\eta_{k'}}} \tag{2.39}$$

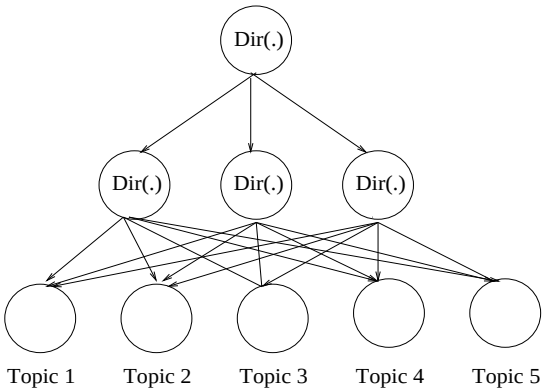

Figure 2.3: Pachinko allocation employs a directed acyclic graph of Dirichlet nodes as a topic prior.

The logistic normal distribution assumes that η is normally distributed and then mapped to a topic distribution. Topic correlations are thus modeled through the covariance matrix of the Normal distribution. Correlated topic model enables modeling pair-wise topic correlation. However, the non-conjugate logistic normal prior poses complication on variational EM due to the computation of the expected log probability of a topic assignment $E_q[\log f(\eta_k)]$ that does not have a closed-form solution. Taylor expansion is employed to approximate this term for variational inference.

2.2.5 Pachinko Allocation

Pachinko allocation (Li and McCallum, 2006) is another extension of latent Dirichlet allocation for modeling topic correlation. Their approach replaces a Dirichlet prior with a direct-acyclic Dirichlet graph where each node in the graph is modeled as a Dirichlet distribution over the outgoing links which connects to other nodes in a top-down and fully connected fashion as shown in Figure 2.3. There are K nodes at the bottom layer of the model to denote the topics. Therefore, probability of a topic $p(z|\theta)$ is computed as product of branching probabilities from a root node to a leaf node analogous to a popular Japanese Pachinko game where a ball follows a random path from the top to the bottom. Pachinko allocation can be interpreted as generalization of latent Dirichlet-Tree allocation that we

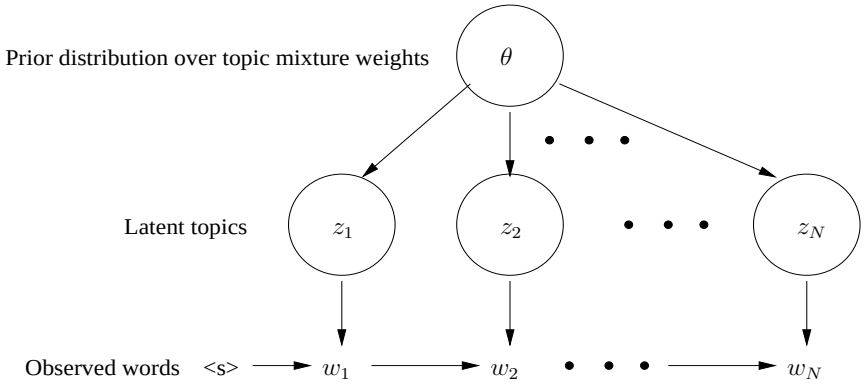

Figure 2.4: Graphical representation of bigram LSA. Adjacent words in a document are linked together to form a Markov chain from left to right.

have proposed independently on different applications. However, Pachinko allocation is more expensive in nature since the structure of the Dirichlet graph is undefined. Moreover, the model requires the Gibbs sampling procedure for model estimation while we present a variational Bayes approach for efficient model estimation.

2.2.6 Bigram Latent Semantic Analysis

Latent semantic analysis makes a "bag-of-word" assumption that word ordering is ignored. For document classification, word ordering may not be important. But from the language modeling perspective, word ordering is crucial since a trigram language model usually outperforms a unigram language model for word prediction. To relax the "bag-of-word" assumption, bigram LSA has been proposed (Wallach, 2006). The model modifies the graphical structure of latent Dirichlet allocation by connecting adjacent words together to form a Markov chain. Figure 2.4 shows the graphical representation of bigram LSA where the top node represents a prior distribution over topic distributions and the middle layer represents topic labels associated with each word. The document generation procedure of bigram LSA is similar to that of latent Dirichlet allocation except that a previous word is taken into consideration for generating a current word:

1. Sample θ from a Dirichlet prior $p(\theta)$.

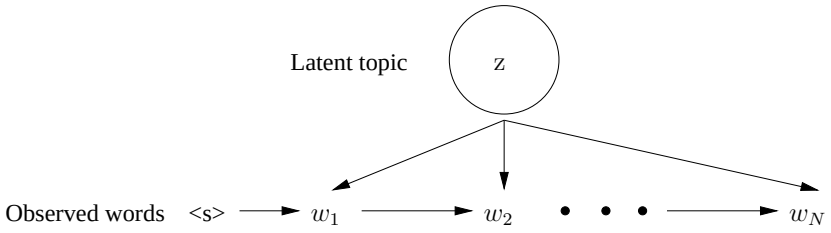

Figure 2.5: Graphical representation of the sentence-level bigram topic mixture model.

2. For each word w_i at the i-th position of a document:

   (a) Sample a topic label: $z_i \sim$ Multinomial(θ).

   (b) Sample w_i given w_{i-1} and the topic label z_i: $w_i \sim p(\cdot|w_{i-1}, z_i)$.

In (Wallach, 2006), the Gibbs sampling procedure is applied for model training, which can be very slow since it requires drawing a significant number of samples for convergence. Usually 500 Gibbs iterations are common in practice for latent Dirichlet allocation (Porteous et al., 2008). Therefore, this approach is difficult to scale up to large training data. Moreover, simple Laplace smoothing is employed which can give poor model smoothing. We address these issues using model bootstrapping and language model smoothing in this thesis, and show the effectiveness for topic adaptation in automatic speech recognition in Chapter 4 and statistical machine translation in Chapter 5.

2.2.7 Sentence-Level Topic Mixtures

Sentence-level topic mixture (Iyer and Ostendorf, 1999) is similar in spirit to bigram LSA that the latent topics can be modeled via a mixture of N-gram language models. However, their model structures are different. They assume that all words in a sentence are assigned with the same topic label, which differs from bigram LSA that each word is assigned with an independent topic label. Moreover, each sentence is assumed to be independent within a document while bigram LSA captures the sentence dependency via a probabilistic prior. Figure 2.5 illustrates the graphical model representation of the sentence-level topic mixture. Specifically, the probability of a word sequence w_1^N using a bigram topic mixture

is given as follows:

$$p(w_1^N) = \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_k \cdot p(w_i|w_{i-1}, k) \tag{2.40}$$

A "hard" document/sentence clustering is employed to train the cluster-specific N-gram language models for model initialization. For instance, each document/sentence can be represented as a vector with each component represented as TF.IDF (term frequency multiplied by inverse document frequency). Then the K-means algorithm can be performed to form the clusters. This follows an Expectation-Maximization procedure to re-estimate the model parameters as follows:

E-steps:

$$p^{(t)}(z = k|w_1^{N_s}) \propto p^{(t)}(w_1^{N_s}|k) \cdot \lambda_k^{(t)} \tag{2.41}$$

M-step:

$$p_{ML}^{(t+1)}(v|u, k) \propto C_k^{(t)}(u, v) \tag{2.42}$$

$$\lambda_k^{(t+1)} \propto \sum_s p^{(t)}(z = k|w_1^{N_s}) \tag{2.43}$$

$$\text{where } C_k^{(t)}(u, v) = \sum_s C_k^{(t)}(u, v|s) \tag{2.44}$$

$$\text{and } C_k^{(t)}(u, v|s) = C^{(t)}(u, v|s) \cdot p^{(t)}(z = k|w_1^{N_s}) \tag{2.45}$$

where s denotes the sentence index; $lambda_k$ denotes the topic mixture weights; $C(u, v|s)$ denotes the bigram count of sentence s and $C_k(u, v|s)$ denotes the fractional bigram count assigned to topic k in sentence s. To avoid a zero probability for an unseen bigram, the bigram model is linearly interpolated with a unigram model in a context dependent fashion as follows:

$$p(v|u, k) = (1 - \phi_{uk}) \cdot p_{ML}(v|u, k) + \phi_{uk} \cdot p_{ML}(v|k) \tag{2.46}$$

$$\text{where } \phi_{uk} = \frac{N_k(u, \cdot)}{N_k(u, \cdot) + C_k(u, \cdot)} \tag{2.47}$$

$$\text{and } N_k(u, \cdot) = \sum_{v:(u,v)} \frac{C_k(u, v)}{C(u, v)} = \sum_{v:(u,v)} p(k|u, v) \tag{2.48}$$

where ϕ_{uk} is estimated using an analogous version of Witten-Bell smoothing for fractional counts. $N_k(u, \cdot)$ denotes the fractional number of "unique" words following word u for topic k which falls back to the integral definition when the mixture model has only one topic. In Chapter 4, we will describe fractional Kneser-Ney smoothing which supports fractional counts for bigram LSA.

2.3 Bilingual Latent Semantic Analysis

The interest of extending latent semantic analysis from monolingual to crosslingual manner comes from the advantage of exploiting information from one language and applying them to another language. This notion applies naturally to statistical machine translation.

2.3.1 Bilingual Latent Semantic Indexing

Bilingual latent semantic indexing (Kim and Khudanpur, 2004) has been proposed to capture a joint latent semantic space of a source and target language via singular value decomposition. In their approach, each of a parallel document pair are concatenated into a super-document vector for singular value decomposition. Their approach creates a LSA-based translation word lexicon described as follows:

$$p_{lsi}(f|e) = \frac{sim(f,e)^{\gamma}}{\sum_{f'} sim(f',e)^{\gamma}} \tag{2.49}$$

where $\gamma >> 1$ as suggested in (Coccaro and Jurafsky, 1998). Incorporation of monolingual documents for bilingual latent semantic indexing is done by filling in zeros for the unknown components of a pseudo-bilingual document vector before singular value decomposition. However, this approach may undermine the basis vectors because of the unjustified zero co-occurrence counts between source and target words that may mislead the result of singular value decomposition.

2.3.2 Bilingual Topic Admixture Model

Bilingual Topic Admixture Model (Zhao and Xing, 2006) (BiTAM) has been proposed to incorporate latent topics into word alignment via topic-dependent translation lexicons. HM-BiTAM (Zhao and Xing, 2007) incorporates a hidden Markov model into BiTAM for word alignment. Among the three proposed variants in BiTAM, BiTAM-3 has yielded the best word alignment accuracy. The generative procedure of a sentence pair is described as follows:

1. Sample topic mixture weights θ from a Dirichlet prior.

2. For each word f_j in a source sentence f_1^J,

   - Sample a latent topic index z_j from Multinomial(θ).
   - Sample an alignment variable $a_j \in [0, I]$ to index a word e_{a_j} in a target sentence e_0^I where e_0 denotes a NULL word.
   - Sample f_j from a topic-dependent translation lexicon $p(f_j|e_{a_j}, z_j)$.

BiTAM can be trained using variational EM. Better results are obtained using the IBM-4 translation lexicon as an initial model. The graphical model representation of BiTAM is illustrated in Figure 2.6. Using the generic form of variational EM in equation 2.9, one can identify the Markov blanket of the latent variables θ, z_j, and a_j to be $\{z_1 ... z_J\}$, $\{\theta, f_j, e_{a_j}\}$ and $\{z_j, f_j, e_{a_j}\}$ respectively. Therefore, the variational E-steps and M-step for a sentence pair can be shown as below:

E-steps:

$$\gamma_k = \alpha_k + \sum_{j=1}^{J} q(z_j = k) \quad (2.50)$$

$$q(z_j = k) \propto e^{E_q[\log\theta;\gamma_k] + E_q[\log p(f_j|e_{a_j},z_j)]_{\setminus z_j}} \quad (2.51)$$

$$q(a_j = i) \propto e^{\log p(a_j) + E_q[\log p(f_j|e_{a_j},z_j)]_{\setminus a_j}} \quad (2.52)$$

$$\text{where } E_q[\log p(f_j|e_{a_j},z_j)]_{\setminus z_j} = \sum_{i=0}^{I} q(a_j = i)\cdot \log p(f_j|e_i, z_j = k) \quad (2.53)$$

$$\text{and } E_q[\log p(f_j|e_{a_j},z_j)]_{\setminus a_j} = \sum_{k=1}^{K} q(z_j = k)\cdot \log p(f_j|e_i, z_j = k) \quad (2.54)$$

$$\text{and } E_q[\log\theta_k] = \Psi(\gamma_k) - \Psi(\sum_{k'=1}^{K}\gamma_{k'}) \quad (2.55)$$

M-step:

$$p(f|e,k) \propto \sum_{j=1}^{J}\sum_{i=0}^{I} \delta(f_j,f)\cdot\delta(e_i,e)\cdot q(z_j = k)\cdot q(a_j = i) \quad (2.56)$$

where $q(\theta;\{\gamma_k\})$, $q(a_j)$ and $q(z_j)$ are the variational posterior distributions for the topic mixture weights, word alignment index and topic index for each sentence pair $\{e_1^I, f_1^J\}$. Although improvement in BLEU has been reported, the model training is computationally expensive. In Chapter 5, we will present a marginal adaptation approach so that latent topics can be incorporated into a background translation word lexicon without the explicit modeling of topic-dependent lexicons.

2.4 Language Model Smoothing

Language model smoothing is essential to avoid assigning zero probabilities to unseen events. The state-of-the-art smoothing is based on the Kneser-Ney approach (Kneser and Ney, 1995). Witten-Bell smoothing (Witten and Bell, 1991) is a competitive approach and generalizes to fractional counts (Iyer and Ostendorf, 1999).

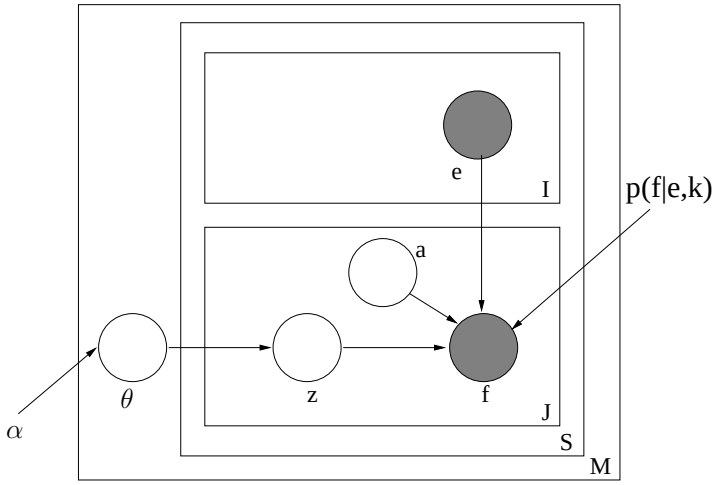

Figure 2.6: Graphical representation of bilingual topic admixture model. S and M denote the number of parallel sentences and the number of documents in parallel corpora respectively. I and J denote the number of words in a target sentence and a source sentence respectively. a is the word alignment variable for f.

2.4.1 Kneser-Ney Smoothing

The state-of-the-art smoothing for a backoff N-gram language model is based on Kneser-Ney smoothing (Kneser and Ney, 1995). The belief of its success comes from the preservation of a marginal distribution. The interpolated form of a bigram language model using the absolute discounting can be expressed as follows:

$$p_{KN}(v|u) \quad = \quad \frac{max\{C(u,v) - D, 0\}}{C(u)} + \lambda(u) \cdot p_{KN}(v) \tag{2.57}$$

where D is a discounting factor between 0 and 1. The lower-order distribution $p_{KN}(v)$ is treated as unknown parameters, which can be estimated using the preservation of a marginal distribution:

$$\hat{p}(v) \quad = \quad \sum_{u} p_{KN}(v|u) \cdot \hat{p}(u) \tag{2.58}$$

where $\hat{p}(v)$ is the marginal distribution estimated from background training data such that $\hat{p}(v) = \frac{C(v)}{\sum_{v'} C(v')}$. After substitution and manipulation of equations, we arrive at the fol-

lowing solution for the lower-order distribution:

$$p_{KN}(v) \quad = \quad \frac{N(\cdot, v)}{\sum_v N(\cdot, v)} \tag{2.59}$$

where $N(\cdot, v)$ denotes the number of unique words preceding v.

2.4.2 Witten-Bell Smoothing

Witten-Bell smoothing (Witten and Bell, 1991) is motivated from the Good-Turing estimate which states that for any event that occurs r times, we should pretend that it occurs r^* times according to the following relationship:

$$r^* \cdot n_r \quad = \quad (r+1) \cdot n_{r+1} \tag{2.60}$$

where n_r represents the number of events that occur r times. With this notion, the total mass of unseen bigrams (u, v) is equal to n_1 corresponding to the number of bigrams which occur once. Therefore, the total predicted mass of seen and unseen bigrams is estimated as $N_1(u, \cdot) + C(u, \cdot)$ where $N_1(u, \cdot)$ denotes the number of word types following u and occur once, and $C(u, \cdot)$ denotes the unigram count of word u. For a backoff language model, the probability mass reserved for the unigram distribution is expected to be $\frac{N_1(u,\cdot)}{N_1(u,\cdot)+C(u,\cdot)}$. Since there may be a chance that $N_1(u, \cdot)$ is equal to zero, $N_{1+}(u, \cdot)$, which denotes the number of word types following word u, is used instead to avoid zero probabilities. In summary, a bigram language model with Witten-Bell smoothing can be expressed as follows:

$$p_{WB}(v|u) \quad = \quad \frac{C(u, \cdot)}{N_{1+}(u, \cdot) + C(u, \cdot)} p_{ML}(v|u) + \frac{N_{1+}(u, \cdot)}{N_{1+}(u, \cdot) + C(u, \cdot)} p_{WB}(v) \tag{2.61}$$

where

$$p_{ML}(v|u) \quad = \quad \frac{C(u, v)}{C(u, \cdot)} \tag{2.62}$$

For comparison, Witten-Bell smoothing employs a maximum likelihood estimate for the bigram distribution while Kneser-Ney smoothing employs a *discounted* maximum likelihood estimate. Moreover, Kneser-Ney smoothing is grounded on the marginal preservation principle while Witten-Bell smoothing is motivated by the Good-Turing scheme. In

Chapter 4, we will generalize the Kneser-Ney smoothing with fractional counts for N-gram latent semantic analysis.

2.5 Unsupervised Language Model Adaptation

Language model adaptation is essential because of the mismatch between the training and the test domains. Besides LSA-based approaches, there are different approaches such as cache-based language model (Kuhn and Mori, 1990; Clarkson and Robinson, 1997), marginal adaptation (Kneser et al., 1997) and word triggering via maximum entropy model (Rosenfeld, 1994), to name just a few.

2.5.1 Word Caching

Word caching (Kuhn and Mori, 1990; Clarkson and Robinson, 1997) is an interesting approach for language model adaptation. The hypothesis is that the recently used words have a higher chance to occur again in the future. A cache-based language model can be implemented by tracking the frequency of recently occurred words in an exponentially decaying N-gram cache, serving as a self-triggering model. This approach is effective for supervised adaptation such as a dictation task where a user can help correct speech recognition errors. However, this approach can be harmful for unsupervised adaptation since it boosts the probability of mis-recognized words by increasing their word frequencies.

In Chapter 4, we show that language model adaptation via latent semantic analysis can be interpreted as a cache-based language model. Instead of caching the frequency of words, the LSA approach caches the frequency of the recent topics and thus is more robust against speech recognition errors for unsupervised adaptation.

2.5.2 Word Triggering

Word triggering has been proposed to capture a long-distance relationship between words (Rosenfeld, 1994; Chen et al., 1998; Wu and Khudanpur, 2002) complementary to an N-gram language model. In this approach, all possible word pairs within a context window are considered and filtered to produce a list of potential word triggers. Then the average mutual information for each word trigger is computed and those with high average mutual information are selected as final word triggers. Finally, the selected word triggers serve as constraints for maximum entropy language modeling or they are linearly interpolated with a background language model.

Word triggering can be extended to crosslingual settings (Kim and Khudanpur, 2004) where the word pairs between a source and a target language are extracted from document-aligned parallel corpora. Similarly, average mutual information $I(f, e)$ are employed to select the crosslingual word triggers. The word-trigger probability of a source word f given a target word e is computed as follows:

$$p(f|e) = \frac{I(f,e)}{\sum_{f'} I(f',e)} \tag{2.63}$$

where $I(f, e)$ is set to zero whenever (f, e) is not a trigger pair. Around 1% relative reduction in character error rate was reported for crosslingual language model adaptation on a Mandarin ASR system for broadcast news with the baseline performance of 28.8% in character error rates (Kim, 2004). For unsupervised language model adaptation, word triggering may be sensitive to speech recognition errors similar to word caching because incorrectly recognized words may also trigger other irrelevant words.

2.5.3 Marginal Adaptation

While linear interpolation is a convenient technique for language model adaptation, interpolating an N-gram language model with a unigram language model may destroy the "syntactic" structure of the N-gram language model especially for function words that are usually position sensitive, e.g. "I am", "You are" etc. Motivated by information theory, marginal adaptation (Kneser et al., 1997) finds an adapted language model $p_a(w|h)$

such that the Kullback-Leibler divergence between $p_a(w|h)$ and the background language model $p_{bg}(w|h)$ is minimized subject to marginal constraints. The objective function to minimize is as in equation 2.64,

$$\begin{array}{lrcl} \text{Minimize} & \sum_h p_a(h) \cdot KL\left(p_a(.|h)||p_{bg}(.|h)\right) & & \\ \text{such that} & \forall w : \sum_h p_a(h) \cdot p_a(w|h) & = & \tilde{p}(w) \\ & \forall h : \sum_w p_a(w|h) & = & 1 \end{array} \tag{2.64}$$

where $\tilde{p}(w)$ is some unigram distribution relevant to the test domain. One can write the Lagrangian of the objective function, take the derivative with respect to $p_a(w|h)$ and set it to zero (equation 2.65–2.66).

$$\begin{array}{rcl} D(p_a(.|.)) & = & \sum_h p_a(h) \cdot \sum_w p_a(w|h) \cdot \log \frac{p_a(w|h)}{p_{bg}(w|h)} \\ & & - \sum_w \lambda_w \left(\sum_h p_a(h) \cdot p_a(w|h) - \tilde{p}(w)\right) \\ & & - \sum_h \mu_h \left(\sum_w p_a(w|h) - 1\right) \end{array} \tag{2.65}$$

$$\begin{array}{rcl} \frac{\partial D(.)}{\partial p_a(w|h)} & = & p_a(h) \cdot (1 + \log \frac{p_a(w|h)}{p_{bg}(w|h)}) - \lambda_w \cdot p_a(h) - \mu_h = 0 \\ \Rightarrow p_a(w|h) & \propto & p_{bg}(w|h) \cdot e^{\lambda_w} \\ & \propto & p_{bg}(w|h) \cdot e^{\sum_j \lambda_j \cdot f_j(h,w)} \end{array} \tag{2.66}$$

where

$$f_j(h,w) = \begin{cases} 1 & \text{if } w = j \\ 0 & \text{otherwise} \end{cases} \tag{2.67}$$

$f_j(h,w)$ is a unigram feature function independent of h. Since the solution of the adapted language model is in an exponential form, the optimization problem is similar to the maximum entropy settings (Rosenfeld, 1994). Therefore, λ_j can be solved using the generalized iterative scaling (GIS) (Darroch and Ratcliff, 1972). In the literature, only one GIS iteration can be applied since $p_a(h)$ is unknown for computing feature expectation using a given model. However, it can be approximated with a background model $p_{bg}(h)$ at the

first iteration as in equation 2.68–2.70,

$$\forall j: \quad \lambda_j^{(1)} = \lambda_j^{(0)} + \log \frac{\tilde{E}[f_j(h,w)]}{E[f_j(h,w)]} \tag{2.68}$$

$$= \lambda_j^{(0)} + \log \frac{\sum_{h,w} \tilde{p}(w,h) \cdot f_j(h,w)}{\sum_{h,w} p_a^{(0)}(w|h) p_a^{(0)}(h) \cdot f_j(h,w)} \tag{2.69}$$

$$= \log \frac{\tilde{p}(w)}{p_{bg}(w)} \tag{2.70}$$

where $p_a^{(0)}(w|h) = p_{bg}(w|h)$, $p_a^{(0)}(h) = p_{bg}(h)$ and $\lambda_j^{(0)} = 0$.
In summary, the form of the adapted model is a rescaled version of the background language model:

$$p_a(w|h) = \frac{\alpha(w) \cdot p_{bg}(w|h)}{Z(h)} \tag{2.71}$$

where $Z(h)$ is a normalization factor to guarantee that probabilities sum to unity. $\alpha(w)$ is a scaling factor that is commonly approximated as follows:

$$\alpha(w) \approx \left(\frac{\tilde{p}(w)}{p_{bg}(w)} \right)^{\epsilon} \tag{2.72}$$

where ϵ is a tuning factor between 0 and 1 to compensate for the approximation due to the limitation of one GIS iteration. In general, marginal adaptation is very expensive due to the computation of the normalization factor $Z(h)$. However, an efficient normalization is available for a backoff N-gram language model (Kneser et al., 1997). The idea is to further impose a constraint that the total probability of the observed transition (h,w) in background training corpora is conserved after language model adaptation:

$$\sum_{w:(h,w)\in T} p_a(w|h) = \sum_{w:(h,w)\in T} p_{bg}(w|h) = Mass(h) \tag{2.73}$$

where the summation is taken *only* on the observed history and the current word (h,w) in training corpora T. Given that the background language model has a standard backoff structure plus the above constraint, the adapted language model has the following recursive backoff formula:

$$p_a(w|h) = \begin{cases} \frac{\alpha(w) \cdot p_{bg}(w|h)}{z_0(h)} & \text{if (h,w)} \in T \\ bo(h) \cdot p_a(w|\hat{h}) & \text{otherwise} \end{cases} \tag{2.74}$$

where

$$\frac{1}{z_0(h)} = \frac{Mass(h)}{\sum_{w:(h,w)\in T} \alpha(w) \cdot p_{bg}(w|h)} \tag{2.75}$$

and

$$bo(h) = \frac{1 - Mass(h)}{1 - \sum_{w:(h,w)\in T} p_a(w|\hat{h})} \tag{2.76}$$

$bo(h)$ denotes the backoff weight of the 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 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 equation 2.71, 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 verbs, articles, conjunctions, sentence boundary markers and punctuations, we do not adapt their N-gram probabilities because predicting stopwords mostly relies on syntactic context but not topical context. We can easily model this effect by inserting a new branch in equation 2.74 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 equation 2.73 and equation 2.75, respectively.

In Chapter 4 we will show that using the LSA marginals for $\tilde{p}(w)$ are effective for unsupervised marginal adaptation. We will describe a computationally inexpensive version of marginal adaptation for incremental lattice rescoring.

2.6 Summary

We have covered background materials relevant to the thesis including variational Bayes, approaches for latent semantic analysis for high-order models, language model smoothing, and language model adaptation techniques. In the following chapters, we will present our unified unsupervised topic adaptation framework for monolingual and crosslingual adaptation. In Chapter 4, we will describe monolingual language model adaptation for

automatic speech recognition based on N-gram latent semantic analysis. In Chapter 5, we will extend our framework to crosslingual adaptation for statistical machine translation based on bilingual N-gram latent semantic analysis.

Chapter 3

Baseline Transcription and Translation Systems

We describe the baseline transcription and translation systems used for our topic adaptation experiments in this thesis. The systems are built mainly for large-scale evaluation including Mandarin transcription, Arabic transcription and Chinese-to-English statistical machine translation.

3.1 Background

Our research effort centers on the GALE (Global Autonomous Language Exploitation) program [1]. The goal is to recognize, translate and extract useful information on broadcast news and broadcast conversation audio in multiple languages. Our research effort in this thesis focuses on transcription and translation parts of the GALE program. In the following sections, we describe our baseline transcription systems for Mandarin Chinese and Arabic as the source languages, and the Chinese-to-English translation systems, which translate an input Chinese text into English.

[1] http://www.darpa.mil/ipto/programs/gale/gale.asp, last consulted: 23 June 2009

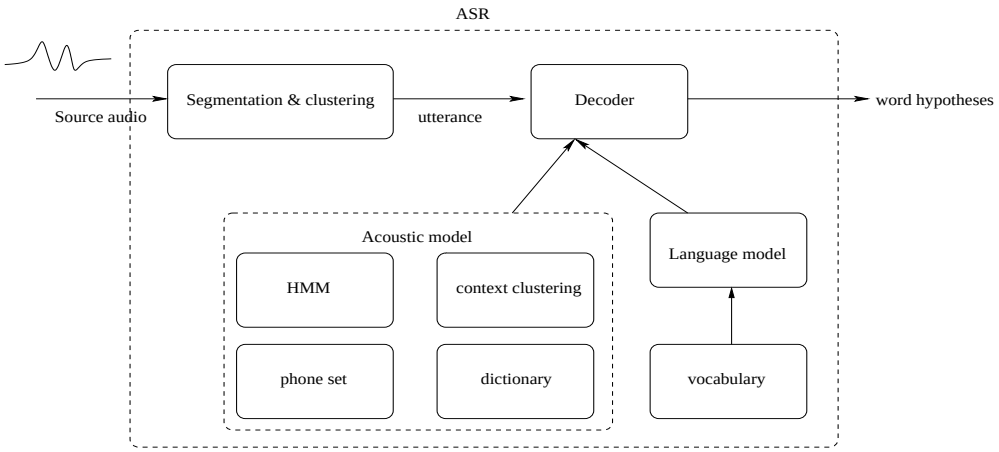

Figure 3.1: Block diagram of automatic speech recognition.

3.2 Mandarin Transcription

Our first Mandarin transcription system was implemented for the RT04 evaluation (Yu et al., 2004). Our recent GALE Mandarin transcription systems (Hsiao et al., 2008) have been implemented based on the RT04 system with significant performance improvement due to the increased amount of training data from the GALE program, improved speech segmentation and clustering (Section 3.2.2), discriminative training on acoustic models (Section 3.2.3), cross adaptation (Section 3.2.6) and system combination. All systems were based on hidden Markov models (HMM) (Rabiner, 1989) implemented using JRTK (Finke et al., 1997). Speech recognition was implemented using the IBIS decoder (Soltau et al., 2001).

The architecture of our transcription system is illustrated in Figure 3.1. An input audio is first segmented into speech utterances followed by automatic speaker clustering. Then a speech recognizer decodes an input speech utterance X into a word hypothesis W using an acoustic model $p(X|W)$ and a language model $p(W)$ using the Bayes decision rule:

$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} \underbrace{p(X|W)}_{\text{acoustic model}} \cdot \underbrace{p(W)}_{\text{language model}} \tag{3.1}$$

Table 3.1: Demi-syllables for Initial-Final modeling for Mandarin Chinese.

Initial	b c ch d f g h j k l m n p q r s sh t w x y z zh
Final	a ai an ang ao e ei en eng er i ia ian iang iao ie in ing iong iu o ong ou u ua uai uan uang ue ui un uo ü üe ng

3.2.1 Chinese-Specific Issues

Chinese text is not segmented at the word level. In other words, a sentence is simply a sequence of characters, with no spaces in between. It is not trivial to segment Chinese text into words. To make matters worse, since the distinction between words and phrases is weak, a sentence can have several acceptable segmentation with the same meaning. For language modeling purposes, it is important to have a good word list and to segment the training data properly. While the number of words can be unlimited, there are only about 6.7K characters in simplified Chinese. A Chinese character is pronounced as a syllable, hence Chinese is a mono-syllabic language. A syllable can have five different tones: flat, rising, dipping, falling, and neutral. There are about 1300 unique tonal syllables, or 408 unique syllables disregarding tones. Studies have shown that the realization of tones is context sensitive, an effect known as tone sandhi. For example, when a word comprises two third-tone characters, the first character will be realized in a second tone.

The out-of-vocabulary issue can be alleviated by adding a closed set of Chinese characters into the search vocabulary of a speech decoder so that if a new word is spoken, there is still a chance that the speech decoder may recognize the individual characters of the word.

Pinyin is the official romanization system for Mandarin Chinese. While most European languages are transcribed at the phone level, Pinyin is essentially a demi-syllable level representation, also known as initial-final: an initial is typically a consonant; a final can be either a monophthong, a diphthong or a triphthong. There are 23 initials and 35 finals in Mandarin as shown in Table 3.1. Since the Pinyin representation is standard, it is easy to find pronunciation lexicons in this format. Alternatively to using pinyin, one can use a phonetic representation for pronunciations. The LDC 1997 Mandarin CallHome lexicon

(LDC96L15) contains phonetic transcriptions for about 44K words, using a phone set of 38 phones. While phonemes are well studied and understood, they are not the most natural representation for Chinese. It also remains unclear whether there is a widely accepted phonetic transcription standard for Chinese.

In Mandarin Chinese, some words can have the same pinyin and tone transcription but with different meanings. For example, "Shu4-Mu4" can mean "tree" or "number" depending on a context. Therefore, it is very useful to use topical information from the context for word disambiguation.

3.2.2 Audio Segmentation and Speaker Clustering

Audio segmentation in our system (Yu et al., 2004) is implemented via an HMM segmenter with four classes: Speech, Noise, Silence, and Music. The speech features are 13-dimension MFCC plus their first and second derivatives. Each class is represented by a Gaussian mixture model with 64 Gaussians. The system is trained on 3 hours of manually annotated HUB4 shows.

The resulting speech segments (the Noise, Silence, and Music segments are ignored) are then grouped into several clusters, each of which ideally correspond to an individual speaker. A hierarchical, agglomerative clustering technique with Bayesian Information Criterion (BIC) is applied (Jin and Schultz, 2004). A tied Gaussian mixture model (TGMM) is built on the whole set of speech segments, serving as a background model. A Gaussian mixture model for each cluster is trained via adaptation of the background TGMM. Each segment is considered as a cluster at an initial step. We define the distance between two clusters by the Generalized Likelihood Ratio (GLR):

$$D(C_1, C_2) = -\log \frac{p(X|\theta)}{p(X_1|\theta_1)\ p(X_2|\theta_2)} \tag{3.2}$$

where X_1, X_2, and X are feature vectors in clusters C_1,C_2, and the merged cluster of C_1 and C_2, respectively. θ_1, θ_2, and θ are statistical models built on X_1, X_2, and X, respectively.

We can see from equation 3.2 that the smaller the distance, the closer the two clusters

Table 3.2: Acoustic model training data for Mandarin transcription.

Mandarin systems	Hour	Source
RT04/GALE-P1-dryrun	96hr	HUB4m (LDC98S73), TDT4 subset (LDC2005S11)
GALE-P1	558hr	+ GALE-Y1 data (LDC2007E99)
GALE-P2/GALE-P3	1300hr	+ GALE-Y2 data (LDC2008E38)

are to each other. At each step, the two closest clusters are merged and the model of a new cluster is re-estimated. Clustering terminates when the BIC stopping threshold is exceeded. Setting the threshold is an art that highly depends on the nature of an input audio. In the GALE-P2 and GALE-P3 evaluations, a snippet show, which is a piece of story on a consistent topic, is relatively shorter (1-2 minutes) compared to the GALE-P1 evaluation. In addition, an input audio file can contain multiple snippet shows. We therefore perform clustering across snippets on the same show, which means that all speech segments from different snippets of the same show are pooled together for clustering and a unique speaker label is shared across different snippets. We apply two different BIC thresholds depending on the number of snippets per show to reduce the chance of underestimating the number of speakers in a multiple-snippet show. Underestimating the number of speakers will affect the performance of acoustic model adaptation since speech segments from an irrelevant speaker may be used undesirably for adaptation.

3.2.3 Acoustic Modeling

Table 3.2 shows the audio data for acoustic model training for the RT04 and GALE transcription systems. We benefit from an increased amount of the in-domain training data from the GALE program [2] so that we can afford more codebook and Gaussian parameters.

For feature extraction, we use 13 Mel-Frequency Cepstral Coefficients (MFCC) per frame. Cepstral mean and variance normalization is performed on a speaker/cluster basis.

[2] http://projects.ldc.upenn.edu/gale/data/DataMatrix.html, last consulted 30 Mar 2009

Dynamic features are extracted by concatenating 15 adjacent frames (current frame ± 7 adjacent frames), then using linear discriminant analysis (Duda et al., 2001) to produce an output feature vector with 42 dimensions. Vocal tract length normalization (VTLN) is performed on a speaker/cluster basis (Zhan and Westphal, 1997). As described before, the acoustic modeling units can be either Initial-Finals (I-F) or phones. In both cases, context-dependent models are built and then clustered using a decision tree with a set of phonetically motivated questions. Each Initial and Final demi-syllables employ a 3-state and 5-state hidden Markov model while each phone has 3 states, all with a strict left-to-right topology. We find that both systems give comparable performance, with the initial-final system slightly better than the phone-based system. We take advantage of cross-adapting among the syllable-based and the phone-based systems to improve the recognition accuracy.

Model Training

Since the manual transcripts of the GALE training audio are manually segmented with speaker labels, speech segmentation and clustering are not involved in the GALE training audio. The model training can be performed by bootstrapping a context-independent model using a legacy model. Then the context-independent model is used to estimate a context-dependent model, which is further refined to a speaker-independent and a speaker-adaptive model for multi-pass decoding.

Given our legacy RT04 system, we perform initial forced alignment on the GALE training utterances to obtain the state alignment. Then we use maximum likelihood estimation to train the acoustic models. For context-independent models, we apply the following steps:

1. Perform linear discriminant analysis (Duda et al., 2001) to project a window of MFCC feature vectors into a 42-dimension feature vector.

2. Extract feature samples for each HMM state according to the state alignment of the training utterances.

3. Perform the K-means clustering to obtain an initial Gaussian mixture model with a fixed number of Gaussians.

4. Run eight iterations of label training (i.e. with a fixed state alignment) to refine the acoustic model. Viterbi or Baum-Welch training can also be applied at the expense of increased computation.

5. Refine the state alignment using the updated model and repeat the above steps as needed.

We initialize an unclustered context-dependent model with the context independent model via sharing the Gaussian parameters but with unique Gaussian mixture weights per unclustered model. Then we train a speaker independent (SI) model for the first-pass decoding:

1. Perform one iteration of Viterbi training to obtain the mixture weights of each unclustered context-dependent model.

2. Perform top-down state clustering based on the Gaussian mixture weights of the unclustered model using the decision tree approach that maximizes the information gain after each node is split (Rogina, 1997). A phonetically motivated question set is applied, producing tied quinphone states (or senone) with a fixed number of codebooks.

3. Perform linear discriminant analysis to project a window of MFCC feature vectors into a 42-dimension feature vector.

4. Extract feature samples for each HMM state according to the state alignment of the context-independent system.

5. Perform the merge and split training to grow the Gaussian mixtures incrementally depending on the minimum occupancy count of a Gaussian and the maximum number of Gaussians allowed (which is set to 100 per codebook for the GALE Mandarin systems).

6. Perform global semi-tied covariance (STC) training (Gales, 1999).

Below are the procedures for speaker-adaptive training:

1. Determine the warping factors for vocal tract length normalization (Zhan and Westphal, 1997) for each training speaker.

2. Perform linear discriminant analysis using the warped MFCC features as inputs.

3. Extract feature samples for each HMM state.

4. Perform the merge and split training for each HMM state followed by semi-tied covariance training.

5. Perform speaker-adaptive training (SAT) using a single feature space transform per speaker (Gales, 1997), known as the feature space adaptation (FSA) or the feature-space maximum likelihood linear regression (fMLLR).

With the syllable-to-phone mapping, we can bootstrap an initial phone model easily. The phone-based systems follow the same training procedures using a phone-based word lexicon. In addition, we attempt to maximize the difference between the phone-based and the syllable-based system via the genre-dependent modeling. We incorporate a binary question to ask whether a phonetic context is from broadcast news (BN) or broadcast conversation (BC). Then the state clustering procedure will determine if this question is applied for node splitting.

Discriminative Training

Since the hidden Markov model is not a correct model for speech, discriminative training is an essential technique to correct the modeling assumption and thus giving significant improvement in recognition accuracy. Maximum mutual information estimation (MMIE) (Valtchev et al., 1997) and boosted MMIE (BMMIE) (Povey et al., 2008) are common techniques for discriminative training and have been applied in our GALE-P3 Mandarin transcription system (Hsiao et al., 2008). Recent advancement in discriminative

training includes better smoothing via controlling the degree of improvement of conditional likelihood on the model space (Hsiao et al., 2009) and the feature space (Hsiao and Schultz, 2009).

Starting with the syllable-based speaker-adaptive model using maximum likelihood estimation, we decode the GALE training utterances to generate the word lattices to represent a compact set of competing hypotheses for each utterance.

MMIE aims at maximizing the posterior probability of a reference compared to the competing hypotheses in a word lattice. The objective function of MMIE is:

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(X_r|M_{s_r}) \cdot p(s_r)}{\sum_{s \in lat} p_\lambda(X_r|M_s) \cdot p(s)} \tag{3.3}$$

where λ represents the acoustic model parameters to be optimized; X_r is the r-th training utterance; s_r is the reference and M_s represents the corresponding HMM state sequence of sentence s. Maximizing $F_{MMI}(\lambda)$ improves the posterior probability of the reference in a lattice. This function can be optimized using the extended Baum-Welch algorithm. The update equations of Gaussian means and covariances, without the smoothing parts, are:

$$\hat{\mu}_r = \frac{x_r^{num} - x_r^{den} + D_r \mu_r}{\gamma_r^{num} - \gamma_r^{den} + D_r} \tag{3.4}$$

$$\hat{\Sigma}_r = \frac{S_r^{num} - S_r^{den} + D_r(\Sigma_r + \mu_r \mu_r')}{\gamma_r^{num} - \gamma_r^{den} + D_r} - \hat{\mu}_r \hat{\mu}_r' \tag{3.5}$$

where x_r and S_r are the weighted sum of features x_t and $x_t x_t'$ for the r-th Gaussian, respectively; γ_r represents the occupancy count; D_r is a constant to control the learning rate and to ensure Σ_r is positive definite. The superscripts **num** and **den** specify the statistics belonging to the numerator or denominator of $F_{MMI}(\lambda)$. For MMIE, the numerator statistics are the same as that of maximum likelihood estimation, while denominator statistics are collected from the word lattices.

Intuitively, some paths may contain more error than other paths in a word lattice. Boosted MMIE boosts the importance of competitors that make large error. The objective function is shown as follows:

$$F_{BMMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(X_r|M_{s_r}) \cdot p(s_r)}{\sum_s p_\lambda(X_r|M_s) \cdot p(s) \cdot e^{-bA(s,s_r)}} \tag{3.6}$$

Table 3.3: Acoustic model configurations of the GALE Mandarin transcription systems. * denotes that (boosted) MMIE and genre-dependent modeling were applied to the GALE-P2 and P3 systems only.

Model	1st-pass	2nd-pass	3rd-pass
Unit	I-F	phone	I-F
Model	SI	SAT-FSA	SAT-FSA
Training	ML	ML	(B)MMIE*
# Codebook	10K	10K	10K
# Gaussian	805K	825K	784K
Genre-dependent	no	yes*	no
Algorithms	-	STC&VTLN	STC&VTLN

where $A(s, s_r)$ is the raw phone accuracy of sentence s (Povey, 2003); b is the boosting factor larger than zero. Hence, the likelihood of the competitors with higher error are boosted. The update equation of BMMIE is the same as MMIE, but the denominator statistics are altered when we compute the forward-backward scores on the lattices. In other words, the likelihood score of each word arc in a lattice is subtracted by $b \cdot A(s, s_r)$. In our system, the boosting factor is set to 0.5. We apply I-smoothing with $\tau = 100$ for both MMIE and BMMIE training. The maximum likelihood model is used as a backoff in I-smoothing. Finally, the discriminatively trained model is used for the third-pass decoding in our GALE system.

Table 3.3 summarizes the acoustic model configurations for the GALE Mandarin transcription systems.

3.2.4 Language Modeling, Text Data, Normalization

We use several corpora for our language model development for the RT04 and the GALE systems: Mandarin Chinese News Text, TDT$\{2,3,4\}$, Mandarin Gigaword corpora v1-2, the HUB4 1997 acoustic training transcript, the CALLHOME Mandarin transcript, the GALE acoustic training and web transcripts, and the web data shared among the com-

Table 3.4: Language model training data for Mandarin transcription.

Mandarin systems	Word token	Source
RT04-small	13M	Xinhua News 2002 (LDC2003T09)
RT04-full/GALE-P1-dryrun	300M	Mandarin Chinese News Text (LDC95T13), TDT2 (LDC2001T57), TDT3 (LDC2001T58), TDT4 (LDC2005T16), Mandarin Gigaword corpora v1 (LDC2003T09), the HUB4 1997 acoustic training transcript (LDC98S73), the CALLHOME Mandarin transcript (LDC96T16)
GALE-P1	800M	+ GALE-Y1 transcripts (LDC2007E100) + Mandarin Gigaword corpora v2 (LDC2005T14)
GALE-P2	1.0B	+ GALE-Y2 transcripts (LDC2008E39) and web data
GALE-P3	1.5B	+ GALE-Y3 transcripts (LDC2008E39) and shared web data

peting teams of the GALE program. We divide the training data into subsets based on their sources, including the acoustic training transcripts for broadcast news and broadcast conversation as separate sources. These give us 14 sources including BBC, CCTV, NTDTV, RFA, Central News Agency, China Radio, People's Daily, Sina News, Xinhua News, Zaobao News, Phoenix TV, broadcast news, broadcast conversation and others. Table 3.4 summarizes the text data for language modeling in different Mandarin systems. The RT04-small and RT04-full systems are employed for small-scale topic-adaptation experiments while the GALE systems are used for large-scale evaluation. Any text that falls into the excluded time frame (specified in the evaluation specification) are removed from the training corpora.

Before training a statistical language model, we first process the Chinese text data to normalize for ASCII numbers, ASCII strings and punctuations. We devise heuristic rules in combination with a Maximum Entropy classifier to normalize numbers. The classifier classifies whether an input number is a digit string (e.g. telephone number) or a number quantity based on the surrounding word context. We map English words to a special token +english+, human noises (such as breath and cough) to +human_noise+. Non-human (environmental) noises are removed from the training transcript. Since punctuations provide word boundary information useful for word segmentation, they are removed only after word segmentation. Word segmentation is based on a maximal substring matching approach that locates the longest possible word segment at each character position. Since proper names are often incorrectly segmented, we later on add the LDC Named-Entity (NE) list (LDC2005T34) into a wordlist provided from the official LDC segmenter [3]. The NE list contains different semantic categories, such as organization, company, person and location names. Having them in the wordlist greatly improves segmentation quality, which translates to more accurate predictions in an N-gram language model. After word segmentation, we choose the vocabulary to be the top-K most frequent words. The commonly used Chinese characters ($6.7k$) are then added into the vocabulary, giving the $63k$ vocabulary for the RT04 system. For cross-site collaboration between IBM, JHU and CMU-InterACT for the GALE evaluation, we merge our word lists plus some frequently occurring English words, giving a fixed $108k$ vocabulary for the GALE systems.

Unless specified, a background 4-gram language model is trained with the modified Kneser-Ney smoothing using the SRI language model toolkit (Stolcke, 2002) in this thesis. Each of the source-dependent language models are linearly interpolated with the interpolation weights estimated using a heldout set.

3.2.5 Pronunciation Lexicon

Our pronunciation lexicon is based on the LDC CallHome Mandarin lexicon, which contains about $44k$ words. Pronunciations for words not covered by the LDC lexicon are

[3] http://projects.ldc.upenn.edu/Chinese/segmenter/Mandarin.fre, last consulted 27 March 2009

generated using a maximal matching method. The idea is similar to our word segmentation algorithm. We first compile a list of all possible character segments for each covered vocabulary word. For each uncovered word, the algorithm repeatedly searches for the longest matching character segment from the beginning to the end of the word, producing a sequence of character segments. Pronunciations of these segments are then concatenated to produce the pronunciation for a new word. There are 23 initials and 35 finals, and 38 phonemes defined by the CallHome lexicon. Eight additional phonemes are used to model human noises, environmental noises and silence. We use the demi-syllable-to-phoneme mappings provided by the CallHome lexicon to convert a demi-syllable lexicon into a phone-based lexicon.

3.2.6 Decoding Strategy

We employ a three-pass decoding strategy for the GALE evaluation. Given the manual segmentation of the test audio, automatic speaker clustering is performed. Then we use the speaker-independent model to decode the test utterances to obtain the initial word hypotheses for acoustic model adaptation using vocal tract length normalization (Zhan and Westphal, 1997), feature-space adaptation (Gales, 1997) and model-space maximum likelihood linear regression (Leggetter and Woodland, 1995) with a maximum of 256 classes in a binary regression tree. The word confidence is also applied to weight the importance of each frame during adaptation. Then we use the adapted phone-based system for the second-pass decoding. Similarly, the word hypotheses from the second-pass decoding are used to adapt the speaker-adaptive syllable-based system for the third-pass decoding.

Cross-Adaptation with IBM

For the GALE evaluation, we perform cross-adaptation between our system and the IBM system via exchanging the word hypotheses from the final self-adapted systems. In other words, each system first goes through a multi-pass decoding procedure on test utterances. Then the word hypotheses from the IBM system are treated as transcription references for acoustic model adaptation on our system. Similarly, the IBM system is adapted using the

Table 3.5: Size of the acoustic model (AM) and language model (LM) training corpora and the size of vocabulary for the Arabic transcription system in terms of number of hours and word tokens.

Arabic	AM	LM	Vocab.
GALE-P3	1500hr	962M	737k (OOV ≈1.0%)

word hypotheses from our system. To facilitate the cross-adaptation procedure, we use a common search vocabulary and a speaker clustering database for decoding.

3.3 Arabic Transcription

Besides Mandarin, we also evaluate our language model adaptation approach on Arabic transcription. Our Arabic transcription system adopts a similar training procedure as our Mandarin transcription system described in Section 3.2.3. More details can be found in (Noamany et al., 2007). The training corpora for acoustic and language model training for the GALE-P3 evaluation are shown in Table 3.5.

3.3.1 Arabic-Specific Issues

Two Arabic-specific issues are worth mentioning. First, the written Modern Standard Arabic (MSA) lacks short vowels in between two consonants. Vowelization is required to restore the vowels for accurate acoustic modeling training. Given all possible configuration of vowelization of a transcribed utterance, forced alignment can be applied to select the most likely configuration. However, it is challenging to vowelize the text data with high accuracy for language modeling. Therefore, it is generally preferred to allow the speech decoder to choose the best vowelized model during decoding. Therefore, our Arabic language model is trained on unvowelized text.

Secondly, Arabic has rich morphology. A new word can be formed by attaching prefixes and suffices to a word stem. As a result, the vocabulary size in Arabic (over 700k)

is much larger than that in Mandarin (108k). This makes the data sparseness issue more critical for language modeling. Large vocabulary is required to maintain an acceptable out-of-vocabulary rate within 1% on a development set. In our preliminary experiment, we reduce the vocabulary size via stemming to alleviate the data sparseness issue. However, this is not effective to improve the lattice rescoring performance since most of the word transitions in a lattice become identical after stemming and thus losing the discriminative power. Therefore, we do not apply word stemming in our experiments.

3.3.2 Decoding Strategy

Similar to our Mandarin transcription system, the Arabic transcription system employs a three-pass decoding strategy. The first-pass decoding employs an unvowelized speaker-independent model in which a vowel between consonants is not restored. The second-pass decoding employs an unvowelized speaker-adaptive model that uses the word hypotheses from the first-pass decoding for acoustic model adaptation using vocal tract length normalization, feature-space adaptation and model-space maximum likelihood linear regression. In the third-pass decoding, we employ a vowelized model in which the vowel is restored in the pronunciation lexicon. Interleaving the use of unvowelized and vowelized models may maximize the system difference in terms of different error patterns so that the adaptation performance may be enhanced.

3.3.3 Performance Metrics

Performance metrics are the word perplexity and the character (word) error rate defined as follows:

$$\text{Perplexity} = e^{-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i|w_{i-3}w_{i-2}w_{i-1})} \tag{3.7}$$

$$\text{Word error rate} = \frac{I+D+S}{N} \tag{3.8}$$

where N denotes the number of word tokens in a reference. A word history is usually represented as a trigram history $w_{i-3}w_{i-2}w_{i-1}$. I, D, and S denote the insertion, deletion and

Table 3.6: Statistics of the Mandarin RT04 test set.

RT04	Duration	Genre
CCTV	0.33hr	BN
NTDTV	0.33hr	BN
RFA	0.33hr	BN plus phone interview

Table 3.7: Statistics of the development and test sets for the GALE evaluations from phase 2 (P2) to phase 3 (P3). "Eval07u" and "Eval07r" stand for unsequestered and re-test portions of Eval07 respectively.

	Mandarin Chinese					Arabic		
	P2		P3			P3		
Genre	Eval06	Dev07	Dev08	Eval07u	Eval07r	Dev07	Dev08	Eval07u
BN	0.59hr	1.07hr	0.49hr	0.63hr	-	1.71hr	-	2.05hr
BC	0.45hr	1.38hr	0.48hr	0.52hr	-	0.87hr	-	2.03hr
ALL	1.04hr	2.53hr	0.98hr	1.15hr	1.64hr	2.58hr	3.04hr	4.08hr

substitution error after aligning the recognized word sequence with the reference using dynamic programming. For fair comparisons between different approaches, an optimal word error rate is reported after tuning an optimal word insertion penalty (lp) and a language model weight (lz) that are usually combined in the following manner:

$$\text{Score(X,W)} \quad = \quad \log p(X|W) + lz \cdot \log p(W) - lp \tag{3.9}$$

where X and W denote a speech utterance and a word sequence respectively.

3.3.4 Evaluation Sets

We have the RT04 eval set and the GALE development/test sets to benchmark the recognition performance as shown in Table 3.6 and Table 3.7 respectively. The Mandarin RT04 eval set are mainly broadcast news while the GALE development/test sets are a mixture of

Table 3.8: Sources of the GALE Mandarin development/test sets.

Set	Source
Eval06	CCTV4, NTDTV, PHOENIX
Dev07	CCTV1, CCTV4, CCTVNEWS, NTDTV, PHOENIX
Dev08	BEIJING, CCTV1, CCTV2, CCTV4, CCTV7, CCTVNEWS, NTDTV, PHOENIX, VOA
Eval07u	ANHUI, CCTV1, CCTV2, CCTV4, CCTV7, CCTVNEWS, NTDTV, PHOENIX
Eval07r	CCTV1, CCTV4, CCTV7, CCTVNEWS, NTDTV, PHOENIX

Table 3.9: Sources of the GALE Arabic development/test sets.

Set	Source
Dev07	ABUDHABI ALAM ALJZ ARABIYA DUBAISCO IRAQIYAH KUWAITTV LBC SCOLA SYRIANTV
Dev08	ALAM ALHIWAR ALHURRA ALJZ ALMANAR ALURDUNYA ARABIYA DUBAISCO IRAQIYAH KUWAITTV LBC OMANTV SAUDITV SCOLA SYRIANTV
Eval07u	ABUDHABI ALAM ALJZ ARABIYA DUBAISCO IRAQIYAH KUWAITTV LBC OMANTV SCOLA SYRIANTV

broadcast news and broadcast conversation. Since broadcast conversation is more spontaneous in speaking style, speech recognition becomes more challenging on broadcast conversation than broadcast news. Table 3.8 shows the sources of the GALE development and test sets. Originally, Eval07 is designed for the GALE-P2 evaluation. Because of system retesting, known as the GALE-P2.5 evaluation, Eval07 is divided into the unsequestered portion (Eval07u) and the retest portion (Eval07r). Eval07u is treated as an internal development/test set while the retest portion is part of the official test set for the GALE-P2.5 evaluation.

Aligning word hypotheses with the reference transcription using hubscr07.pl [4] produces a SGML (Standard Generalized Markup Language) file for a baseline system and an alternative system. Then the Matched Pairs Sentence-Segment Word Error (MAPSSWE) approach (Gillick and Cox, 1989) is performed for significance testing using the NIST scoring tool (sclite) (Pallett et al., 1990) with the following command:

cat baseline.sgml alternative.sgml | sc_stats -p -t mapsswe -v -u -n result.txt

3.4 Statistical Machine Translation

Our baseline statistical machine translation system is trained using parallel training sentences. We first introduce different components of statistical machine translation and describe our baseline systems.

3.4.1 Basic Components

Statistical machine translation (SMT) usually consists of three components: a translation model $p(F|E)$, a language model $p(E)$ and a distortion model $d(F, E)$ where F is an input sentence of a source language and E is an output sentence of a target language. From the Bayes point of view, the decision rule for statistical machine translation is identical to that

[4] http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/tools/hubscr07.pl, last consulted 9 April 2009

for automatic speech recognition:

$$\hat{E} = \arg\max_{E} p(E|F) = \arg\max_{E} \underbrace{p(F|E)}_{\text{translation model}} \cdot \underbrace{p(E)}_{\text{target language model}} \quad (3.10)$$

The translation model is analogous to an acoustic model, while a language model is required on both tasks. However, statistical machine translation requires a distortion model to help re-order output words/phrases on the target language. In this thesis, topic adaptation applies to the translation model and the language model of the target language.

3.4.2 Word Alignment

Parallel sentences are required to obtain a word translation lexicon $p(f|e)$ in statistical machine translation. Therefore, we need to know how a source word f_j at position j of a source sentence $F = f_1^J$ aligns to a target word e_i at position i of a target sentence $E = e_0^I$. Prevalent word alignment models include the IBM Model 1–5 (Brown et al., 1994), the HMM model (Vogel et al., 1996) and Model 6 (Och and Ney, 2003). A latent alignment variable a_j of a source word f_j is a position index of a target word e_{a_j} in a target sentence E. Below is a generative procedure for a parallel sentence pair using the IBM Model 1:

For each position j of a source sentence f_1^J,

- Sample an alignment variable $a_j \in [0, I]$ uniformly to pick a word e_{a_j} in a target sentence e_0^I where e_0 denotes a NULL word, meaning that f_j does not align to any target word.
- Sample f_j from a word translation lexicon $p(f|e_{a_j})$.

The generative procedure defines the joint likelihood of the source sentence f_1^J and the word alignment sequence a_1^J given the target sentence e_0^I:

$$p(f_1^J, a_1^J|e_0^I) \propto \prod_{j=1}^{J} p(a_j|I) \cdot p(f_j|e_{a_j}) \quad (3.11)$$

$$\propto \left(\frac{1}{I+1}\right)^J \prod_{j=1}^{J} p(f_j|e_{a_j}) \quad (3.12)$$

The IBM Model 1 samples each alignment variable a_j independently and uniformly. IBM Model 2 replaces the uniform distribution with a distribution $p(a_j|I)$. The HMM model introduces the first-order dependence of the alignment variables via $p(a_j|a_{j-1}, I)$. IBM Model 3-5 further improves the word alignment via the notion of fertility and the inverted alignment set B_i. With fertility, a target word e_i can generate multiple source words, with B_i containing a set of positions of the source words. Model 3 makes a zero-order distortion model over B_i using $p(B_i|e_i)$ while Model 4-5 employ a first-order model $p(B_i|B_{i-1}, e_i)$ which depends on the previous inverted alignment set B_{i-1}. The generative procedure of the IBM Model 4 for $p(F|E)$ is described as follows:

- For each position i of a target sentence e_1^I, sample a fertility factor ϕ_i for e_i of how many source words f are generated (which can be zero).
- Sample a fertility factor ϕ_0 for the NULL word e_0 from a binomial distribution Binomial$(\phi|\sum_{i=1}^{I}\phi_i)$.
- For each position i of e_0^I, sample the source words given e_i up to the fertility count using the word translation lexicon, that is $f_{ik} \sim p(f|e_i)\ \forall k = 1...\phi_i$.
- For each position i of e_1^I, sample the inverted word alignment set B_i. To start with, the position of the first source word f_{i1} is determined by sampling the jump distance Δ_1 using the probability distribution $p_{=1}(\Delta|\text{class}(f_{i1}), \text{class}(e_i))$ relative to the center position of the previous non-empty $B_{c(i)}$ denoted as $\overline{B_{c(i)}}$. In other words, $B_{i1} = \overline{B_{c(i)}} + \Delta_1$. The word classes of the source and target words can be determined using a word clustering algorithm (Brown et al., 1992). The use of word classes is to reduce the number of model parameters and improve the model generalization. The jump distances Δ_k for other remaining source words f_{ik} (with $k > 1$) are sampled using another probability distribution $p_{>1}(\Delta|\text{class}(f_{ik}))$ monolingually. In other words, $B_{ik} = B_{ik-1} + \Delta_k$.
- The inverted alignment positions B_{0k} corresponding to the NULL word e_0 is sampled with a uniform distribution $\frac{1}{\phi_0!}$, assuming that each permutation is equally likely.

The generative procedure of IBM Model 4 corresponds to the joint likelihood of the source sentence f_1^J and the word alignment sequence a_1^J given the target sentence e_0^I:

$$
\begin{aligned}
p(f_1^J, a_1^J|e_0^I) &= p(f_1^J, B_0^I|e_0^I) \qquad (3.13)\\
&= \underbrace{p(\phi_0|\sum_{i=1}^{I}\phi_i)\prod_{i=1}^{I}p(\phi_i|e_i)}_{\text{fertility model}} \cdot \underbrace{\prod_{i=0}^{I}\prod_{k=1}^{\phi_i}p(f_{ik}|e_i)}_{\text{translation model}} \cdot \\
&\quad \underbrace{\prod_{i=1}^{I}p_{=1}(B_{i1}-\overline{B_{c(i)}}|\text{class}(f_{i1}),\text{class}(e_i))}_{\text{distortion model for the first source position}} \cdot \\
&\quad \underbrace{\prod_{i=1}^{I}\prod_{k=2}^{\phi_i}p_{>1}(B_{ik}-B_{ik-1}|\text{class}(f_{ik}))}_{\text{distortion model for the remaining source positions}} \cdot \\
&\quad \underbrace{\frac{1}{\phi_0!}}_{\text{distortion model for } B_0} \qquad (3.14)
\end{aligned}
$$

Model 3-4 is said to be deficient because both models ignore whether a source word has been chosen while sampling B_i. In addition, probability mass is reserved for source positions outside a sentence boundary. Therefore, the probability distribution does not sum to unity. Model 5 addresses this deficiency by excluding the source positions which have been already sampled. Model 6 is a log-linear combination of Model 4 and the HMM model.

Given the word alignment of parallel sentences, estimating a word translation lexicon can be performed via the word alignment counts $C(f, e)$ of how many times a source word f is aligned to a target word e in training corpora:

$$p(f|e) = \frac{C(f,e)}{\sum_{f'}C(f',e)} \qquad (3.15)$$

A backward translation lexicon $p(e|f)$ can be estimated in a similar fashion. Expectation-Maximization or Viterbi training is involved to refine the word alignment and then re-estimate the translation lexicons iteratively until convergence.

3.4.3 Phrase Extraction

Word alignment is an essential step towards state-of-the-art phrase-based statistical machine translation (Koehn et al., 2003; Vogel et al., 2003). The advantage of using phrase translations is that local word re-ordering can be captured within a phrase pair $< \tilde{f}, \tilde{e} >$, which can be extracted from the word alignment between a parallel sentence pair. The aligned phrase pairs that are consistent with the word alignment are collected: The words in a legal phrase pair are only aligned to each other, and not to words outside (Och et al., 1999).

Similar to estimating a word translation lexicon, we can use a phrase alignment count $C(\tilde{f}, \tilde{e})$ to estimate the phrase translation probability known as the phrase score:

$$p(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} C(\tilde{f}', \tilde{e})} \tag{3.16}$$

Alternatively, the translation probability can be computed using the best word alignment of a phrase pair, known as the lexical weighting, as follows:

$$p_w(\tilde{f}|\tilde{e}) \approx \max_a p(\tilde{f}|\tilde{e}, a) \tag{3.17}$$

$$= \max_a \prod_j \frac{1}{|\{i : (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(f_j|e_i) \tag{3.18}$$

where a is some word alignment configuration of a phrase pair observed in all parallel sentence pairs. Typically, phrase extraction creates a phrase table with four scores for each phrase pair: the phrase scores and the lexical weightings in both translation directions.

3.4.4 Minimum Error Rate Training

Motivated from the maximum entropy modeling, a direct modeling approach (Och and Ney, 2002) is prevalent to model the posterior probability $p(e_1^I|f_1^J)$ via a set of feature functions $h_m(e_1^I, f_1^J)$:

$$p(e_1^I|f_1^J) = \frac{e^{\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)}}{Z(f_1^J)} \tag{3.19}$$

where each feature function $h_m(\cdot)$ takes a sentence pair as an input and produces a real number as an output. For instance, feature functions include language model, distortion model, word count, phrase count, phrase translation scores and lexical weightings. For instance, the feature function for a language model takes the target sentence (and ignore the source sentence) and returns the total language model score of the target sentence e_1^I. Minimum error rate training (Och, 2003) is applied to optimize the feature weights $\{\lambda_m\}$ with an optimization criterion such as BLEU, which will be discussed in Section 3.4.6. Given the N-best translation candidates of each of the input sentences in a development set, the feature weights are adjusted iteratively to re-rank the N-best lists such that the cost function is optimized.

3.4.5 Language Modeling

Similar to automatic speech recognition, an N-gram language model is usually employed for statistical machine translation. The language model is trained only on text from the target language. Since our baseline system translates from Chinese to English, our target language model is trained on English text from monolingual non-parallel corpora and the English side of bilingual parallel training corpora. The SRILM toolkit (Stolcke, 2002) is used for language model training using the modified Kneser-Ney smoothing. Similar to language modeling for automatic speech recognition, the English text are partitioned according to sources so that a source-dependent language model is trained. The language models are linearly interpolated with the interpolation weights estimated using a heldout set.

3.4.6 Performance Metrics

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) measures the quality of translation based on the statistical closeness of translated sentences $\{s_i\}$ to reference trans-

lations $\{r_i\}$ with I sentence pairs in a test set. The BLEU metric is defined as follows:

$$\text{BLEU} = \text{BP} \cdot e^{\sum_{n=1}^{N} \frac{1}{N} \cdot \log p_n} \quad (3.20)$$

$$\text{where } p_n = \frac{\sum_{i=1}^{I} \sum_{\text{n-gram} \in s_i} \text{count(n-gram)}}{\sum_{i=1}^{I} \sum_{\text{n-gram} \in s_i} \text{count}_{sys}\text{(n-gram)}} \quad (3.21)$$

$$\text{and BP} = e^{-\max(0, \frac{L_{ref}}{L_{sys}} - 1)} \quad (3.22)$$

where count(n-gram) is the n-gram co-occurrence in a translated sentence and a corresponding reference translation. count$_{sys}$(n-gram) is the n-gram count in the translated sentence only. p_n is the modified n-gram precision. BP is the brevity factor to penalize shorter translation than the reference translation where L_{ref} and L_{sys} denote the total length of the reference and the system translation on the test set respectively. Usually, N is set to 4, known as the 4-gram BLEU score.

The NIST metric (Doddington, 2002) attempts to weight the n-gram co-occurrence based on information since some n-gram may be more informative than the others. The formulation is defined as follows:

$$\text{NIST} = \sum_{n=1}^{N} \text{BP} \cdot \frac{\sum_{\text{all n-grams that co-occur}} \text{info(n-gram)}}{\sum_{\text{n-gram} \in s_i} 1} \quad (3.23)$$

$$\text{info(n-gram)} = \log_2 \frac{\text{count((n-1)gram)}}{\text{count(n-gram)}} \quad (3.24)$$

$$\text{BP} = e^{\text{b} \cdot \log_2 \min(\frac{L_{sys}}{L_{ref}}, 1)} \quad (3.25)$$

where count(n-gram) is the count of occurrences of $(w_1\ w_2\ \ldots\ w_n)$ and count((n-1)gram) is the count of occurrences of $(w_1\ w_2\ \ldots\ w_{n-1})$ in all reference translations. b is chosen to make the brevity penalty (BP) equal to 0.5 when the number of words in the system output is 2/3rds of the average number of words in the reference translation.

3.4.7 Chinese-To-English Translation System

Our baseline systems include a small-scale RT04 system (Tam et al., 2007a), medium-scale GALE development system, and large-scale GALE-P2.5 SMT system (Hildebrand

Table 3.10: Size of the language model training corpora and the parallel training corpora for phrase extraction in terms of number of words.

	LM training		// training corpora	
SMT System	English	Source	Chinese	English
RT04	80M	Donga, Xinhua 2004	35M	43M
GALE-dev	500M	Xinhua (1995-2006)	59M	67M
GALE-P2.5	2.7B	Gigaword V3 (LDC2007T07)	232M	260M

et al., 2008) translating from Chinese to English. The parallel training corpora for system development are shown in Table 3.10. Part of the Chinese-English bilingual corpora for the GALE system are available from the LDC [5].

The RT04 system employs online phrase extraction using the PESA approach (phrase pair extraction as sentence splitting) (Vogel, 2005). To facilitate the efficiency of online phrase extraction, parallel training sentences are indexed via a suffix array (Manber and Myers, 1993; Zhang and Vogel, 2006) and pre-loaded into memory before decoding. The IBM Model-1 lexicon is used for scoring the phrase pairs during decoding.

For the medium-scale and large-scale GALE systems, the IBM Model-4 is used for word alignment using a parallel version of GIZA++ toolkit (Och and Ney, 2003). Phrase extraction and scoring are performed using the Moses toolkit (Koehn et al., 2007)

4-gram English language models are employed for the RT04 and the GALE development systems while a 5-gram language model is trained for the GALE-P2.5 system. The text pre-processing steps include tokenization on the English side and on the Chinese side: automatic word segmentation using a revised version of the Stanford Chinese word segmenter (Tseng et al., 2005), replacement of traditional Chinese characters by their simplified equivalent and 2byte to 1byte ASCII character normalization. Sentence pairs with unbalanced sentence length are removed from the training corpora.

[5] http://projects.ldc.upenn.edu/gale/data/catalog.html, last consulted 2 April 2009

Table 3.11: Statistics of the development sets and the test sets for statistical machine translation containing newsgroup (NG), newswire (NW) and broadcast news (BN). Eval07u.BN stands for the unsequestered BN portion of Eval07 for speech translation. Confusion network (CN) is used to represent multiple translation options of a target phrase in Eval07 test set.

Set	Sentence	Document/Show	Reference	Genre
RT04-dev	272	4	1	BN
RT04-eval	522	3	1	BN, BC
MT03 (dev)	919	100	4	NW
MT06 (test)	1664	79	4	NG, NW, BN
Eval07u.BN (ASR test)	314	32	CN	BN

Decoding

Decoding is performed by constructing a translation lattice which contains all possible matched bilingual phrase pairs of an input source sentence. In the GALE-P2.5 system, part-of-speech based word re-ordering (Rottmann and Vogel, 2007) is performed on an input sentence to produce an input source lattice before building the translation lattice. Search is then performed on this lattice using our STTK beam-search decoder (Vogel et al., 2003). The word re-ordering window is set to 4 and 3 for the RT04 system and the medium-scale system respectively while monotonic decoding is applied for the GALE-P2.5 system since word re-ordering is already applied in the source lattices. An optimal path is returned with a maximum translation score consisting of a log-linear combination of feature functions including a language model probability, distortion penalty, word-count penalty, phrase count and phrase-alignment scores.

Development/Evaluation Set

Table 3.11 shows the development sets and the evaluation sets. The RT04 sets are used to tune and evaluate the RT04 system while MT03 and MT06 sets are used to tune and

Table 3.12: Configurations of the baseline statistical machine translation systems.

	RT04	GALE-dev	GALE-P2.5
Scale	small	medium	large
Translation model	phrase-based (online)	phrase-based	phrase-based
Word alignment	-	IBM Model 4	IBM Model 4
Target language model	4-gram	4-gram	5-gram
Distortion model	distance-based	distance-based	part-of-speech
Input format	sentence	sentence	word lattice
# feature functions	10	8	9

evaluate the GALE systems respectively. The Mandarin RT04 sets are originally designed for the RT04 broadcast news evaluation for automatic speech recognition. MT03 contains newswire documents while MT06 comprises other genre such as newsgroup and broadcast news. The development sets are used to tune the weights of the feature functions via minimum error rate training. For end-to-end speech translation, we use the unsequestered broadcast news portion of Mandarin Eval07 for evaluation. The English translation reference of Eval07 employs a confusion-network-like representation to encapsulate multiple translation options of an English phrase. For instance, the following sentence "*Over 50 car models with domestic brands have reduced their prices [at the same time//collectively//all].*" encapsulates different translation options compactly using a confusion network.

We perform significance testing using a bootstrapping approach (Zhang and Vogel, 2004) that repeatedly draws random subsets of translated sentences from a baseline system so that the score of a performance metric is computed for each random subset. As a result, an empirical distribution over the scores is formed and a 95% confidence interval with respect to the baseline system is constructed. We report statistical significance when the score of an alternative approach exceeds the baseline confidence interval.

Table 3.12 shows the summary of the baseline RT04, GALE development and GALE-P2.5 systems.

3.5 Summary

We have described our baseline Mandarin and Arabic transcription systems and the baseline Chinese-to-English statistical machine translation systems. Our systems employ the current state-of-the-art techniques for training and decoding. In the following chapters, we describe our unified topic adaptation framework for the baseline transcription and translation systems.

Chapter 4

Monolingual N-gram LSA Based Language Model Adaptation

We present unsupervised language model adaptation based on latent semantic analysis. Firstly, we propose a topic caching approach that caches topic counts of a word context in contrast to traditional word caching that caches word counts. We introduce incremental marginal adaptation for lattice rescoring that is analogous to full marginal adaptation on a background language model. We propose latent Dirichlet-Tree allocation for modeling topic correlation to generalize latent Dirichlet allocation for latent semantic analysis. Lastly, we extend latent Dirichlet-Tree allocation to its N-gram version to relax the "bag-of-word" assumption and address the model training and smoothing issues. We evaluate our approaches for unsupervised language model adaptation on large scale GALE evaluations on Mandarin and Arabic.

4.1 Topic Caching

Cache-based language model (Kuhn and Mori, 1990; Clarkson and Robinson, 1997) enables rapid language model adaptation by capturing the dynamics of natural languages via caching the frequency of the recently occurred words. This approach is computationally

efficient since only word counts are needed to store and manipulate online. It offers significant improvement in word perplexity for supervised language model adaptation. However, caching the word counts from speech recognition hypotheses is not appropriate for *unsupervised* language model adaptation since the probability of a mis-recognized word is increased after word caching. We present a topic caching approach via latent Dirichlet allocation (Tam and Schultz, 2005) that employs a Dirichlet prior. The Dirichlet prior can be interpreted as a dynamic cache to store the fractional topic counts in the E-steps described in Section 2.2.3:

E-steps:

$$\gamma_k = \alpha_k + \sum_{i \in h} q(z_i = k) \tag{4.1}$$

$$q(z_i = k) \propto e^{E_q[\log \theta_k]} \cdot p(w_i|k) \tag{4.2}$$

where α_k denotes the prior pseudo-count for topic k and $q(z_i = k)$ is the fractional topic count of the i-th word in the context h. Equation 4.1 means caching the fractional topic counts from a word context. After topic caching, we generate an adapted unigram language model via linear interpolation as follows:

$$p_{lda}(w|h) = \int_\theta \sum_{k=1}^{K} p(w|k) \cdot p(k|\theta) \cdot q(\theta|h; \{\gamma_k\}) \tag{4.3}$$

$$= \sum_{k=1}^{K} p(w|k) \cdot \int_\theta \theta_k \cdot q(\theta|h; \{\gamma_k\}) \tag{4.4}$$

$$= \sum_{k=1}^{K} p(w|k) \cdot E_q[\theta_k|h] \tag{4.5}$$

$$\text{where } E_q[\theta_k|h] = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k} \quad (k = 1...K) \tag{4.6}$$

$q(\theta|h; \{\gamma_k\})$ denotes a variational Dirichlet posterior over the topic mixture weights θ with K topics. Given the word hypotheses decoded from past speech utterances, unsupervised language model adaptation can be performed as follows:

1. Cache the fractional topic counts.

2. Re-compute an adapted unigram model.

3. Update the topic cache table $\{\alpha_k\}$ in the Dirichlet prior as:

$$\alpha_k \leftarrow \lambda \cdot \alpha_k + \sum_{i \in h} c_i \cdot q(z_i = k) \tag{4.7}$$

   where $\lambda \in [0, 1]$ is a scaling factor of the history which can be tuned on a heldout set, and c_i denotes the confidence score of the i-th word from the context buffer h.

4. Perform (log) linear interpolation with a background language model and then decode the next utterance.

5. If a topic boundary (e.g. at the end of an audio show) is given, clear the context buffer and reset the cache table to the background $\{\alpha_k\}$.

Discounting the prior counts with λ in equation 4.7 is desirable since an audio show, such as in broadcast news, can contain multiple independent stories and the information from the past utterances crossing an unknown topic boundary are irrelevant to the current topic.

For example, the following sentences extracted from CCTV audio news are marked with topic boundaries at the sentence level using latent Dirichlet allocation:

okay let 's break in to a piece of news that we just received

<TOPIC BOUNDARY>

according to a report by south korean ytn cable tv two trains carrying flammable materials collided and exploded at the ryongchon train station in p yong - an - buk - do in north korea at 1 pm on the 22 nd local time which was 2 pm beijing time

the explosion might have killed several thousands of people and injured 3000 others

<TOPIC BOUNDARY>

okay let 's continue our focus on financial news

Moreover, the automatic topic assignments can vary within a sentence:

Table 4.1: Sample latent topics extracted from latent Dirichlet allocation.

Latent topic	Top words (translated from Chinese)
"economy"	development, economy, country, society, world, globe
"sport"	competition, candidate, rank, sport, result, champion
"health"	disease, therapy, AIDS, hospital, health, patient, people
"technology"	company, information, network, system, technology
"education"	hong kong, education, mainland, student, expert

<topic 1> *okay let 's continue our focus on* </topic 1> <topic 2> *financial* </topic 2><topic 3> *news* </topic 3>

The computational complexity of the E-steps is O(TMK) where T denotes the number of iterations in the E-steps and M denotes the number of words in the context buffer. Log-linear interpolation is computationally efficient during decoding since the scores are usually expressed in logarithm and thus the computation only involves few floating point operations.

4.1.1 Experiment

We compared topic caching with word caching for incremental unsupervised language model adaptation on different language model training scenarios from scarce data (1M words) to large data (300M words) drawn from the Chinese Gigaword corpora V1. The word cache-based language model was a unigram language model that dynamically adapted to the past decoded hypotheses using the decaying word counts, and was then interpolated with the background trigram language model. We trained a background trigram language model and latent Dirichlet allocation using the same amount of data on each test scenario. The training corpora for latent Dirichlet allocation were organized into documents where each document was roughly a piece of news story annotated in the corpora. We did not remove function words from training otherwise the unigram probability of function words would be under-estimated. Table 4.1 shows examples of latent topics in latent Dirichlet allocation sorted by the unigram probability $p(w|k)$. The number of latent topics K was

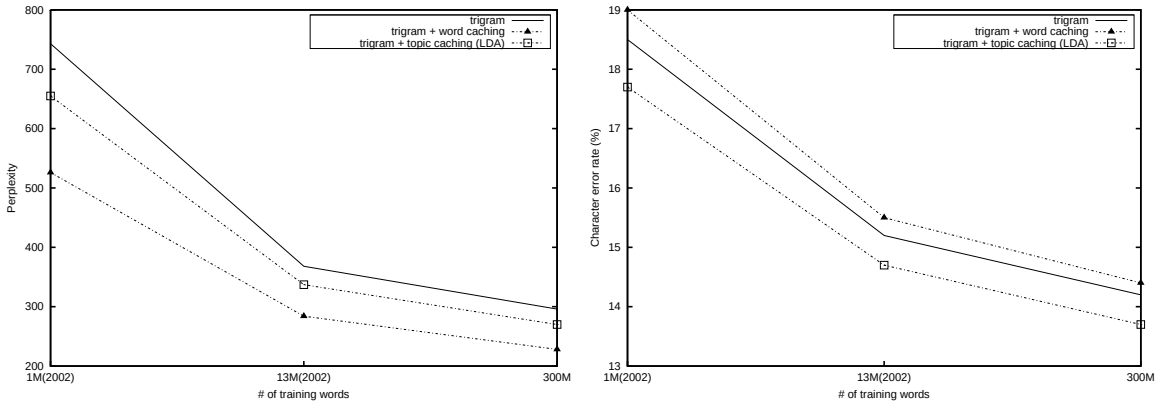

Figure 4.1: Perplexity (Left) and the character error rate (Right) for topic and word caching on CCTV of the RT04 test set.

set to 50 motivated by (Blei et al., 2003).

We used the official Mandarin RT04 development set for parameter tuning such as the interpolation weight between the background language model and the dynamic unigram language model generated from latent Dirichlet allocation, and the history scaling factor λ. Optimal weight for the trigram language model was between $0.7-0.9$ and the word history scaling factor λ was between $0.3-0.4$ based on word perplexity of the RT04 development set. We employed our Mandarin RT04 transcription system to decode the CCTV show of the RT04 test set. Since the topic boundary was not given in the test set, the word history buffer was cleared only at the end of the audio file.

4.1.2 Results

Figure 4.1 shows the language model adaptation performance on perplexity and character error rate with different sizes of the training corpora. Although the word caching approach was more effective in reducing perplexity compared to topic caching, degradation in character error rate was obtained in Figure 4.1 (Right) , which corresponded to the results reported in (Clarkson and Robinson, 1998). On the other hand, topic caching reduced the recognition error rate on different training scenarios. When the language model training data are scarce, the recognition performance degrades and the word hypotheses contain

LM (300M)	CCTV	NTDTV	RFA	ALL	Rel. Δ
background	13.1%	17.5	35.7	21.5	-
50 topic	13.1	17.6	35.2	21.4	0.5
100 topic	13.2	17.3	34.6	21.1*	1.9
200 topic	**12.7**	17.1	34.9	**21.0***	2.3
300 topic	13.2	**17.0**	34.7	21.1*	1.9
400 topic	12.9	17.3	34.7	21.1*	1.9
500 topic	13.1	17.2	**34.4**	**21.0***	2.3

Table 4.2: Character error rate (%) on the RT04 test set with different number of topics in latent Dirichlet allocation using the GALE-P1-dryrun Mandarin transcription system. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported. * denotes that the approach is statistically significant at $\leq 5\%$ significance level compared to the unadapted baseline.

more recognition errors. However, topic caching still achieved improvement in the recognition performance compared to the unadapted baseline. The observation suggests that topic caching is more robust against speech recognition errors than word caching, making it suitable for unsupervised language model adaptation.

4.1.3 Optimal Number of Topics

The next question is the optimal number of topics for latent Dirichlet allocation. Empirically, we varied the number of topics for LSA training from 50 to 500 and performed unsupervised language model adaptation using the GALE-P1-dryrun Mandarin transcription system. All models were trained on the same training corpora with 300M Chinese words. We compared the recognition performance on the full RT04 test set having three audio shows: CCTV, NTDTV and RFA.

Table 4.2 shows the character error rate with different number of topics for topic caching. Better recognition performance was achieved by increasing the number of topics. The optimal number of topics was 200 in terms of optimal recognition performance and

minimal model size. The error reduction was statistically significant at a $\leq 5\%$ significance level compared to the unadapted baseline. The setting of $K = 200$ is employed for the rest of the experiments.

4.2 Latent Dirichlet-Tree Allocation

One assumption in latent Dirichlet allocation is the use of a Dirichlet prior, which asserts that the topics are independent. 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 in a newspaper website are usually organized into the main-topic and sub-topic hierarchy. Intuitively, it would be advantageous to model the topic correlation, which motivates the extension of latent Dirichlet allocation into latent Dirichlet-Tree allocation (LDTA) (Tam and Schultz, 2007b). Latent Dirichlet-Tree allocation captures the topic correlation via a structural Dirichlet-Tree prior (Minka, 1999; III, 1991). 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 with depth one. Sampling from a Dirichlet distribution becomes labeling the branches under a node with probability values summing to unity. In general, a Dirichlet-Tree can have different depth and structure. Figure 4.2 illustrates a depth-two Dirichlet-Tree where the root node is a Dirichlet distribution with more than two branches while the Dirichlet nodes at the bottom only allow binary branches.

Given a Dirichlet-Tree of a fixed structure parametrized by a set of Dirichlet parameters $\{\alpha_{jc}\}$, a document w_1^N is generated as follows:

1. Sample a vector of branch probabilities $b_j \sim Dirichlet(\cdot; \{\alpha_{jc}\})$ for each node $j = 1...J$ where $\{\alpha_{jc}\}$ denotes the parameter of a Dirichlet distribution at node j, that is, the pseudo-counts of the outgoing branch c at node j.

2. Compute the topic distribution as in equation 4.8,

$$\theta_k = \prod_{jc} b_{jc}^{\delta_{jc}(k)} \tag{4.8}$$

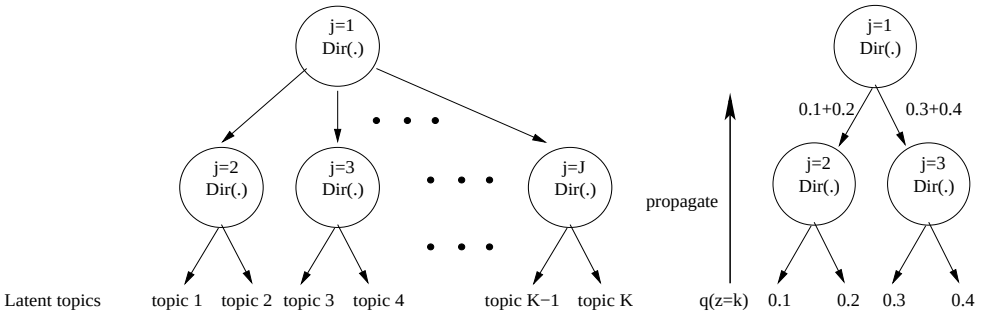

Figure 4.2: Left: Dirichlet-Tree prior of depth 2: Each internal node is represented by a Dirichlet distribution over the branches. Right: Variational E-step as bottom-up propagation and summation of fractional topic counts.

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 their child branches b_j) and an observed document w_1^N can be written as equation 4.9,

$$p(w_1^N, z_1^N, b_1^J) \;=\; p(b_1^J|\{\alpha_{jc}\}) \prod_{i=1}^{N} p(w_i|z_i) \cdot \theta_{z_i} \tag{4.9}$$

where

$$p(b_1^J|\{\alpha_{jc}\}) \;=\; \prod_{j=1}^{J} \text{Dirichlet}(b_j; \{\alpha_{jc}\}) \propto \prod_{jc} b_{jc}^{\alpha_{jc}-1} \tag{4.10}$$

Similar to training latent Dirichlet allocation, we apply variational Bayes to optimize the lower bound of the marginalized document likelihood using the Jensen's inequality (equation 4.11):

$$\begin{aligned}
\log p(w_1^N;\Lambda) &= \log \int_{b_1^J} \sum_{z_1^N} q(z_1^N, b_1^J;\Gamma) \cdot \frac{p(w_1^N, z_1^N, b_1^J;\Lambda)}{q(z_1^N, b_1^J;\Gamma)} && (4.11)\\
&\geq \int_{b_1^J} \sum_{z_1^N} q(z_1^N, b_1^J;\Gamma) \cdot \log \frac{p(w_1^N, z_1^N, b_1^J;\Lambda)}{q(z_1^N, b_1^J;\Gamma)} && (4.12)\\
&= Q(w_1^N;\Lambda,\Gamma) && (4.13)
\end{aligned}$$

where

$$\begin{aligned}
Q(w_1^N;\Lambda,\Gamma) &= E_q[\log \frac{p(w_1^N, z_1^N, b_1^J;\Lambda)}{q(z_1^N, b_1^J;\Gamma)}] && (4.14)\\
&= E_q[\log p(w_1^N|z_1^N)] + E_q[\log \frac{p(z_1^N|b_1^J)}{q(z_1^N)}] + E_q[\log \frac{p(b_1^J;\{\alpha_j\})}{q(b_1^J;\{\gamma_j\})}] && (4.15)
\end{aligned}$$

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

$$\begin{aligned}
p(b_1^J|z_1^N) &\propto p(z_1^N|b_1^J) \cdot p(b_1^J;\{\alpha_{jc}\}) && (4.16)\\
&\propto \left(\prod_{i=1}^N \prod_{jc} b_{jc}^{\delta_{jc}(z_i)}\right) \cdot \prod_{jc} b_{jc}^{\alpha_{jc}-1} && (4.17)\\
&= \prod_{jc} b_{jc}^{\left(\alpha_{jc}+\sum_{i=1}^N \delta_{jc}(z_i)\right)-1} && (4.18)\\
&= \prod_{j=1}^J Dirichlet(b_j;\{\gamma'_{jc}\}) && (4.19)
\end{aligned}$$

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, the E-steps of latent Dirichlet-Tree allocation is similar to latent Dirichlet allocation:

E-steps:

$$\gamma_{jc} = \alpha_{jc} + \sum_{i=1}^{N} E_q[\delta_{jc}(z_i)] \tag{4.20}$$

$$= \alpha_{jc} + \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{q_{ik}} \cdot \delta_{jc}(k) \tag{4.21}$$

$$q_{ik} \propto p(w_i|k) \cdot e^{E_q[\log \theta_k; \{\gamma_{jc}\}]} \tag{4.22}$$

where

$$E_q[\log \theta_k] = \sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}] \tag{4.23}$$

$$= \sum_{jc} \delta_{jc}(k) \left(\Psi(\gamma_{jc}) - \Psi(\sum_{c} \gamma_{jc}) \right) \tag{4.24}$$

where q_{ik} denotes $q(z_i = k|w_1^N)$ meaning the variational topic posterior of word w_i. Equation 4.20 and equation 4.22 are executed iteratively until convergence is reached. Equation 4.20 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 Figure 4.2 (Right).

M-step:

$$\hat{p}(w|k) \propto \sum_{i=1}^{N} q_{ik} \cdot \delta(w_i, w) \propto C_k(w) \tag{4.25}$$

The re-estimation formula for $\{p(w|k)\}$ is the weighted relative word frequency in equation 4.25 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. Appendix B provides a full derivation based on variational Expectation-Maximization algorithm.

Table 4.3: Sample contiguous fragment of latent topics extracted from latent Dirichlet-Tree allocation.

Latent topic index	Top words (translated from Chinese)
"topic-61"	education, student, school, teacher, learning
"topic-62"	university, expert, high-level, education, training
"topic-63"	employment, expert, labor, work, career
"topic-64"	research, china, science, technology, scientist
"topic-65"	gene, human, clone, research, biology
"topic-66"	research, discover, cell, gene, treatment
"topic-67"	transplant, surgery, patient, liver, hospital
"topic-68"	information, network, service, web, client
"topic-69"	system, computer, technology, computer, chip, software

4.2.1 Experiment

We compared latent Dirichlet-Tree allocation with latent Dirichlet allocation via unsupervised marginal adaptation. For rapid benchmarking, we first employed the small-scale Mandarin RT04 transcription system for experiments followed by a large-scale evaluation using the Mandarin GALE-P1 transcription system. The LSA marginals were computed separately for each approach at the show level on the RT04 test set. Then the LSA marginals were employed for marginal adaptation. For latent Dirichlet-Tree allocation, a balanced binary tree was employed.

Table 4.3 shows the correlated topics extracted via latent Dirichlet-Tree allocation. We observe contiguous fragments of correlated topics corresponding to the leaf nodes of the tree from a left to right fashion. Topics 61–63 are closely related to a general topic "education" and topics 68–69 are closely related to a general topic "information technology". The results suggest that the tree structure enforces proximity constraint over the topics.

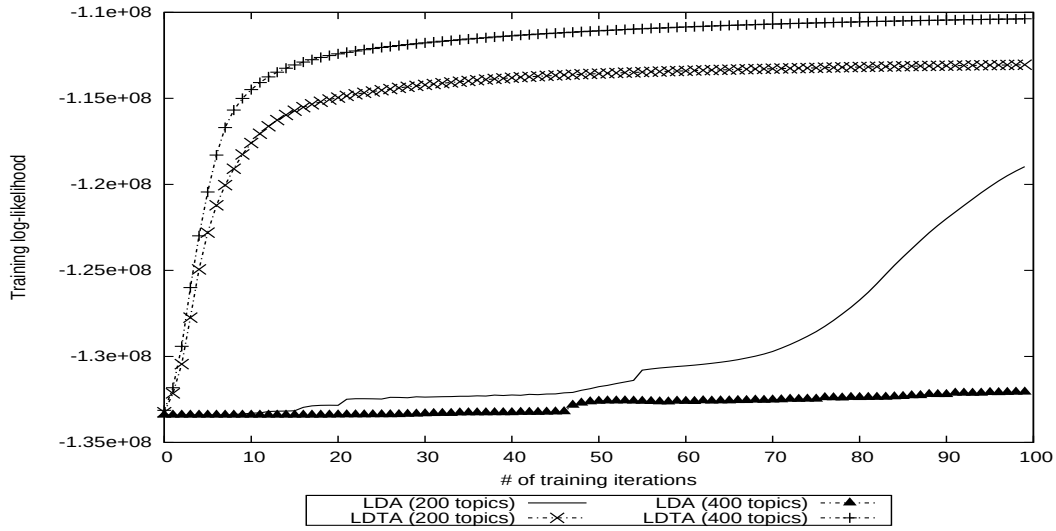

Figure 4.3: Training log-likelihood of latent Dirichlet allocation (LDA) and latent Dirichlet-Tree allocation (LDTA) using the Xinhua News 2002 corpora (13M words).

4.2.2 Training Convergence

Figure 4.3 shows the training convergence of latent Dirichlet allocation and latent Dirichlet-Tree allocation in terms of the training log likelihood using 200 and 400 topics. Both training approaches started with the same model $p(w|k)$ that were initialized with uniform distributions while their prior distributions were initialized randomly. Latent Dirichlet-Tree allocation converged significantly faster than latent Dirichlet allocation in terms of the number of training iterations. This effect was more significant when the number of topics increased to 400. The rapid convergence is attributed to the structured Dirichlet-Tree prior that restricts the model space compared to the unstructured Dirichlet prior. In other words, an observed topic triggers its correlated topics via the tree structure while the topic independence assumption in the Dirichlet prior lacks this effect. We conclude that latent Dirichlet-Tree allocation is useful because it does not suffer from model initialization issue.

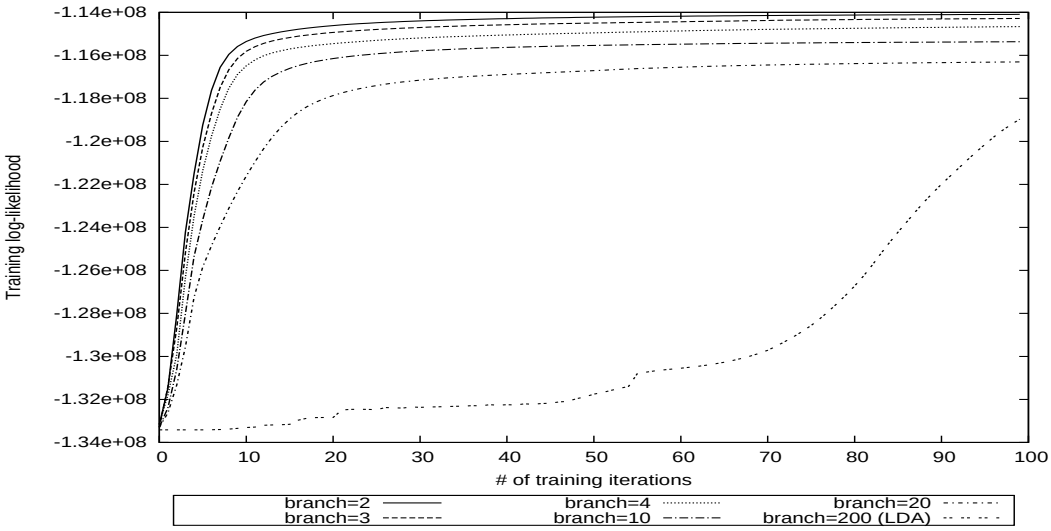

Figure 4.4: Training log-likelihood of latent Dirichlet-Tree allocation with different number of branches in a Dirichlet node using the Xinhua News 2002 corpora (13M words).

4.2.3 Effect of Dirichlet-Tree Structure

Figure 4.4 illustrates the effect of the tree structure in terms of the training likelihood by varying the number of branches in a Dirichlet node. Results show that the training convergence is optimal when a balanced binary tree is employed. As the number of branches increases, the independence assumption among topics becomes stronger and thus slowing down the training convergence. In an extreme case in latent Dirichlet allocation, the training convergence is the slowest.

4.2.4 Results

Table 4.4 shows the word perplexity and the character error rate after LSA marginal adaptation using a small-scale Mandarin RT04 transcription system trained on the 13M corpora. Latent Dirichlet-Tree allocation reduces the overall perplexity and character error rate relatively by 7–12% and 2% respectively compared to latent Dirichlet allocation, and by 10%–17.5% and 4.0% compared to the unadapted 4-gram language model. Latent Dirichlet-Tree allocation performs better than latent Dirichlet allocation and the unadapted

Table 4.4: Marginal adaptation results on character error rate (word perplexity) on the RT04 test set using the small-scale Mandarin RT04 ASR system (13M). Latent Dirichlet allocation (LDA) and latent Dirichlet-Tree allocation (LDTA) were compared. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported. * denotes that the approach is statistically significant at $\leq$ 5% significance level compared to the unadapted baseline.

LM (13M)	CCTV	NTDTV	RFA	ALL	Rel. Δ
background	15.6% (748)	22.1 (1718)	40.0 (3655)	25.3	-
+LDA (100 iter)	15.1 (695)	21.7 (1669)	39.6 (3451)	24.8*	2.0 (5.6)
+LDTA (100 iter)	14.4 (629)	21.5 (1547)	38.9 (3015)	24.3*	4.0 (17.5)

baseline at a $\leq$ 5% significance level.

Table 4.5 shows the adaptation performance for a large-scale evaluation using the Mandarin GALE-P1 transcription system trained on the 800M corpora. Due to the limitation of computation resources, we performed only 20 and 50 training iterations for latent Dirichlet-Tree allocation and latent Dirichlet allocation respectively. Due to the slow training convergence, latent Dirichlet allocation required more training iterations before yielding an acceptable adaptation performance. Latent Dirichlet-Tree allocation yielded 5% relative perplexity reduction compared to latent Dirichlet allocation with no degradation in the overall character error rate. Compared to the unadapted baseline, latent Dirichlet-Tree allocation reduced the relative perplexity and the character error rate by 8.9%–14.5% and 2.5% respectively, which were statistically significantly at a $\leq$ 5% significance level. Therefore, we use latent Dirichlet-Tree allocation for the rest of our experiments due to its stable performance in terms of training convergence and language model adaptation.

4.3 Incremental Marginal Adaptation

Marginal adaptation is useful for integrating an in-domain knowledge via latent semantic marginals (Federico, 2002; Tam and Schultz, 2006). Incremental marginal adaptation for

Table 4.5: Marginal adaptation results on character error rates (word perplexity) on the RT04 test set using the Mandarin GALE-P1 ASR system (800M). Latent Dirichlet allocation (LDA) and latent Dirichlet-Tree allocation (LDTA) were compared. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported. * denotes that the approach is statistically significant at $\leq 5\%$ significance level compared to the unadapted baseline.

LM (800M)	CCTV	NTDTV	RFA	ALL	Rel. Δ
background	8.3% (359)	14.4 (868)	26.3 (778)	15.9	-
+LDA (50 iter)	8.1 (332)	14.0 (834)	25.6 (703)	15.5*	2.5 (9.6)
+LDTA (20 iter)	8.3 (313)	14.2 (791)	25.3 (665)	15.5*	2.5 (14.5)

decoding is computationally expensive due to the computation of the normalization factor for all N-gram entries in a background language model. Therefore, marginal adaptation is usually applied for domain adaptation where the background language model is adapted offline. We propose an incremental marginal adaptation approach for lattice rescoring. The cost of marginal adaptation for lattice rescoring is computationally inexpensive since only a few outgoing word links actually emerge from a context node in a lattice. Thus, computing the normalization factor can be done efficiently. The adapted language model scores for each word link (i, j) is analogous to equation 2.71 for full marginal adaptation:

$$lm_a(i,j) = \frac{\alpha(w_{ij}) \cdot lm_{bg}(i,j)}{\sum_{j' \in Out(i)} \alpha(w_{ij'}) \cdot lm_{bg}(i,j')} \cdot Mass(i) \tag{4.26}$$

$$\text{where } Mass(i) = \sum_{j' \in Out(i)} lm_{bg}(i,j') = \sum_{j' \in Out(i)} lm_a(i,j') \tag{4.27}$$

$$\text{and } \alpha(w_{ij}) = \left(\frac{p_{lda}(w_{ij})}{p_{bg}(w_{ij})} \right)^{\epsilon} \tag{4.28}$$

where w_{ij} is the word label associated to the link (i, j). $Out(i)$ denotes a set of links from node i. $Mass(i)$ is introduced to ensure that the total probability mass from node i is conserved after adaptation, which is similar in spirit to fast marginal adaptation in equation 2.73.

	Dev07				Eval06			
	BN	BC	ALL	Rel. Δ	BN	BC	ALL	Rel. Δ
background	7.5%	18.8	13.9	-	15.1	26.8	20.4	-
+LSA decode	7.4	18.6*	13.8*	1.4	14.7*	26.5*	20.0*	2.0
+LSA decode & rescore	**7.3***	**18.1***	**13.4***	**4.3**	**14.4***	**26.2***	**19.7***	**3.4**

Table 4.6: Character error rate (%) after applying LSA for decoding (denoted as LSA decode) using the GALE-P2 Mandarin transcription system followed by incremental marginal adaptation (rescore) for lattice rescoring after cross-adapting with the IBM Mandarin transcription system on Dev07 and Eval06 test sets. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported. * denotes the approach is significantly better than the unadapted baseline at $\leq$5% level of significance.

4.3.1 Experiment

The language model adaptation experiment was performed during the GALE-P2 evaluation for Mandarin. Firstly, we cross-adapted the GALE-P2 Mandarin transcription system using the word hypotheses from the IBM Mandarin transcription system. Then we applied topic caching via latent Dirichlet-Tree allocation during decoding to generate word lattices followed by incremental LSA-marginal adaptation.

4.3.2 Results

Table 4.6 shows the recognition performance after applying topic caching (LSA decode) followed by incremental marginal adaptation (LSA rescore). Topic caching reduced the character error rate relatively by 1.4% and 2.0% on Dev07 and Eval06 respectively compared to the unadapted baseline. Incremental marginal adaptation further reduced the character error rate relatively by 2.0% and 1.5% on Dev07 and Eval06 respectively compared to topic caching. The total relative reductions after applying both techniques were 4.3% and 3.4% compared to the unadapted baseline. All reductions were statistically significant at a $\leq$ 5% significance level.

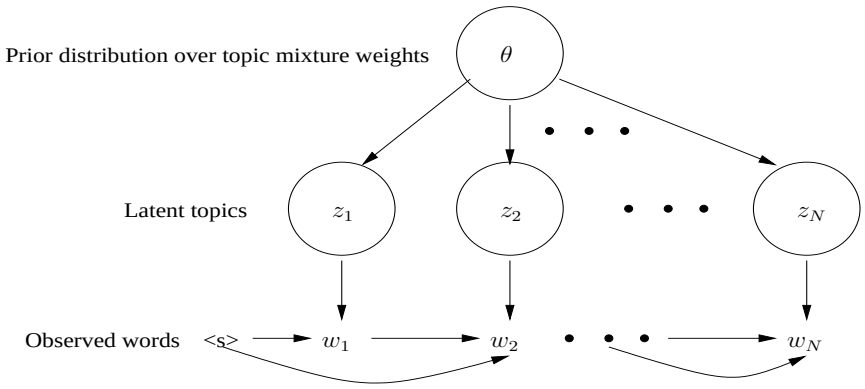

Figure 4.5: Graphical model representation of a trigram LSA.

4.4 N-gram Latent Dirichlet-Tree Allocation

One issue in latent semantic analysis is the "bag-of-word" assumption that ignores word ordering. For document classification, word ordering may not be important. But for language modeling, word ordering is crucial since a trigram language model usually outperforms a unigram language model for word prediction. In Chapter 2, we describe a bigram topic model to relax this assumption by connecting adjacent words in a document together to form a Markov chain in latent Dirichlet allocation. We present the N-gram latent Dirichlet-Tree allocation (Tam and Schultz, 2008) based on the bigram topic model (Wallach, 2006) and latent Dirichlet-Tree allocation (Tam and Schultz, 2007b). The graphical model representation of a trigram LSA is shown in Figure 4.5.

The original formulation of the bigram topic model does not address two important issues: efficient model training and smoothing. We propose an efficient training algorithm for the N-gram latent Dirichlet-Tree allocation via variational Expectation-Maximization algorithm and model bootstrapping which are scalable to large data sets in Section 4.4.1. We formulate the fractional Kneser-Ney smoothing [1] for model smoothing. Our formulation generalizes the original Kneser-Ney formulation (Kneser and Ney, 1995) which supports only integral counts in Section 4.4.2. We apply the N-gram latent Dirichlet-Tree

[1]This method was briefly mentioned in (Xu et al., 2003) without detail in a different context. (Bisani and Ney, 2008) formulated this method independently in a grapheme-to-phoneme setting.

allocation for the large-scale GALE evaluation for automatic speech recognition in this chapter, and in statistical machine translation in Chapter 5.

4.4.1 Model Training

Gibbs sampling is employed in the original bigram topic model. Despite its simplicity, it can be slow and inefficient since it usually requires hundreds of sampling iterations for convergence. We present a variational Bayes approach for model training. For simplicity, we only show the formulation for bigram LSA, but it is straightforward to generalize to N-gram LSA. The joint likelihood of a document w_1^N, the latent topic sequence z_1^N and θ using bigram LSA can be written as follows:

$$p(w_1^N, z_1^N, \theta) = p(\theta) \cdot \prod_{i=1}^{N} p(z_i|\theta) \cdot p(w_i|w_{i-1}, z_i) \tag{4.29}$$

By introducing a factorizable variational posterior distribution $q(z_1^N, \theta; \Gamma) = q(\theta) \cdot \prod_{i=1}^{N} q(z_i)$ over the latent variables and applying the Jensen's inequality, the lower bound of the marginalized document likelihood can be derived as follows:

$$\begin{aligned}
\log p(w_1^N; \Lambda, \Gamma) &= \log \int_\theta \sum_{z_1 \ldots z_N} q(z_1^N, \theta; \Gamma) \cdot \frac{p(w_1^N, z_1^N, \theta; \Lambda)}{q(z_1^N, \theta; \Gamma)} \\
&\geq \int_\theta \sum_{z_1 \ldots z_N} q(z_1^N, \theta; \Gamma) \cdot \log \frac{p(w_1^N, z_1^N, \theta; \Lambda)}{q(z_1^N, \theta; \Gamma)} \\
&= E_q[\log \frac{p(\theta)}{q(\theta)}] + \sum_{i=1}^{N} E_q[\log \frac{p(z_i|\theta)}{q(z_i)}] + \sum_{i=1}^{N} E_q[\log p(w_i|w_{i-1}, z_i)] \\
&= Q(w_1^N; \Lambda, \Gamma)
\end{aligned}$$

where the expectation is taken using the variational posterior $q(z_1^N, \theta)$. For the E-steps, we compute the partial derivative of the auxiliary function $Q(\cdot)$ with respect to $q(z_i)$ and the parameter γ_{jc} in the Dirichlet-Tree posterior $q(\theta)$. Setting the derivatives to zero yields:

E-steps:

$$
\begin{aligned}
q(z_i = k) &\propto p(w_i|w_{i-1}, k) \cdot e^{E_q[\log \theta_k; \{\gamma_{jc}\}]} \text{ for } k = 1..K && (4.30)\\
\gamma_{jc} &= \alpha_{jc} + \sum_{i=1}^{N} E_q[\delta_{jc}(z_i)] && (4.31)\\
&= \alpha_{jc} + \sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k) \cdot \delta_{jc}(k) && (4.32)\\
\text{where } E_q[\log \theta_k] &= \sum_{jc} \delta_{jc}(k) \cdot E_q[\log b_{jc}] && (4.33)\\
&= \sum_{jc} \delta_{jc}(k) \left(\Psi(\gamma_{jc}) - \Psi(\sum_{c} \gamma_{jc}) \right) && (4.34)
\end{aligned}
$$

where equation 4.31 is motivated from the conjugate property that the Dirichlet-Tree posterior given the topic sequence z_1^N has the same form as the Dirichlet-Tree prior, which has been introduced in Section 4.2. Equation 4.30 and equation 4.31 are applied iteratively until convergence is reached. For the M-step, we compute the partial derivative of the auxiliary function $Q(\cdot)$ over all training documents d with respect to a topic bigram probability $p(v|u, k)$ and set it to zero:

M-step (unsmoothed):

$$
\begin{aligned}
p(v|u, k) &\propto \sum_{d} \sum_{i=1}^{N_d} q(z_i = k|d) \cdot \delta(w_{i-1}, u)\delta(w_i, v) && (4.35)\\
&= \frac{\sum_d C_d(u, v|k)}{\sum_d \sum_{v'=1}^{V} C_d(u, v'|k)} && (4.36)\\
&= \frac{C(u, v|k)}{\sum_{v'=1}^{V} C(u, v'|k)} && (4.37)
\end{aligned}
$$

where N_d denote the number of words in document d and $\delta(w_i, v)$ is a $0 - 1$ Kronecker Delta function to test if the i-th word in document d is vocabulary v. $C_d(u, v|k)$ denotes the fractional counts of a bigram (u, v) belonging to topic k in document d. Intuitively, equation 4.37 simply computes the relative frequency of the bigram (u, v). However, this solution is not practical since the model assigns a zero probability to an unseen bigram.

Therefore, bigram LSA should be smoothed properly. One simple approach is to use the Laplace smoothing by adding a small count δ to all the bigrams (Wallach, 2006). However, this approach can lead to worse performance since it will bias the bigram probability towards a uniform distribution when the vocabulary size V gets large. Our approach is to represent $p(v|u,k)$ as a standard backoff language model smoothed by fractional Kneser-Ney smoothing as described in Section 4.4.2.

Model initialization is crucial for variational EM. We employ a bootstrapping approach using a well-trained LSA as an initial model for bigram LSA so that $p(w_i|w_{i-1},k)$ is approximated by $p(w_i|k)$ in equation 4.30. It saves computation and avoids keeping the full bigram LSA in memory during the Expectation-Maximization training. To make the training procedure more practical, we apply bigram pruning during statistics accumulation in the M-step when the bigram count in a document is less than a threshold, say 0.1. This heuristic is reasonable since only a small portion of topics are "active" to a bigram. With the sparsity, there is no need to store K copies of accumulators for each bigram and thus reducing the memory requirement significantly. For simplicity, the pruned bigram counts are re-assigned to the most likely topic of the current document so that the counts are conserved. For practical implementation, accumulators are saved into a disk in batches for count merging using the SRILM toolkit. In the final step, each topic-dependent language model is smoothed individually using the merged count file.

4.4.2 Fractional Kneser-Ney Smoothing

The state-of-the-art smoothing for a backoff language model is based on the Kneser-Ney smoothing (Kneser and Ney, 1995). The belief of its success is due to the preservation of marginal distributions. However, the original formulation is defined only on integral counts, which is not suitable for bigram LSA using fractional counts. We investigate the fractional Kneser-Ney smoothing as a generalization of the original formulation.

The interpolated form using absolute discounting can be expressed as follows:

$$p_{KN}(v|u) = \frac{max\{C(u,v) - D, 0\}}{C(u)} + \lambda(u) \cdot p_{KN}(v) \tag{4.38}$$

where D is a discounting factor. In the original formulation, D lies between 0 and 1. But in our formulation, D can be any positive number. Intuitively, D controls the degree of smoothing. If D is set to zero, the model is unsmoothed; If D is too big, bigram counts smaller than D are pruned from the language model. $\lambda(u)$ ensures that the bigram probability sums to unity. After summing over all possible v on both sides of equation 4.38 and re-arranging terms, $\lambda(u)$ becomes:

$$
\begin{aligned}
1 &= \sum_v \frac{max\{C(u,v) - D, 0\}}{C(u)} + \lambda(u) && (4.39)\\
\Longrightarrow \lambda(u) &= 1 - \sum_v \frac{max\{C(u,v) - D, 0\}}{C(u)} && (4.40)\\
&= 1 - \sum_{v:C(u,v)>D} \frac{C(u,v) - D}{C(u)} && (4.41)\\
&= \frac{C(u) - \sum_{v:C(u,v)>D} C(u,v) + D \sum_{v:C(u,v)>D} 1}{C(u)} && (4.42)\\
&= \frac{\sum_{v:C(u,v)\leq D} C(u,v) + D \sum_{v:C(u,v)>D} 1}{C(u)} && (4.43)\\
&= \frac{C_{\leq D}(u,\cdot) + D \cdot N_{>D}(u,\cdot)}{C(u)} && (4.44)
\end{aligned}
$$

where $C_{\leq D}(u,\cdot)$ denotes the sum of bigram counts following u and smaller than D. $N_{>D}(u,\cdot)$ denotes the number of word types following u with the bigram counts bigger than D.

In the Kneser-Ney smoothing, the lower-order distribution $p_{KN}(v)$ is treated as an unknown parameter that can be estimated using the preservation of marginal distributions:

$$
\hat{p}(v) = \sum_u p_{KN}(v|u) \cdot \hat{p}(u) \qquad (4.45)
$$

where $\hat{p}(v)$ is the marginal distribution estimated from background training data so that

$\hat{p}(v) = \frac{C(v)}{\sum_{v'} C(v')}$. Therefore, we substitute equation 4.38 into equation 4.45:

$$
\begin{aligned}
C(v) &= \sum_u \left(\frac{max\{C(u,v) - D, 0\}}{C(u)} + \lambda(u) \cdot p_{KN}(v) \right) \cdot C(u) && (4.46)\\
&= \left(\sum_u max\{C(u,v) - D, 0\} \right) + p_{KN}(v) \sum_u C(u) \cdot \lambda(u) && (4.47)\\
\implies p_{KN}(v) &= \frac{C(v) - \sum_u max\{C(u,v) - D, 0\}}{\sum_u C(u) \cdot \lambda(u)} && (4.48)\\
&= \frac{C(v) - C_{>D}(\cdot, v) + D \cdot N_{>D}(\cdot, v)}{\sum_u C(u) \cdot \lambda(u)} && (4.49)\\
&= \frac{C_{\leq D}(\cdot, v) + D \cdot N_{>D}(\cdot, v)}{\sum_u C_{\leq D}(u, \cdot) + D \cdot N_{>D}(u, \cdot)} \text{ (using equation 4.44)} && (4.50)\\
&= \frac{C_{\leq D}(\cdot, v) + D \cdot N_{>D}(\cdot, v)}{\sum_{v'} C_{\leq D}(\cdot, v') + D \cdot N_{>D}(\cdot, v')} = \frac{C''(v)}{\sum_{v'} C''(v')} && (4.51)
\end{aligned}
$$

Equation 4.51 generalizes the Kneser-Ney smoothing to both integral and fractional counts. In the original formulation, $C_{\leq D}(u, \cdot)$ equals to zero since each observed bigram count must be at least one by definition with D less than one. As a result, the D term cancels out yielding the original formulation that counts the number of word type preceding v and thus recovering the original formulation. Intuitively, the numerator in equation 4.51 measures the total discounts of the observed bigrams ending at v. In other words, the fractional Kneser-Ney smoothing estimates the lower-order probability distribution using the relative frequency over *discounts* instead of word counts. With this approach, each topic-dependent language model in bigram LSA can be smoothed using our formulation.

In general, the fractional Kneser-Ney smoothing can be applied to a higher-order back-off language model including a factored language model via propagating the discounts to the lower-order distribution for estimation. Figure 4.6 illustrates that discounts are propagated from a trigram language model to a bigram language model, and then from a bigram language model to a unigram language model for model estimation in a recursive manner. In this case, the modified bigram and unigram counts, and the corresponding models are

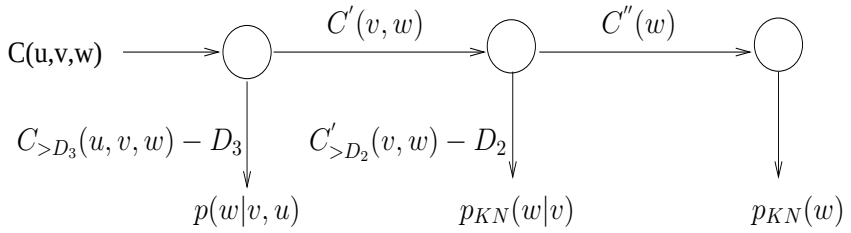

Figure 4.6: Fractional Kneser-Ney smoothing via propagation of discounts from a trigram language model to a lower-order bigram and a unigram language model.

computed as follows:

$$C'(v,w) = C_{\leq D_3}(\cdot,v,w) + D_3 \cdot N_{>D}(\cdot,v,w) \quad (4.52)$$

$$C''(w) = C'_{\leq D_2}(\cdot,w) + D_2 \cdot N'_{>D_2}(\cdot,w) \quad (4.53)$$

and

$$p_{KN}(w|v) = \frac{C'(v,w) - D_2}{C'(v,\cdot)} \quad (4.54)$$

$$p_{KN}(w) = \frac{C''(w)}{C''(\cdot)} \quad (4.55)$$

where D_i denotes the discounting constant at each language model order. As a sanity check, as D_3 goes to infinity, trigram LSA should fall back to bigram LSA. In this case, one can tune on the discounting constants so that trigram LSA would perform at least as well as bigram LSA.

4.4.3 Two-stage Unsupervised Language Model Adaptation

Unsupervised language model adaptation is performed by first inferring a topic distribution using word hypotheses from the first-pass decoding via variational inference in equation 4.30–4.31. Relative frequency over the branch posterior counts γ_{jc} is applied on each Dirichlet node j. The maximum a posteriori topic mixture weight $\hat{\theta}$ and the adapted

unigram, bigram and trigram LSA models are computed as follows:

$$\hat{\theta}_k \propto \prod_{jc} \left(\frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \quad \text{for } k = 1...K \tag{4.56}$$

$$p_a(w) = \sum_{k=1}^{K} p(w|k) \cdot \hat{\theta}_k \tag{4.57}$$

$$p_a(w|v) = \sum_{k=1}^{K} p(w|v,k) \cdot \hat{\theta}_k \tag{4.58}$$

$$p_a(w|u,v) = \sum_{k=1}^{K} p(w|u,v,k) \cdot \hat{\theta}_k \tag{4.59}$$

The LSA marginals are integrated into a background N-gram language model $p_{bg}(w|h)$ via marginal adaptation as follows:

$$p_a^{(1)}(w|h) \propto \left(\frac{p_a(w)}{p_{bg}(w)} \right)^{\epsilon} \cdot p_{bg}(w|h) \tag{4.60}$$

Marginal adaptation has a close connection to maximum entropy modeling since the marginal constraints can be encoded as unigram features. Intuitively, bigram LSA would be integrated in the same fashion by introducing bigram marginal constraints. However, we found that integrating bigram features via marginal adaptation did not offer further improvement compared to only integrating unigram features. Marginal adaptation corresponds to only one iteration of generalized iterative scaling (GIS). Due to millions of bigram features, one GIS iteration may not be sufficient for convergence. On the other hand, simple linear interpolation is effective in our experiment. The final language model adaptation formula is provided using equation 4.57– 4.60 as a two-stage process:

$$p_a^{(2)}(w|h) = \lambda_1 \cdot p_a^{(1)}(w|h) + \lambda_2 \cdot p_a(w|v) + \lambda_3 \cdot p_a(w|u,v) \tag{4.61}$$

$$\text{where } \lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{4.62}$$

$$\text{and } \lambda_i \geq 0 \ \ \forall i \tag{4.63}$$

where $\{\lambda_i\}$ are tuned to optimize the performance on a development set.

4.4.4 Experiment

As motivated from the previous experiments, the number of latent topics were set to 200 for LSA, bigram LSA and trigram LSA unless specified. The discounting factor D for the fractional Kneser-Ney smoothing was set to 0.4 for bigram LSA while the higher-order discounting factor for trigram LSA was set to 2.4 to maintain a reasonably compact model. We did a sanity check that trigram LSA fell back to bigram LSA when the higher-order discounting factor was large.

For rapid benchmarking, we first evaluated N-gram LSA using the small-scale Mandarin RT04 transcription system with unsupervised marginal adaptation at a show level similar to the experiments on latent Dirichlet-Tree allocation in Section 4.2.1. However, re-decoding was applied after language model adaptation instead of lattice rescoring.

Then, we evaluated the adaptation approach on Mandarin and Arabic languages using the GALE-P3 transcription systems. Topic caching was applied for decoding on the Mandarin system but not on the Arabic system. The word hypotheses from the final decoding passes of a discriminatively trained Initial-Final Mandarin system and a vowelized Arabic system were taken for incremental LSA-marginal adaptation and N-gram LSA lattice rescoring as described in Section 4.4.3. Dev08 and Dev07 were employed as the development set for the Mandarin and Arabic respectively. Statistical significance tests were applied to compare the LSA and the N-gram LSA performance.

4.4.5 RT04 Mandarin Results

Table 4.7 shows the correlated bigram topics sorted by the joint bigram probability $p(v|u,k) \cdot p(u|k)$. Most of the top bigrams appear either as phrases or words attached with a stopword such as 的(’s in English).

Table 4.8 shows the language model adaptation results in word perplexity and character error rate. Applying both LSA and bigram LSA yielded consistent improvement over LSA in the range of 6.4%–8.5% relative reduction in perplexity and 2.5% relative reduction in the overall character error rate. The reduction in character error rate was statistically

Table 4.7: Correlated bigram topics extracted from bigram LSA using the Xinhua news 2002 corpora (13M).

Latent topic	**Top bigrams sorted by** $p(u, v|k)$
"topic-61"	的+学生('s student), 的+教育('s education), 教育+的(education 's)
	学校+的(school 's), 少年+班(youth class), 素质+教育(quality of education)
"topic-62"	人才+培养(expert cultivation), 大学+校长(university chancellor)
	着+名(famous), 所+高校(high-school), 的+学生('s student)
"topic-63"	和+社会保障(and social security), 的+就业('s employment),
	失业+人员(unemployed officer), 就业+岗位(employment position)
"topic-64"	的+研究('s research), 专家+学者(expert people), 等+领域(etc area)
	生物+技术(biological technology), 研究+成果(research result)
"topic-65"	人类+基因组(Human DNA sequence), 的+基因('s DNA)
	生物+技术(biological technology), 胚胎+干细胞(embryo stem cell)

significant at a 0.1% significance level. We compared fractional Kneser-Ney smoothing with Witten-Bell smoothing which also supports fractional counts. The results showed that Kneser-Ney smoothing performed slightly better than Witten-Bell smoothing in word perplexity and character error rate. Increasing the number of topics from 30 to 200 in bigram LSA helped despite model sparsity. We applied extra EM iterations initialized with the bootstrapped bigram LSA but no further performance improvement was observed.

4.4.6 GALE-P3 Results

Mandarin Results

The upper section of Table 4.9 shows the overall LM adaptation results on Mandarin before cross adaptation with the IBM Mandarin system. To illustrate the effect of LSA during decoding, the first background setting was compared to the second background setting which enabled topic caching for decoding similar to the experiments in Section 4.1.1. Both set-

Table 4.8: Character error rate (word perplexity) on the Mandarin RT04 test set. Bigram LSA (biLSA) was applied in addition to LSA. Unless specified, the LSA and bigram LSA models employ 200 topics. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported. * denotes that bigram LSA is significantly better than LSA at 0.1% significance level in terms of overall character error rate.

LM (13M)	CCTV	NTDTV	RFA	ALL	Rel. Δ
background	15.3% (748)	21.8 (1718)	39.5 (3655)	24.9	-
+LSA	14.4 (629)	21.5 (1547)	38.9 (3015)	24.3	2.4
+biLSA (Kneser, K=30)	14.5 (604)	**20.7** (1502)	39.0 (2736)	24.1	3.2
+biLSA (Witten)	14.1 (594)	20.9 (1452)	38.3 (2628)	23.8	4.4
+biLSA (Kneser)	**14.0 (587)**	20.8 **(1448)**	**38.2 (2586)**	**23.7***	**4.8**

tings were then followed by lattice rescoring using LSA marginal adaptation, and linear interpolation of bigram LSA (biLSA) and trigram LSA (triLSA) with the LSA-adapted language model. Applying LSA for decoding helped on all test sets compared to the unadapted baseline. The reduction in character error rate was additive to those obtained from lattice rescoring. Moreover, applying LSA for lattice rescoring yielded further reduction in character error rate. Bigram-LSA lattice rescoring yielded additional reduction in character error rate compared to LSA rescoring which was statistically significant at a $\leq$5% significance level in all test cases. Compared to the unadapted baseline, the relative reduction in character error rate after LSA and bigram LSA adaptations were between 5%–6.9% which were statistically significant at a 0.1% significance level in all test cases. Replacing bigram-LSA with trigram LSA yielded similar recognition performance. By combining bigram-LSA and trigram-LSA together via simple score averaging, we achieved slight reduction in character error rate on Eval07r without degrading the performance on the other sets. The final relative reduction in character rate rate on Eval07r was 7.6% after applying all adaptations. We did not attempt 4-gram LSA rescoring since the additive reduction from trigram LSA was marginal. The results after cross adaptation followed a similar trend with 3.3%–6.7% relative reduction in character error rate compared to the unadapted baseline which were statistically significant at a $\leq$ 5% significance level.

Table 4.9: Lattice rescoring results in character error rate using the Mandarin GALE P3 system. Overall relative reduction (Rel. Δ) compared to the unadapted baseline (background) is reported. * denotes that bigram LSA (biLSA) and trigram LSA (triLSA) are significantly better than LSA at $\leq$5% level of significance.

Mandarin	Dev08	Rel. Δ	Eval07u	Rel. Δ	Eval07r	Rel. Δ
background	11.6%	-	14.0	-	11.8	-
+LSA	11.4	1.7	13.9	0.7	11.6	1.7
+biLSA (Kneser)	11.0*	5.2	13.6*	2.9	11.1*	5.9
background (LSA)	11.5	0.9	13.8	1.4	11.7	0.8
+LSA	11.2	3.4	13.7	2.1	11.5	2.5
+biLSA (Witten)	10.9*	6.0	**13.3***	**5.0**	11.1*	5.9
+biLSA (Kneser)	**10.8***	**6.9**	**13.3***	**5.0**	11.0*	6.8
+triLSA (Kneser)	**10.8***	**6.9**	**13.3***	**5.0**	11.1*	5.9
+bi & triLSA (Kneser)	**10.8***	**6.9**	**13.3***	**5.0**	**10.9***	**7.6**
Cross-adaptation with IBM						
background	9.0	-	10.8	-	9.0	-
+LSA	8.9	1.1	10.6	1.9	8.7	3.3
+biLSA (Kneser)	**8.6***	**4.4**	10.4	3.7	8.5*	5.6
background (LSA)	9.0	0.0	10.6	1.9	8.8	2.2
+LSA	8.8	2.2	10.5	2.8	8.6	4.4
+biLSA (Kneser)	8.7	3.3	**10.3***	**4.6**	**8.4***	**6.7**

Table 4.10: Lattice rescoring results in character error rate using the word lattices from the IBM P3 Mandarin system. Overall relative reduction (Rel. Δ) compared to the IBM system is reported.

Mandarin	Dev08	Rel. Δ	Eval07u	Rel. Δ	Eval07r	Rel. Δ
Rescoring the best single IBM system						
IBM (neural LM)	6.7	-	8.3	-	6.6	-
+biLSA (Kneser)	6.7	0.0	**8.1**	**2.4**	**6.5**	**1.5**

Table 4.11: Lattice rescoring results in word error rate using the Arabic GALE P3 system. Overall relative reduction (Rel. Δ) compared to the unadapted baseline (background) is reported. * denotes that bigram LSA (biLSA) is significantly better than LSA at $\leq$ 5% level of significance.

Arabic	Dev07	Rel. Δ	Dev08	Rel. Δ	Eval07u	Rel. Δ
background	14.3%	-	16.4	-	22.7	-
+LSA	14.2	0.7	16.4	0.0	22.7	0.0
+biLSA (Witten)	13.9	2.8	**15.9***	**3.0**	**22.4***	**1.3**
+biLSA (Kneser)	**13.8***	**3.5**	**15.9***	**3.0**	22.5	0.9
Cross-adaptation with IBM						
background	11.8	-	13.9	-	20.3	-
+LSA	11.8	0.0	13.8	0.7	20.3	0.0
+biLSA (Kneser)	**11.7**	**0.8**	**13.6***	**2.2**	**20.1**	**1.0**

During the GALE-P3 evaluation, we rescored the word lattices generated from the best single IBM Mandarin system, which was rescored with a neural network language model (Bengio et al., 2003; Schwenk, 2007). The results are shown in Table 4.10. Our adaptation approach yielded further relative reduction in character error rate by 2.4% and 1.5% on Eval07u and Eval07r respectively compared to the IBM baseline system. This implies that neural network language model and bigram LSA may capture complimentary information and thus combining two approaches yielded additional gain. Given a well-tuned state-of-the-art IBM system, the gain from bigram LSA rescoring is reasonable.

Arabic Results

The performance trend was similar on Arabic as shown in Table 4.11. LSA rescoring gave slight reduction in word error rate compared to the unadapted baseline. Moreover, bigram LSA achieved additional reduction in word error rate compared to LSA on the unseen Dev08 which was statistically significant at a 0.1% significance level. After cross adaptation with the IBM Arabic system, bigram LSA yielded 2.2% relative reduction in

Table 4.12: Lattice rescoring and system combination results in word error rate using the word lattices from the IBM P3 systems. Overall relative reduction (Rel. Δ) compared to the unadapted baseline is reported.

Arabic	Dev07	Rel. Δ	Dev08	Rel. Δ	Eval07u	Rel. Δ
The best single IBM system						
U-BN	9.9%	-	11.1	-	14.3	-
UBM	9.3	-	10.6	-	13.7	-
UBM+biLSA	**9.1**	**2.2**	10.6	0.0	**13.6**	**0.7**
IBM system combination						
UBM + U-BN	9.1	-	10.3	-	13.4	-
UBM+biLSA + U-BN	**8.9**	**2.2**	10.3	0.0	13.4	0.0

word error rate compared to the unadapted baseline which was statistically significant. The reductions on Dev07 and Eval07u were not statistically significant.

Similar to the Mandarin evaluation, we rescored the word lattices generated from the best single IBM Arabic system as shown in Table 4.12. Again, our approach achieved further reduction in word error rate by 2.2% and 0.7% relative on Dev07 and Eval07u compared to the IBM baseline system. Finally, lattice combination [2] (Hsiao et al., 2008) was applied on two IBM systems, named as UBM and U-BN, with the UBM system rescored with bigram LSA. With bigram LSA rescoring, the word error rate was further reduced by 2.2% relative on Dev07 after lattice combination.

LM Smoothing

Not only fractional Kneser-Ney smoothing is comparable to Witten-Bell smoothing in performance, but the model is also more compact. Table 4.13 shows the compressed size of bigram LSA with different smoothing schemes. Fractional Kneser-Ney smoothing produced a more compact model than Witten-Bell smoothing with over 35% relative

[2]Results were obtained from Ian Lane using the tool implemented by Mark Fuhs.

Table 4.13: Comparison of the size of bigram LSA language model using the Witten-Bell and the fractional Kneser-Ney smoothing on Arabic and Chinese.

Scheme	Arabic LM	Chinese LM
Witten-Bell	3.7Gb	3.4Gb
Kneser-Ney	2.4Gb	2.1Gb

reduction in model size for Arabic and Chinese. The reduction in model size is due to the absolute discounting scheme employed in fractional Kneser-Ney smoothing where the bigram counts smaller than the discounting constant D are pruned.

4.4.7 Discussion

The performance breakdown in terms of broadcast news (BN) and broadcast conversation (BC) genre are shown in Table 4.14 and Table 4.15 for Mandarin and Arabic respectively. It is interesting that bigram LSA generally works better on BN than BC in terms of relative reduction in character error rate although the recognition accuracy is much better on BN than BC. One explanation is that BN is similar to the language model training text that contains a large amount of newspaper text and only a limited amount of audio transcript. On the other hand, BC is more spontaneous in nature with more repetition and hesitation which are rare events on newspaper text.

Figure 4.7 and Figure 4.10 show the recognition error rates per show on the Mandarin and Arabic development sets respectively. Spikes with positive magnitude mean that bigram LSA is effective or vice versa. Both figures show that bigram LSA is effective across the majority of the shows. The relative reduction fluctuates for the "easy" shows with error rates less than 6%. A small fraction of misrecognition can result in a big change in the relative error rates which explains the fluctuation. On the other hand, there are more shows with positive performance with error rates between 6% to 20% (named as "medium" difficulty). As the error rate of a show increases, the relative reduction decreases. In general, the results suggest that bigram LSA is more effective on the shows with "medium" diffi-

Table 4.14: Lattice rescoring results on broadcast news (BN) and broadcast conversation (BC) in character error rate using the CMU-InterACT Mandarin transcription system for the GALE Phase-3 evaluation. * denotes that bigram LSA (biLSA) and trigram LSA (triLSA) are significantly better than LSA at $\leq$5% level of significance.

Mandarin	Dev08				Eval07u			
	BN	Rel. Δ	BC	Rel. Δ	BN	Rel. Δ	BC	Rel. Δ
background	5.4%	-	17.8	-	5.6	-	24.8	-
+LSA	5.0	7.4	17.7	0.6	5.6	0.0	24.7	0.4
+biLSA (Kneser)	5.0	7.4	17.0*	4.5	5.4	3.6	24.3	2.0
background (LSA)	5.5	-ve	17.4	2.2	5.7	-ve	24.3	2.0
+LSA	5.2	3.7	17.1	3.9	5.5	1.8	24.2	2.4
+biLSA (Witten)	5.0	7.4	16.6*	6.7	**5.2***	**7.1**	23.9	3.6
+biLSA (Kneser)	**4.9***	**9.3**	16.7	6.2	**5.2***	**7.1**	**23.8***	**4.0**
+triLSA (Kneser)	5.1	5.6	**16.3***	**8.4**	**5.2***	**7.1**	23.9	3.6
+bi&triLSA (Kneser)	5.0	7.4	16.5*	7.3	**5.2***	**7.1**	23.9	3.6
Cross-adaptation with IBM								
background	4.2	-	13.7	-	4.1	-	19.6	-
+LSA	4.1	2.4	13.5	1.5	3.9	4.9	19.4	1.0
+biLSA	3.9*	7.1	**13.3**	**2.9**	3.6*	12.2	19.3	1.5
background (LSA)	4.2	0.0	13.7	0.0	3.9	4.9	19.4	1.0
+LSA	4.1	2.4	13.6	0.7	3.8	7.3	19.2	2.0
+biLSA	**3.8***	**9.5**	13.5	1.5	**3.5***	**14.6**	**19.1**	**2.6**

Table 4.15: Lattice rescoring results on broadcast news (BN) and broadcast conversation (BC) in word error rate using the CMU-InterACT Arabic transcription system for the GALE Phase-3 evaluation. * denotes that bigram LSA (biLSA) is significantly better than LSA at $\leq$5% level of significance.

Arabic	Dev07				Eval07u			
	BN	Rel. Δ	BC	Rel. Δ	BN	Rel. Δ	BC	Rel. Δ
background	11.6	-	19.4	-	20.8	-	24.7	-
+LSA	11.5	0.9	19.2	1.0	20.6	1.0	24.8	-ve
+biLSA (Witten)	**11.0***	**5.2**	19.0	2.1	**20.4***	**1.9**	**24.6**	0.4
+biLSA (Kneser)	**11.0***	**5.2**	**18.9**	2.6	**20.4***	**1.9**	24.7	0.0
Cross-adaptation with IBM								
background	9.9	-	15.1	-	18.7	-	22.0	-
+LSA	**9.8**	1.0	15.5	-ve	18.6	0.5	21.9	0.5
+biLSA (Kneser)	**9.8**	**1.0**	15.2	-ve	**18.2***	**2.7**	22.1	-ve

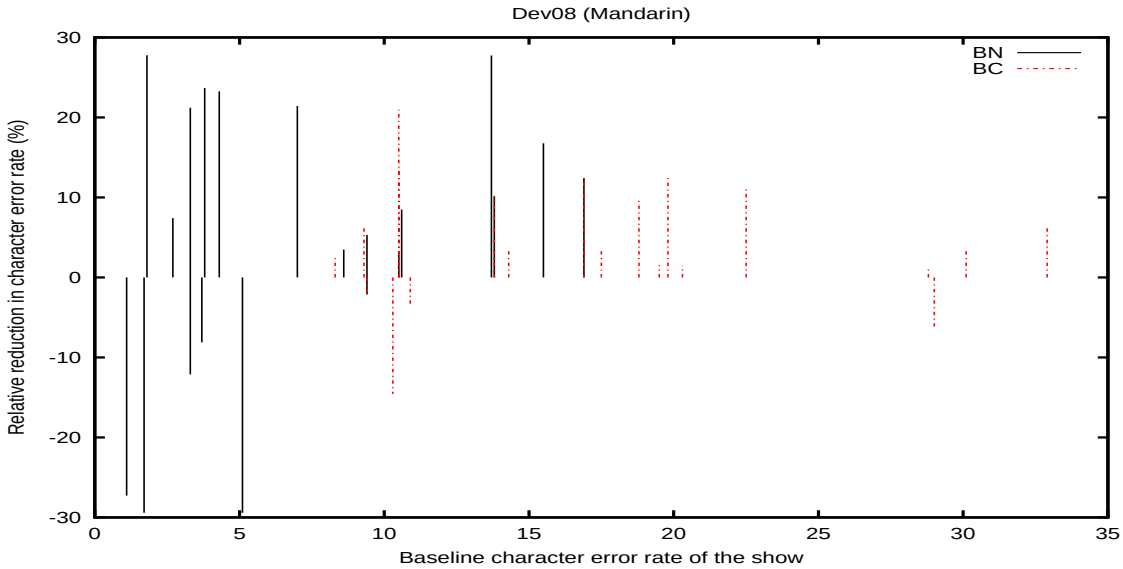

Figure 4.7: Relative reduction in character error rate after bigram-LSA rescoring on the Mandarin Dev08 development set.

Figure 4.8: Relative reduction in character error rate after bigram-LSA rescoring on the Mandarin Eval07u (unsequestered) test set.

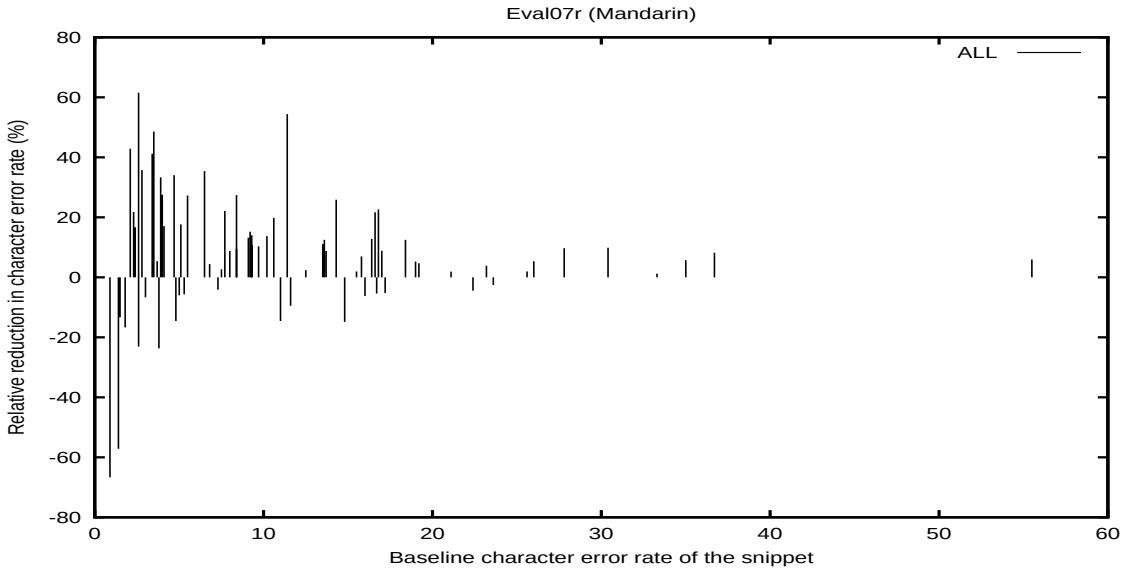

Figure 4.9: Relative reduction in character error rate after bigram-LSA rescoring on the Mandarin Eval07r (retest) test set.

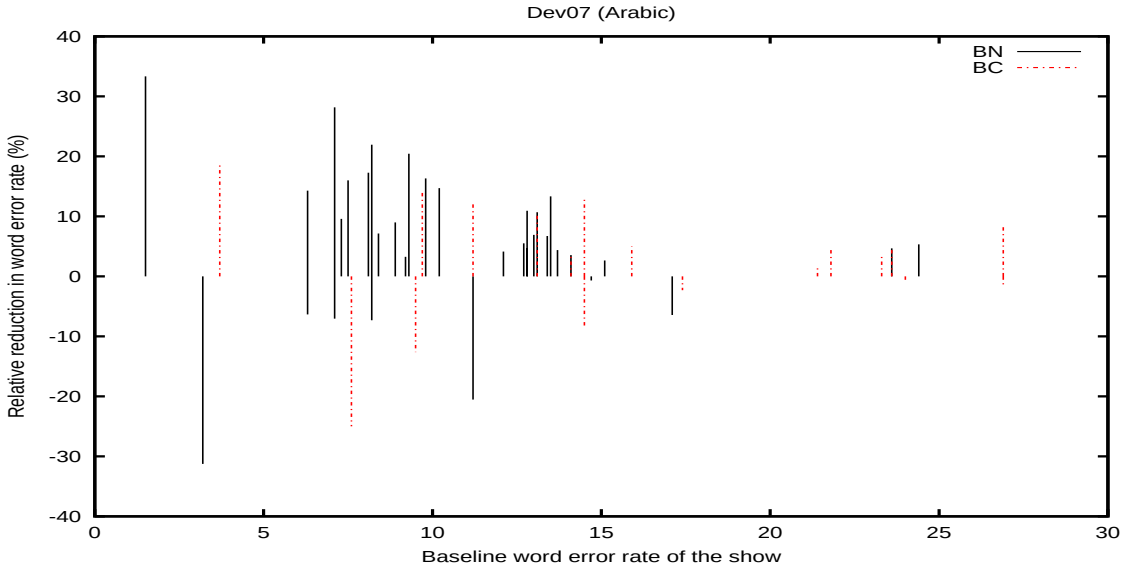

Figure 4.10: Relative reduction in word error rate after bigram-LSA rescoring on the Arabic Dev07 development set.

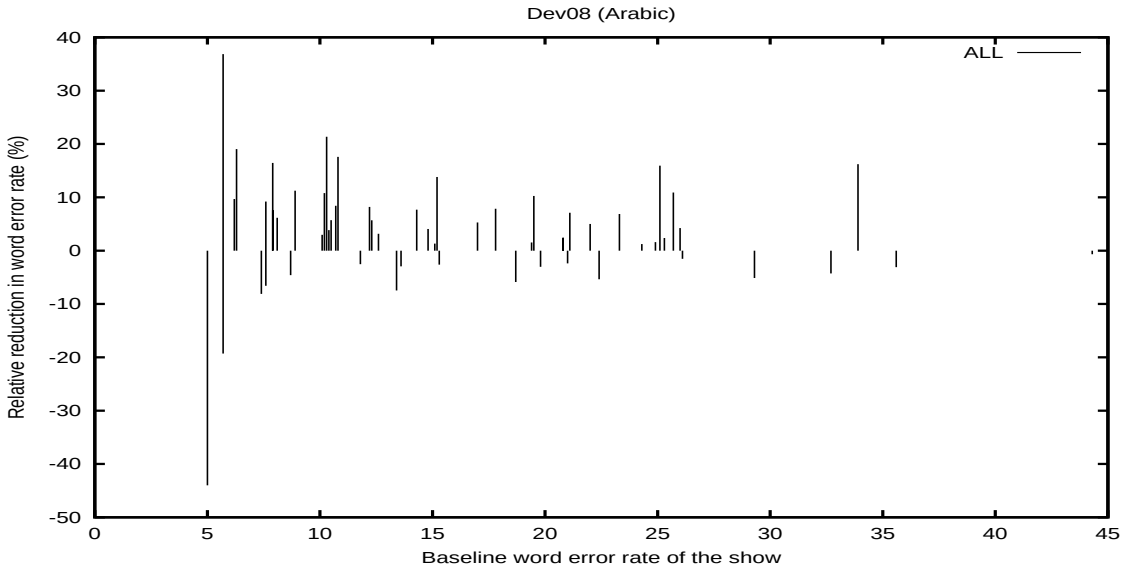

Figure 4.11: Relative reduction in word error rate after bigram-LSA rescoring on the Arabic Dev08 set (unseen).

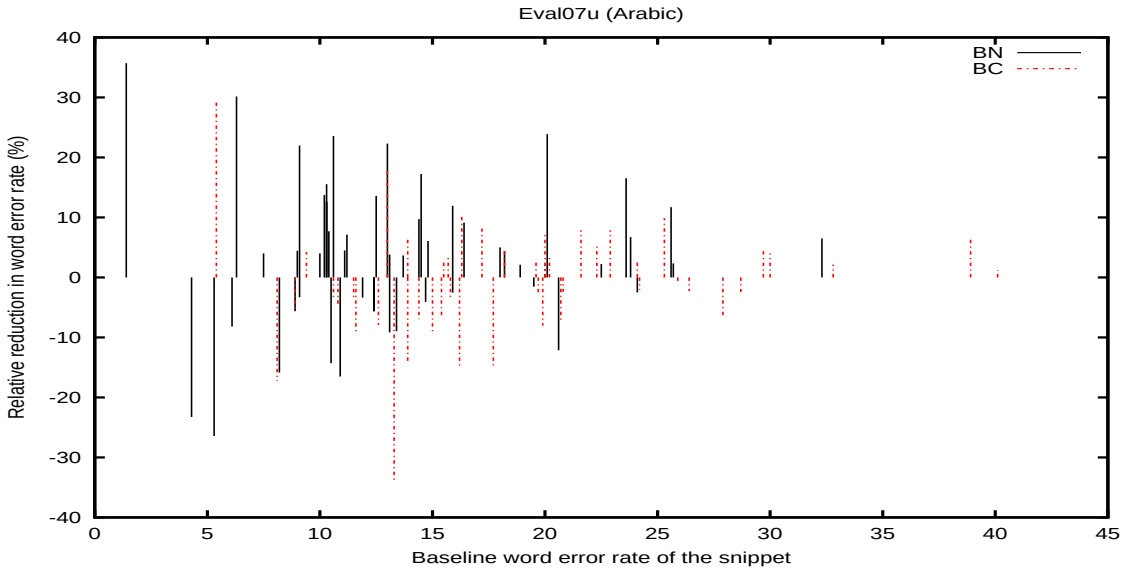

Figure 4.12: Relative reduction in word error rate after bigram-LSA rescoring on the Arabic Eval07u (unsequestered) test set.

culty than the "easy" shows. In addition, the results conform with an observation that most of the recognition errors on the "easy" shows are related to function words but not topical words. Similar trend is observed on the test sets as shown in Figure 4.8 and Figure 4.9 for Mandarin, and Figure 4.11 for Arabic. One exception is the Arabic Eval07u test set showing that adaptation is not effective, especially on broadcast conversation. Broadcast conversation is more disfluent and spontaneous in speaking style compared to broadcast news. In addition, we lack sufficient training data for better modeling. Therefore, part of our future work is to have a better model for broadcast conversation.

4.4.8 Practical Issues

Several points are worth mentioning to make bigram LSA work practically. First of all, the size of bigram LSA can be too big to fit into memory for large-scale evaluation. One solution is to limit the size of vocabulary to a subset occurring only in word lattices. For instance, the base vocabulary of our GALE-P3 Arabic transcription system is $737k$ while the subset on Dev07 is only $11k$. Therefore, it is sufficient to load only the bigram entries

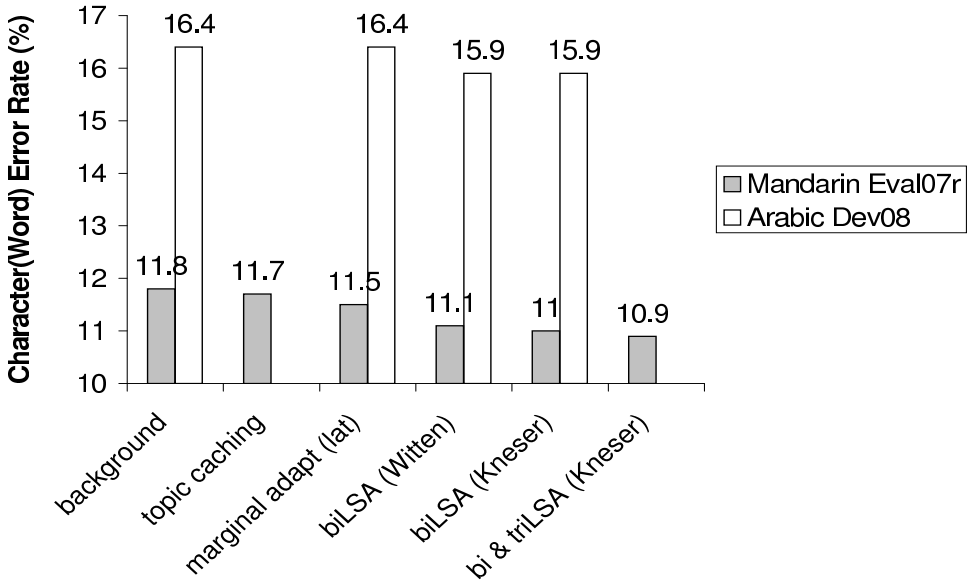

Figure 4.13: Overall performance summary after applying the proposed unsupervised language model adaptation for the large-scale GALE-P3 evaluation on Mandarin and Arabic.

covered by the subset of vocabulary.

Sentence boundaries do not exist in a word lattice. We employ a simple approach to detect sentence boundaries by mapping a silence token $< SIL >$ into $< s >$ when the silence duration exceeds a threshold value, say 0.2 second. This prevents bigram LSA from looking up a bigram which results in a wrong backoff to a unigram model.

For stopwords like auxiliary verbs, articles, conjunctions, sentence boundary markers and punctuations, we do not adapt their N-gram probabilities because predicting stopwords mostly relies on the syntactic context but not the topical context.

The amount of training data from audio transcripts and newspaper text are unbalanced. Therefore, it is desirable to put a higher weight on the audio transcripts than the newspaper text via reweighting the N-gram counts in the M-step of the bigram LSA training.

4.5 Summary

We have investigated a Bayesian latent semantic approach for unsupervised language model adaptation for automatic speech recognition via the N-gram latent Dirichlet-Tree allocation. Topic caching is more robust against speech recognition errors compared to word caching for unsupervised language model adaptation. Incremental marginal adaptation for lattice rescoring is computationally inexpensive and has yielded improvement in recognition performance. Latent Dirichlet-Tree allocation generalizes latent Dirichlet allocation via modeling topic correlation in a tree-based hierarchy, showing rapid training convergence and competitive language model adaptation performance. N-gram latent Dirichlet-Tree allocation has yielded additive gains over latent Dirichlet-Tree allocation via relaxing the "bag-of-word" assumption. Efficient model bootstrapping and smoothing have made this approach applicable for the large-scale evaluation. Figure 4.13 summarizes our contributions towards better recognition performance using our GALE Mandarin and Arabic systems. Empirical results have demonstrated the effectiveness of N-gram latent Dirichlet-Tree allocation for unsupervised language model adaptation, achieving statistically significant reduction in recognition error rates on two different languages.

Chapter 5

Bilingual N-gram LSA Based Adaptation

In Chapter 4, we have shown that monolingual N-gram LSA is effective for unsupervised language model adaptation for automatic speech recognition. In this chapter, we extend this idea to crosslingual adaptation for statistical machine translation.

5.1 Bilingual Latent Semantic Analysis

The success of language model adaptation on automatic speech recognition has motivated applying the same monolingual language model adaptation approach on the target language in statistical machine translation (SMT). Former adaptation approach employs an initial translation of an input text (Kim and Khudanpur, 2003; Paulik et al., 2005a). However, this scheme may depend on the quality of the initial translation. Moreover, it requires two decoding passes.

We present a novel bilingual LSA framework (Tam et al., 2007b) to perform language model adaptation (Tam et al., 2007a) across languages, enabling adaptation from one language based on an adaptation text of another language in a single decoding pass. Bilingual LSA consists of two models based on latent Dirichlet-Tree allocation: one for each lan-

guage trained on parallel document corpora. The key feature of bilingual LSA is a one-to-one topic correspondence between a source and target LSA model. For instance, say topic 10 of a source LSA model is about politics. Then topic 10 of a target LSA model also corresponds to politics and so forth. During language model adaptation, we first infer topic mixture weights of a source text using the source LSA model. We then transfer the inferred mixture weights into LSA (and N-gram LSA) on the target language for language model adaptation. 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. For the same reason, propagation of translation errors can be avoided by using the source text for adaptation. The challenge in bilingual LSA is to enforce a one-to-one topic correspondence. Our proposal is to assume that the topic distribution among a parallel document pair is identical. The assumption is reasonable for a parallel document pair that are faithful translations. In the variational Expectation-Maximization algorithm, this can be easily achieved via sharing the variational topic posteriors between a parallel document pair so that a common latent topic space is enforced 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 a translation lexicon via marginal adaptation (Tam and Schultz, 2007a) so that the likelihood of a bilingual phrase is sensitive to the topics of an input source text. Thus, a background phrase table is enhanced with additional phrase scores computed using the adapted translation lexicon. The weights for the additional phrase and language model feature functions are then optimized via the minimum error rate training. Figure 5.1 illustrates the idea of topic transfer from a source to target LSA followed by language model adaptation, translation lexicon adaptation and phrase table adaptation.

5.1.1 Bilingual LSA Training

Bilingual LSA training is based on sharing the document-level topic posterior distribution between a parallel document pair. It consists of two stages: In the first stage, we perform

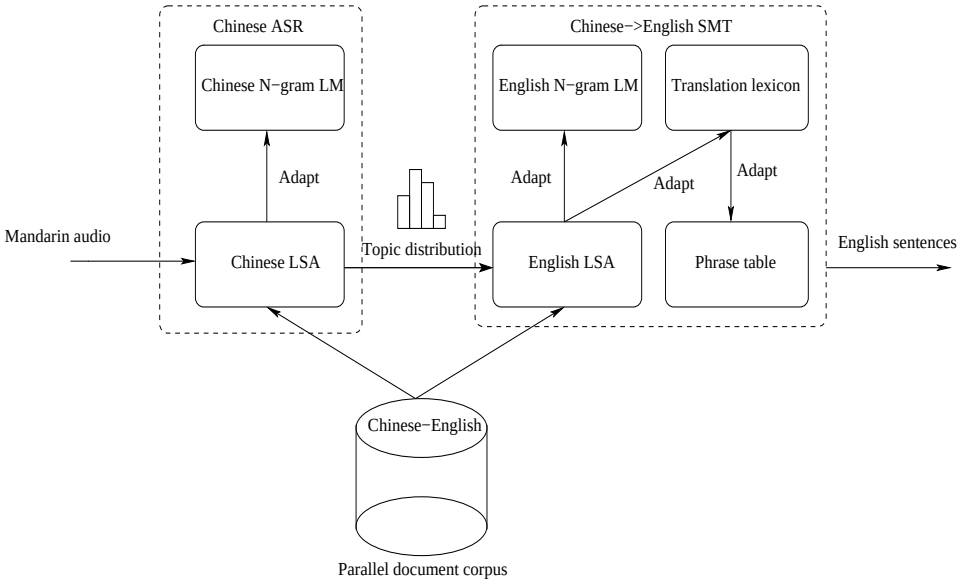

Figure 5.1: Bilingual LSA-based adaptation via transfer of topic distribution from a source language to a target language for speech translation.

monolingual LSA training using the variational Expectation-Maximization algorithm (see equations 4.20–4.25 for latent Dirichlet-Tree allocation) on source documents in parallel corpora. We use the source LSA to compute the term $e^{E_q[\log \theta_k]}$ in equation 4.22 for each source document. In the second stage, we apply the same term $e^{E_q[\log \theta_k]}$ to *bootstrap* the target LSA, which is the key to enforce a one-to-one topic correspondence. The hyper-parameters 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 equation 4.22 with fixed $e^{E_q[\log \theta_k]}$ in the E-step, and equation 4.25 in the M-step to estimate $\{p(w|k)\}$ for the target LSA model. Figure 5.2 illustrates the idea of enforcing a one-to-one topic correspondence of parallel document pairs during bootstrapping a target LSA model from a source LSA model denoted as $bLSA_{(src,tgt)}$. Since the topic posteriors are pre-computed, the E-step is non-iterative resulting in rapid LSA training. In short, given a monolingual LSA, we can rapidly bootstrap LSA models of new languages using parallel corpora. Since the topic transfer can be bi-directional, we can perform the bilingual LSA training in a reverse manner, that is, training a target LSA model followed by bootstrapping a source

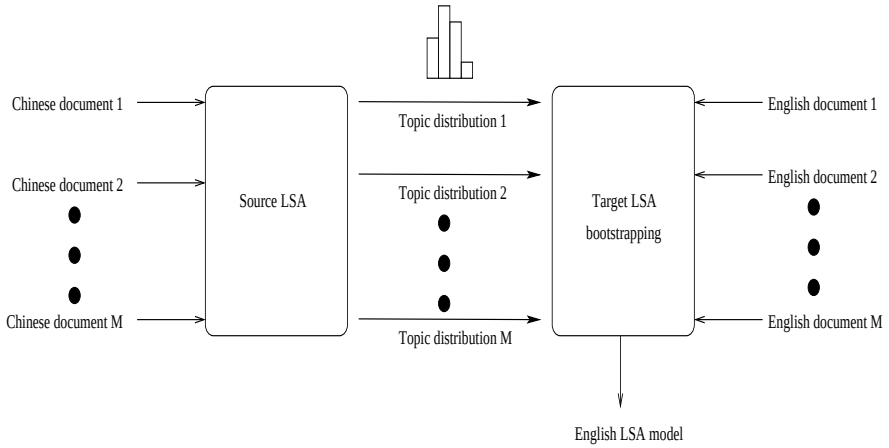

Figure 5.2: LSA bootstrapping via sharing of variational topic posteriors for parallel documents.

Table 5.1: Size of the parallel training corpora for bilingual LSA training.

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

LSA model denoted as $bLSA_{(tgt,src)}$. Bilingual LSA training procedure is general and can be applied to different probabilistic models such as latent Dirichlet allocation and probabilistic LSA. In our experiments, we employ latent Dirichlet-Tree allocation due to its stable training convergence and competitive test performance for automatic speech recognition in Chapter 4.

5.1.2 Experiment

Bilingual LSA was trained using the Chinese–English parallel document corpora consisting of the FBIS corpus, Xinhua News, Hong Kong News, Donga News [1] and Sinorama articles. The combined corpora contained $96k$ parallel documents with $41M$ Chinese words

[1] http://{china,english}.donga.com

and $50M$ English words as shown in Table 5.1.

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 for bilingual LSA was set to 200 based on our best knowledge of language model adaptation for automatic speech recognition. A balanced binary Dirichlet-tree prior was used. The source and target vocabulary 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. For N-gram LSA training, we used the English side of the bilingual LSA to bootstrap the bigram LSA and the trigram LSA as described in Section 4.4.1. Since this is a monolingual N-gram LSA training on the English side, the background language model training data were included.

5.1.3 Results

The proposed bilingual LSA training approach enforced a one-to-one topic correspondence successfully and extracted parallel topics as shown in Table 5.2. 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 crosslingual word triggers via topics. Figure 5.3 demonstrates that our proposed approach leads to rapid training convergence due to sharing of the variational Dirichlet posteriors with the Chinese LSA model compared to the monolingual English LSA starting with the same flat model. On the other hand, the monolingual LSA training had a better training likelihood when more training iterations were applied, which is reasonable since the bootstrapping approach constrain the parameter space so that a one-to-one topic correspondence is satisfied while the parameter space of monolingual LSA training is unconstrained.

Table 5.2: Parallel topics extracted by $bLSA_{(CH,EN)}$. Top words on the Chinese side are translated into English for illustration purposes.

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

5.2 Crosslingual Language Model Adaptation

Marginal language model adaptation in crosslingual settings can be performed in almost the same manner as in monolingual settings as described in Section 2.5.3 except that a source text is used for adaptation in the crosslingual case. Firstly, we estimate the topic weights of latent Dirichlet-Tree allocation on the source language using equation 5.1.

$$\hat{\theta}_k^{(CH)} \propto \prod_{jc} \left(\frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \tag{5.1}$$

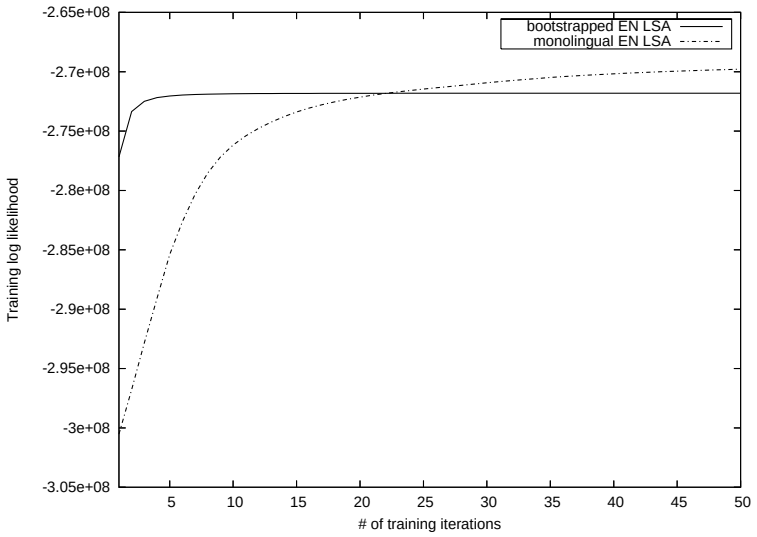

Figure 5.3: Training log likelihood of bootstrapped English LSA from Chinese LSA compared to flat monolingual English LSA.

Then we apply the source topic weights into the target LSA and the N-gram LSA to obtain in-domain marginals as in equation 5.2.

$$\begin{aligned} p_{EN}(w) &= \sum_{k=1}^{K} \hat{\theta}_k^{(CH)} \cdot p^{(EN)}(w|k) \\ p_{EN}(w|v) &= \sum_{k=1}^{K} \hat{\theta}_k^{(CH)} \cdot p^{(EN)}(w|v,k) \\ p_{EN}(w|u,v) &= \sum_{k=1}^{K} \hat{\theta}_k^{(CH)} \cdot p^{(EN)}(w|u,v,k) \end{aligned} \tag{5.2}$$

Finally, we apply marginal adaptation to incorporate LSA into a background language model as described in Section 2.5.3. The adapted bigram or trigram LSA are added as additional language model feature functions to compute the posterior probability of a target sentence given a source sentence in statistical machine translation (See Section 3.4.4 for background information).

Table 5.3: Target word perplexity on MT06 using 67M-word (260M-word) parallel corpora for phrase extraction and 500M-word (2.7B-word) English corpora for language model training. Vocabulary size of the target language model is 1.3M (4.1M).

Language model	Perplexity	Rel. Δ
Baseline EN 4-gram (500M)	154	-
bilingual LSA-adapted	127	17.5%
mono LSA-adapted	125	18.8%
Baseline EN 5-gram (2.7B)	147	-
bilingual LSA-adapted	131	10.9%

5.2.1 Experiment

The marginal adaptation approach described in Section 2.5.3 was applied to an English background language model for each source test document. Words on the stopword list[2] plus punctuation were filtered out from language model adaptation since the usage of stopwords usually does not depend on topical context.

5.2.2 Results

Table 5.3 shows that the proposed approach effectively reduced 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 helped even on a huge 5-gram language model trained on a large amount of text. We also performed monolingual language model adaptation using the translated hypotheses from the decoder. This gave slightly better performance than bilingual LSA adaptation but with a two-pass decoding scheme.

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

5.3 Translation Lexicon Adaptation

Not only the bilingual LSA approach applies for language model adaptation, but it also applies for translation lexicon adaptation. Translation lexicon adaptation is motivated from an observation that a source word can be translated into different target words depending on a topical context. One popular example is the word "bank" which can be related to either a "financial bank" or a "river bank". The adapted translation lexicon can be used to score the phrase pairs depending on the topical context of an input document. 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 source marginals $p(c|d_{ch})$ that are estimated *a priori* using the source document d_{ch}. Thus the objective function to minimize is as in equation 5.3,

$$\begin{array}{lrcl} \text{Minimize} & \sum_e p_a(e) \cdot KL\left(p_a(.|e)||p_{bg}(.|e)\right) & & \\ \text{such that} & \forall c : \sum_e p_a(e) \cdot p_a(c|e) & = & p(c|d_{ch}) \\ & \forall e : \sum_c p_a(c|e) & = & 1 \end{array} \tag{5.3}$$

We write the Lagrangian of the objective function, take the derivative with respect to $p_a(c|e)$ and set it to zero (equation 5.4–5.5):

$$\begin{array}{rcl} D(p_a(.|.)) & = & \sum_e p_a(e) \cdot \sum_c p_a(c|e) \cdot \log \frac{p_a(c|e)}{p_{bg}(c|e)} \\ & & - \sum_c \lambda_c \left(\sum_e p_a(e) \cdot p_a(c|e) - p(c|d_{ch})\right) \\ & & - \sum_e \mu_e \left(\sum_c p_a(c|e) - 1\right) \end{array} \tag{5.4}$$

$$\begin{array}{rcl} \frac{\partial D(.)}{\partial p_a(c|e)} & = & p_a(e) \cdot (1 + \log \frac{p_a(c|e)}{p_{bg}(c|e)}) - \lambda_c \cdot p_a(e) - \mu_e = 0 \\ \Rightarrow p_a(c|e) & \propto & p_{bg}(c|e) \cdot e^{\lambda_c} \propto p_{bg}(c|e) \cdot e^{\sum_j \lambda_j \cdot f_j(c,e)} \end{array} \tag{5.5}$$

where

$$f_j(c,e) = \begin{cases} 1 & \text{if } c = j \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

$f_j(c,e)$ is a unigram feature function independent of e. Since the solution of the adapted lexicon is in an 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 equation 5.7–5.11,

$$\begin{aligned}
\forall j: \quad \lambda_j^{(t+1)} &= \lambda_j^{(t)} + \log \frac{\tilde{E}[f_j(c,e)]}{E[f_j(c,e)]} && (5.7)\\
&= \lambda_j^{(t)} + \log \frac{\sum_{c,e} \tilde{p}(c,e|d_{ch}) \cdot f_j(c,e)}{\sum_{c,e} p_a^{(t)}(c|e) \cdot p_a(e) \cdot f_j(c,e)} && (5.8)\\
&= \lambda_j^{(t)} + \log \frac{\sum_{e} \tilde{p}(c=j,e|d_{ch})}{\sum_{e} p_a^{(t)}(c=j|e) \cdot p_a(e)} && (5.9)\\
&= \lambda_j^{(t)} + \log \frac{p(c=j|d_{ch})}{\sum_{e} p_a^{(t)}(c=j|e) \cdot p_a(e)} && (5.10)\\
&\approx \lambda_j^{(t)} + \log \frac{p(c=j|d_{ch})}{\sum_{e} p_a^{(t)}(c=j|e) \cdot \mathbf{p_{blsa}}(\mathbf{e}|\mathbf{d_{ch}})} && (5.11)
\end{aligned}$$

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 (5.11) 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 a Chinese and English sentence respectively.

5.3.1 Phrase Table Adaptation

Ideally, an adapted translation lexicon can be applied directly during phrase extraction. But this involves an extra implementation work into a phrase extraction algorithm. An alternative approach is to take a 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 the background phrase table similar to the IBM Model1 (equation 5.12),

$$p_a(c_1^I|e_1^J) = \prod_{i=1}^{I} \frac{1}{J_i} \cdot \sum_{j}^{J_i} p_a(c_i|e_j) \tag{5.12}$$

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 entry 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 making the phrase score more discriminative. The NULL model $p(c|NULL)$ or the minimum of $p(c|NULL)$ is used as a backoff model to avoid a zero probability for an 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 and SMT decoding.

5.3.2 Results

Marginal adaptation resulted in a sharper translation lexicon in which the uncertainty of a word-to-word translation was reduced. With the word context "*according to a report by south korean ytn cable tv*", 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 a source text.

5.4 Text Translation Results

The upper section of Table 5.4 shows the translation performance in BLEU and NIST on MT06 using the medium-scale SMT system. 2% relative improvement in BLEU was

Table 5.4: MT06 evaluation results on BLEU and NIST using 67M-word (260M-word) parallel corpora for phrase extraction and 500M-word (2.7B-word) English corpora for language model training. Vocabulary size of the target language model is 1.3M (4.1M). Four English references are used for scoring. English bigram LSA (biLSA) and trigram LSA (triLSA) are applied. * denotes that bilingual LSA adaptation is significantly better than the unadapted baseline at 95% confidence interval.

Language model	BLEU (%)	Rel. Δ	NIST	Rel. Δ
Baseline EN 4-gram (500M)	28.06	-	8.71	-
bilingual LSA-adapted	28.62	2.0	8.80	1.0
bilingual LSA-adapted lexicon	28.59	1.9	8.92*	2.4
bilingual LSA-adapted + lexicon	28.91*	3.0	8.97*	3.0
mono LSA-adapted	28.41	1.2	8.81	1.1
mono LSA-adapted lexicon	28.72	2.4	8.96*	2.9
mono LSA-adapted + lexicon	28.97*	3.2	9.00*	3.3
bilingual LSA-adapted + biLSA	29.08*	3.6	8.99*	3.2
bilingual LSA-adapted + triLSA	**29.42***	**4.8**	9.05*	3.9
bilingual LSA-adapted + bi & triLSA	**29.42***	**4.8**	**9.08***	**4.2**
bilingual LSA-adapted + lexicon + triLSA	29.38*	4.7	9.04*	3.8
Baseline EN 5-gram (2.7B)	31.49	-	9.23	-
bilingual LSA-adapted	31.94	1.4	9.31	0.9
bilingual LSA-adapted lexicon	32.03	1.7	9.34	1.2
bilingual LSA-adapted + lexicon	32.09	1.9	9.37	1.5
bilingual LSA-adapted + triLSA	32.13	2.0	9.37	1.5
bilingual LSA-adapted + lexicon + triLSA	**32.28**	**2.5**	**9.38***	**1.6**

achieved compared to the unadapted baseline after applying bilingual LSA-based language model adaptation and translation lexicon adaptation separately. When both techniques were applied simultaneously, the gain was additive giving 3% relative improvement in BLEU compared to the unadapted baseline. The improvement was statistically significant at a 95% confidence interval [27.29%,28.84%] with respect to the unadapted baseline. The same performance trend in NIST was also observed with 3% relative improvement compared to the unadapted baseline. The improvement was statistically significant at a 95% confidence interval [8.61,8.85] with respect to the unadapted baseline.

The middle section of Table 5.4 shows that monolingual LSA adaptation using the first-pass translated hypotheses achieved a similar performance compared to bilingual LSA adaptation using a source text. In other words, the source text and the initial MT hypotheses were 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 5.5 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*.

We applied bigram and trigram LSA as the additional language model feature functions. Table 5.4 shows that bigram LSA yielded additive relative improvement by 1.6% and 2.2% in BLEU and NIST respectively compared to bilingual LSA. Replacing bigram LSA with trigram LSA achieved further relative improvement by 1.2% and 0.7% in BLEU and NIST respectively compared to bigram LSA. Adding bigram LSA and trigram LSA together yielded slight gain in NIST but equal performance in BLEU. This implies that trigram LSA already covers most of the information from bigram LSA. Incorporating only trigram LSA is sufficient for good performance and avoids the bigram LSA training.

The lower section of Table 5.4 shows the translation performance using the GALE-P2.5 SMT system. The performance trend was similar to the medium-scale system. Improvement in BLEU and NIST were observed after applying bilingual LSA-based language model adaptation and translation lexicon adaptation. Additive gain was obtained

Table 5.5: 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”

after applying both techniques together, yielding 1.9% relative improvement in BLEU compared to the unadapted baseline. Adding trigram LSA further improved the BLEU score with 2.5% relative improvement compared to the unadapted baseline. The gain in BLEU was reduced compared to the results on the medium-scale setting, and it was not statistically significant marginally at a 95% confidence [30.70,32.34] with respect to the unadapted baseline. This may be explained by having a stronger baseline 5-gram language model with an increased amount of training text and a better word reordering strategy in the GALE system. The overall improvement in NIST followed a similar trend with 1.6% relative improvement compared to the unadapted baseline. The gain was statistically significant at a 95% confidence interval [9.147,9.378] with respect to the unadapted baseline.

5.4.1 Human Evaluation

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 non-overlapping 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.

Table 5.6 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 achieved a better average score than the unadapted baseline although the gain was not statistically significant. Table 5.7 shows an example in which bilingual LSA gives a better fluency than the unadapted baseline.

It is surprising that slight degradation of adequacy was observed in the human eval-

Table 5.6: Human evaluation results on sentence fluency and adequacy on MT06 using the GALE Phase-2.5 SMT system compared with the bilingual LSA (bLSA). Worst score is 1 and the best score is 5.

	Fluency		Adequacy		Average	
Grader ID	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

Table 5.7: 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.

Table 5.8: MT06 evaluation results on the average recall using the GALE-P2.5 SMT system.

Language model	Recall (%)	Rel. Δ
Baseline EN 5-gram (2.7B)	46.99	-
bilingual LSA-adapted	47.45	1.0
bilingual LSA-adapted lexicon	47.44	1.0
bilingual LSA-adapted + lexicon	47.74	1.6
bilingual LSA-adapted + triLSA	**48.02**	**2.2**
bilingual LSA-adapted + lexicon + triLSA	47.93	2.0

uation results. Perhaps the number of graders is not large enough to represent the actual performance. Therefore, we employ recall to measure meaning preservation as follows:

$$\text{Recall} = \frac{\text{\# of matched unigram}}{\text{\# of unigram in a reference}} \tag{5.13}$$

Before calculating the recall, stopwords and punctuations were removed before the calculation. Since MT06 has four English references, recall of each reference was computed and the average value is shown in Table 5.8 using the translated hypotheses from the GALE-P2.5 system. Our approaches achieved better recall compared to the unadapted baseline, suggesting that bilingual LSA adaptation may preserve the meaning of source text better.

5.4.2 Discussion

Table 5.9 and Table 5.10 shows the performance breakdown in newsgroup (NG), newswire (NW) and broadcast news (BN) in BLEU and NIST respectively. Adaptation consistently helped on the newswire documents using the medium-scale and the GALE-P2.5 SMT system. This observation is reasonable since the training data are mostly from newswire. On the other hand, adaptation performance on broadcast news and newsgroup are inconsistent. Figure 5.4 and Figure 5.5 shows the corresponding relative improvement in BLEU per document using the medium-scale and the GALE-P2.5 SMT system respectively. The trends

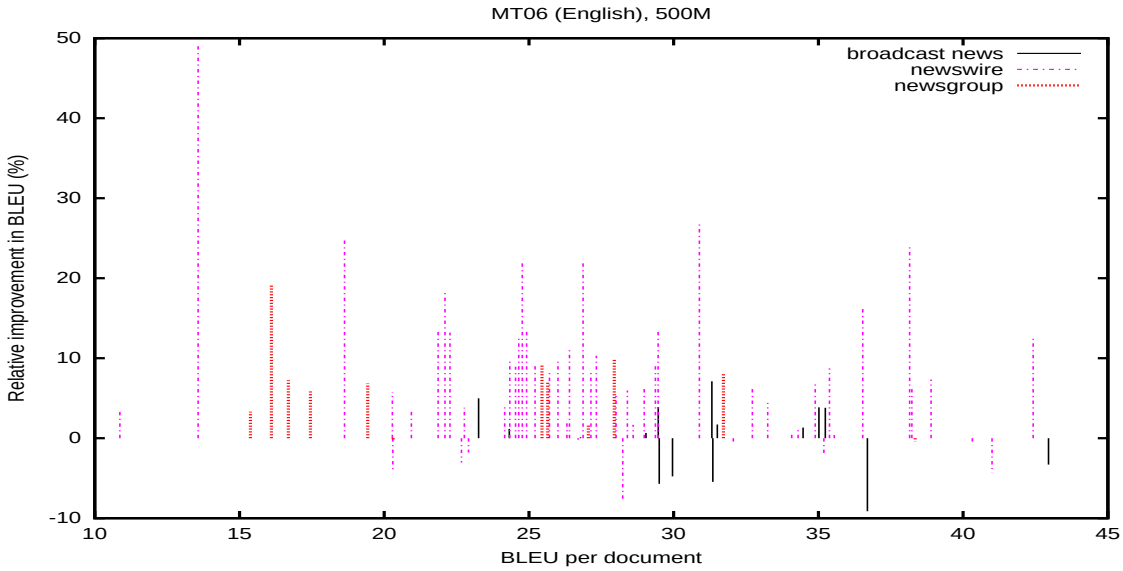

Figure 5.4: Relative BLEU improvement of LSA adaptation compared to the unadapted baseline per document on MT06 using the medium-scale SMT system (500M).

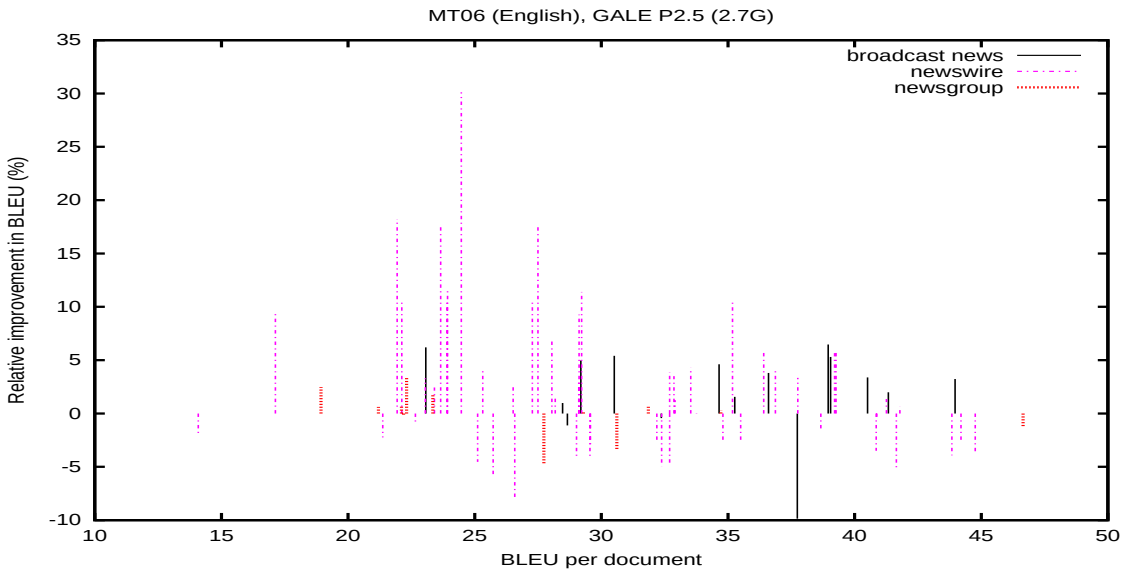

Figure 5.5: Relative BLEU improvement of LSA adaptation compared to the unadapted baseline per document on MT06 using the GALE-P2.5 SMT system (2.7B).

Table 5.9: MT06 evaluation results on newsgroup (NG), newswire (NW) and broadcast news (BN) genre measured on BLEU using 67M-word (260M-word) parallel corpora for phrase extraction and 500M-word (2.7B-word) English corpora for language model training. Vocabulary size of the target language model is 1.3M (4.1M). Four English references are used for scoring.

	BLEU (%)					
Language model	NG	Rel. Δ	NW	Rel. Δ	BN	Rel. Δ
Baseline EN 4-gram (500M)	23.62	-	28.38	-	30.98	-
bilingual LSA-adapted	23.93	1.3	29.19	2.9	31.29	1.0
bilingual LSA-adapted lexicon	24.22	2.5	28.81	1.5	31.60	2.0
bilingual LSA-adapted + lexicon	24.76	4.8	29.38	3.5	31.33	1.1
mono LSA-adapted	23.78	0.7	28.95	2.0	31.10	0.4
mono LSA-adapted lexicon	24.30	2.9	29.25	3.1	31.24	0.8
mono LSA-adapted + lexicon	24.79	5.0	29.51	4.0	31.30	1.0
bilingual LSA-adapted + biLSA	24.73	4.7	29.72	4.7	31.39	1.3
bilingual LSA-adapted + triLSA	25.10	6.3	30.22	6.5	31.42	1.4
bilingual LSA-adapted + bi & triLSA	**25.22**	**6.8**	**30.35**	**6.9**	31.22	0.8
bilingual LSA-adapted + lexicon + triLSA	25.14	6.4	30.09	6.0	**31.50**	**1.7**
Baseline EN 5-gram (2.7B)	27.71	-	31.34	-	34.72	-
bilingual LSA-adapted	27.97	0.9	31.95	1.9	35.02	0.9
bilingual LSA-adapted lexicon	27.73	0.1	31.85	1.6	35.68	2.8
bilingual LSA-adapted + lexicon	27.70	-ve	32.08	2.4	35.52	2.3
bilingual LSA-adapted + triLSA	**28.01**	**1.1**	32.10	2.4	35.41	2.0
bilingual LSA-adapted + triLSA + lexicon	27.75	0.1	**32.26**	**2.9**	**35.84**	**3.2**

Table 5.10: MT06 evaluation results on newsgroup (NG), newswire (NW) and broadcast news (BN) genre measured on NIST using 67M-word (260M-word) parallel corpora for phrase extraction and 500M-word (2.7B-word) English corpora for language model training. Vocabulary size of the target language model is 1.3M (4.1M). Four English references are used for scoring.

	NIST					
Language model	NG	Rel. Δ	NW	Rel. Δ	BN	Rel. Δ
Baseline EN 4-gram (500M)	7.15	-	8.51	-	8.32	-
bilingual LSA-adapted	7.25	1.4	8.61	1.2	8.39	0.8
bilingual LSA-adapted lexicon	7.39	3.4	8.71	2.4	8.46	1.7
bilingual LSA-adapted + lexicon	7.45	4.2	8.77	3.1	8.46	1.7
mono LSA-adapted	7.25	1.4	8.63	1.4	8.38	0.7
mono LSA-adapted lexicon	7.44	4.1	8.78	3.2	8.43	1.3
mono LSA-adapted + lexicon	7.50	4.9	8.83	3.8	8.45	1.6
bilingual LSA-adapted + biLSA	7.48	4.6	8.78	3.2	**8.49**	**2.0**
bilingual LSA-adapted + triLSA	7.55	5.6	8.88	4.3	8.44	1.4
bilingual LSA-adapted + bi & triLSA	**7.61**	**6.4**	**8.94**	**5.1**	8.46	1.7
bilingual LSA-adapted + lexicon + triLSA	7.58	6.0	8.88	4.3	8.44	1.4
Baseline EN 5-gram (2.7B)	7.79	-	8.96	-	8.73	-
bilingual LSA-adapted	7.83	0.5	9.08	1.3	8.75	0.2
bilingual LSA-adapted lexicon	7.91	1.5	9.10	1.6	8.78	0.6
bilingual LSA-adapted + lexicon	7.90	1.4	**9.15**	**2.1**	8.78	0.6
bilingual LSA-adapted + triLSA	7.87	1.0	9.12	1.8	**8.84**	**1.3**
bilingual LSA-adapted + triLSA + lexicon	**7.92**	**1.7**	9.13	1.9	8.83	1.1

Table 5.11: Translation results after crosslingual language model adaptation on the unsequestered broadcast news portion of the Mandarin Eval07 test set (Eval07u.BN) using the GALE-P2.5 SMT system with different number of latent topics K in bilingual LSA.

Source input	BLEU (%)	Rel. Δ	NIST	Rel. Δ
Reference (OOV=0.075%)	17.37	-	5.53	-
K=100	17.69	1.8	5.58	0.9
K=200	**17.74**	**2.1**	**5.59**	**1.1**
K=300	17.51	0.8	5.56	0.5

on both graphs are different: Adaptation helps on documents with high BLEU scores (say $\geq$35%) using the medium-scale system while it is the opposite using the GALE-P2.5 system. On the other hand, adaptation generally helped on documents with BLEU scores between 20%–30% on both systems.

5.5 End-to-End Translation

For end-to-end speech translation, we evaluated the effectiveness of topic adaptation using different source inputs on our GALE-P2.5 Mandarin-English SMT system without part-of-speech reordering feature. To investigate the effect of topic adaptation on transcription towards downstream translation, we translated word hypotheses from unadapted and LSA-adapted GALE-P3 Mandarin transcription systems with character error rates 5.6% and 5.2% respectively on the unsequestered broadcast news portion of Eval07 test set. We also translated manual transcription to serve as an upper-bound performance for comparison.

5.5.1 Optimal Number of Topics

Table 5.11 shows the performance of crosslingual language model adaptation with different number of latent topics in bilingual LSA. With the number of topics set to 200, bilingual LSA yielded the best translation performance in terms of BLEU and NIST. Same

result was found in monolingual language model adaptation for automatic speech recognition in Section 4.1.3.

5.5.2 Results

Table 5.12 shows the development of Mandarin-English speech translation using the GALE transcription and translation systems. Translation on the N-gram LSA-adapted word hypotheses with lower character error rate translated to better translation performance successfully, yielding 0.9% relative improvement in both BLEU and NIST compared to the background unadapted word hypotheses. Using manual transcription as inputs gave the best translation performance with 5.9% and 4.7% relative improvement in BLEU and NIST respectively compared to the background unadapted word hypotheses.

Although the Chinese side of the parallel corpora and the inputs were segmented using the Stanford segmenter consistently, the out-of-vocabulary (OOV) rate on the manual transcription was 1.7%, which was moderately high due to the ability of the Stanford segmenter to hypothesize new vocabulary during segmentation. Unfortunately, the out-of-vocabulary Chinese words cannot be translated because of the possible segmentation mismatch with a phrase table. Therefore, we attempted to reduce the mismatch via segmentation refinement. First, we extracted the Chinese words from the phrase table to form a word list. Then, a maximal matching segmenter with this word list was applied on the Chinese test inputs that were segmented with a Stanford segmenter. As a result, the OOV terms were further segmented into smaller terms which may conform better to the segmentation of the phrase table. As shown in the second sub-table of Table 5.12, the OOV rate dropped significantly to less than 0.1%. In addition, segmentation refinement translated to better translation performance, yielding 3% relative improvement in BLEU on the manual transcription compared to the corresponding counterpart with Stanford segmentation only. Similar results were obtained on the automatic transcription with 2.8% relative improvement in BLEU compared to the corresponding counterparts. In other words, the benefit of using N-gram LSA-adapted word hypotheses was maintained after segmentation refinement.

Table 5.12: Speech translation results on the unsequestered broadcast news portion of the Mandarin Eval07 test set (Eval07u.BN) on BLEU and NIST using the GALE-P2.5 SMT system. Source inputs are word hypotheses from an unadapted background GALE-P3 Mandarin transcription system (CER=5.6%), an N-gram LSA-adapted (CER=5.2%), and a manual reference (CER=0%). Confusion-network-like English references are used for scoring. Relative improvement in BLEU and NIST are reported with respect to the unadapted background word hypotheses before segmentation refinement.

Source input	BLEU (%)	Rel. Δ	NIST	Rel. Δ
ASR hypo (background)	15.92	-	5.28	-
ASR hypo (N-gram LSA)	16.07	0.9	5.33	0.9
Reference (OOV=1.7%)	16.86	5.9	5.53	4.7
After Segmentation Refinement				
ASR hypo (background)	16.36	2.8	5.28	0.0
ASR hypo (N-gram LSA)	16.52	3.8	5.34	1.1
Reference (OOV=0.075%)	17.37	9.1	5.53	4.7
After LM Adaptation				
ASR hypo (background)	16.67	4.7	5.34	1.1
ASR hypo (N-gram LSA)	16.80	5.5	5.37	1.7
Reference (OOV=0.075%)	17.74	11.4	5.59	5.9
After Lexical Adaptation				
ASR hypo (background)	16.52	3.8	5.32	0.8
ASR hypo (N-gram LSA)	16.90	6.2	5.40	2.3
Reference (OOV=0.075%)	17.72	11.3	5.58	5.7
After LM + Lexical Adaptation				
ASR hypo (background)	16.88	6.0	5.37	1.7
ASR hypo (N-gram LSA)	17.18	7.9	5.43	2.8
Reference (OOV=0.075%)	17.99	13.0	5.63	6.6

After segmentation refinement, we applied crosslingual adaptation on target language models, translation lexicons and phrase tables incrementally as shown in Table 5.12. Similar to the results on text translation, language model adaptation and translation model adaptation improved the translation performance individually. When both adapted models were applied simultaneously, the gain was additive, yielding 3.2–3.6% relative improvement in BLEU and 1.7–1.8% relative improvement in NIST compared to the corresponding counterparts of the unadapted SMT systems. However, the improvement was not statistically significant at a 95% confidence interval [15.12%,17.65%] in BLEU and [5.100,5.459] in NIST compared to the unadapted speech translation system. Overall, it is beneficial to apply topic adaptation to improve recognition accuracy that translates to better translation performance. In addition, the gain from better recognition accuracy and sharper translation models after topic adaptation were additive, yielding 1.8% and 1.1% relative improvement in BLEU and NIST respectively compared to the background unadapted word hypotheses. Comparing with the translation results on manual transcription, there is still much room for further improvement in speech translation. In other words, we expect that improving the upstream recognition accuracy will improve the downstream translation performance.

5.6 Non-Parallel Bilingual Latent Semantic Analysis

We have shown the effectiveness of bilingual latent semantic analysis for crosslingual adaptation for statistical machine translation. The limitation of bilingual latent semantic analysis is the requirement of parallel corpora for model training. Since parallel corpora are more expensive to collect than monolingual non-parallel corpora, incorporating non-parallel corpora into bilingual latent semantic analysis is attractive. Moreover, non-parallel corpora generally cover a broader range of topics and vocabulary than parallel corpora. The main issue is that blind incorporation of non-parallel corpora may destroy a one-to-one topic correspondence in bilingual latent semantic analysis since the alignment between a source and target monolingual document is unknown or even non-existent.

To work around the issue of the unknown document alignment, we employ a semi-supervised learning approach where some parallel seed documents are given. The smooth-

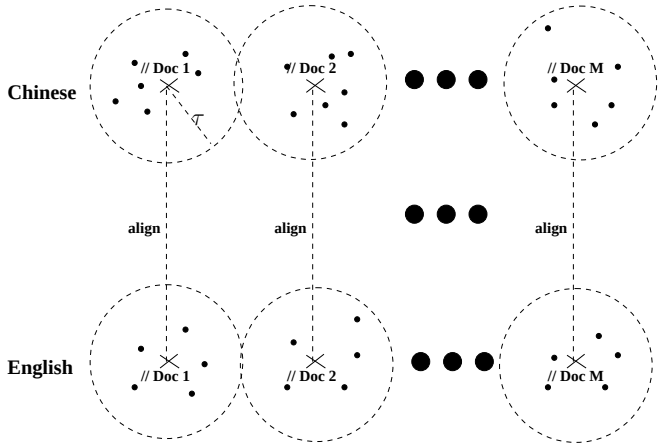

Figure 5.6: Parallel clusters formed by monolingual documents (black dots) using M parallel seed documents.

ness assumption (Chapelle et al., 2006) says that if two points x_1 and x_2 are close in a high-density region, the corresponding outputs y_1 and y_2 should be close as well. In our setting, each parallel source-target document pair is treated as an input-output point in some spaces. With the smoothness assumption, we associate a monolingual non-parallel document to the closest parallel document via a document similarity measure and discard those that are distant to all parallel documents. As a result, a partial alignment between a source and target monolingual non-parallel documents is recovered at the document cluster level. The parallel clusters then serve as constraints for the cluster-based bilingual LSA training via the Lagrangian theory (Tam and Schultz, 2009).

5.6.1 Parallel Clusters

We propose a platform for integrating monolingual non-parallel documents via parallel clusters. The concept of parallel clusters is depicted in Figure 5.6. The idea is to use parallel seed documents to form the initial clusters containing only a single document. Then a parallel cluster is populated by associating each monolingual source and target documents to the corresponding closest parallel document based on a similarity measure. We represent each document d as a K-dimensional topic posterior vector $p(k|d)$ inferred

by monolingual LSA. The distance between two documents is computed as follows:

$$D(d, d_c) = \sum_{k=1}^{K} \sqrt{p(k|d) \cdot p(k|d_c)} \quad (5.14)$$

where d_c is a parallel seed document in cluster c. When the distance is equal to one, the input documents are considered identical in the topic sense. Consequently, partial document alignment between monolingual documents is recovered at the cluster level. Intuitively, monolingual documents within a cluster are expected to come from similar topics. We prevent "noisy" monolingual documents from folding into a cluster by setting a threshold τ so that any monolingual document with distance larger than τ from all cluster centroids is removed.

5.6.2 Cluster-based Bilingual LSA Training

The development of cluster-based bilingual LSA training assumes that the average topic distribution between a source and target cluster is identical. In other words, a one-to-one topic correspondence among a pair of parallel cluster is assumed. Given a pair of parallel cluster $C = \{C^{(i)}, C^{(j)}\}$ where i and j represent the index of the source and target language respectively, this assumption can be encoded as follows:

$$\forall k : E[p^{(i)}(k|C^{(i)})] = E[p^{(j)}(k|C^{(j)})] \quad (5.15)$$

$$\Longrightarrow \frac{\sum_{d \in C^{(i)}} p^{(i)}(k|d)}{|C^{(i)}|} = \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \quad (5.16)$$

where d and k denote a document and a topic index respectively. For monolingual LSA training using latent Dirichlet allocation (Blei et al., 2003), the lower bound of the log likelihood of a document $W_d = w_1 ... w_{N_d}$, denoted as Q(W_d), is:

$$\begin{aligned}
\log \int_\theta \sum_Z p(W_d, Z, \theta) &= \log \int_\theta p(\theta) \prod_{n=1}^{N_d} p(z_n|\theta) \cdot p(w_n|z_n) \\
&\geq E_q[\log \frac{p(\theta)}{q(\theta)}] + \sum_{n=1}^{N_d} \left(E_q[\log \frac{p(z_n|\theta)}{q(z_n)}] + E_q[\log p(w_n|z_n)] \right) \\
&= \text{Q}(W_d)
\end{aligned}$$

where $Z = z_1 ... z_{N_d}$ and θ denote the latent topic sequence and the topic distribution vector sampled from a Dirichlet prior respectively. The lower-bound value is achieved via the Jensen's inequality using a factorizable variational posterior distribution over the latent variables $q(Z, \theta|d) = q(\theta) \prod_{n=1}^{N_d} q(z_n)$. Therefore, the objective function for bilingual LSA training with a pair of cluster is the sum of the lower-bound log likelihood of the documents in the source and target cluster subject to the topic correspondence constraint in equation 5.16. With the Lagrange multipliers λ_{Ck}, the objective function is shown as follows:

$$\begin{aligned} Q(W; \Lambda, \Gamma) &= \sum_{d \in C^{(i)}} Q^{(i)}(W_d; \Lambda, \Gamma) + \sum_{d \in C^{(j)}} Q^{(j)}(W_d; \Lambda, \Gamma) \\ &+ \sum_{C,k} \lambda_{Ck} \left(\frac{\sum_{d \in C^{(i)}} p^{(i)}(k|d)}{|C^{(i)}|} - \frac{\sum_{d \in C^{(j)}} p^{(j)}(k|d)}{|C^{(j)}|} \right) \\ \text{where } p(k|d) &\approx \frac{\sum_{n=1}^{N_d} q(z_n = k|d)}{N_d} \text{ for large } N_d \end{aligned}$$

To derive the E-steps, we compute the partial derivative of $Q(W; \Lambda, \Gamma)$ with respect to $q^{(i)}(z_n = k|d)$ subject to $\sum_{k=1}^{K} q^{(i)}(z_n = k|d) = 1$ which yields the following solution:

$$q^{(i)}(z_n = k|d) = p^{(i)}(w_{dn}|k) \cdot e^{E_q[\log \theta_k^{(i)}] + \mu_{dn}^{(i)}} \cdot e^{\frac{\lambda_{Ck}}{|C^{(i)}| \cdot N_d^{(i)}}} \tag{5.17}$$

where $\mu_{dn}^{(i)}$ is the Lagrange multiplier for probability normalization in $q^{(i)}(z_n = k|d)$. If we assume that each document has the same number of words so that $N_d \approx N$, we can use equation 5.17 to construct the estimated $p(k|d)$ which are put back to the left hand side of equation 5.16. After rearranging terms of the resulting equation, we obtain the following result:

$$e^{\frac{\lambda_{Ck}}{|C^{(i)}| \cdot N}} = \frac{E[p(k|C^{(j)})]}{\frac{1}{|C^{(i)}| \cdot N} \sum_{d \in C^{(i)}} \sum_n p^{(i)}(w_{dn}|k) \cdot e^{E_q[\log \theta_k^{(i)}] + \mu_{dn}^{(i)}}} \tag{5.18}$$

$$\approx \frac{E[p(k|C^{(j)})]}{E[p(k|C^{(i)})]} = r_{j/i}(k|C) \tag{5.19}$$

where $r_{j/i}(k|C)$ is the topic ratio between the target and source cluster in C. Substituting $r_{j/i}(k|C)$ into equation 5.17 and using the E-steps of latent Dirichlet allocation introduced in Chapter 2, we arrive at the following variational E-steps for a source document in $C^{(i)}$:

E-steps:

$$q^{(i)}(z_n = k|d) \propto p^{(i)}(w_{dn}|k) \cdot e^{E_q[\log \theta_k^{(i)}]} \cdot \mathbf{r_{j/i}}(\mathbf{k}|\mathbf{C}) \quad (5.20)$$

$$\gamma_{dk}^{(i)} = \alpha_k^{(i)} + \sum_{n=1}^{N_d^{(i)}} q^{(i)}(z_n = k|d) \quad (5.21)$$

The E-steps resemble those in latent Dirichlet allocation except that an extra term $r_{j/i}(k)$ is introduced to enforce a one-to-one topic correspondence between $C^{(i)}$ and $C^{(j)}$ in equation 5.15. By symmetry, the E-steps for documents on the target language j can be proceeded in a similar fashion. After performing the E-steps on all monolingual documents, $r_{j/i}(k|C)$ is updated using equation 5.19 which are then substituted back to the E-steps iteratively until convergence is reached. The M-step is the same as the derivation in latent Dirichlet allocation.

5.6.3 Experiment

The bilingual LSA training employed parallel Chinese–English corpora from the Donga news websites containing 28k parallel documents with 13M Chinese characters and 9M English words. We applied latent Dirichlet-Tree allocation with 50 latent topics. We employed a small-scale RT04 SMT system to evaluate the performance of crosslingual language model adaptation.

To show the progress of incorporating the pseudo and the real monolingual non-parallel corpora into bilingual LSA, we randomly split the corpora into two parts: 10% of the documents (2.8k) as parallel seed documents and the remaining 90% as pseudo-monolingual documents (25k) where a one-to-one document correspondence was omitted. We compared different bilingual LSA training scenarios from A to F as shown in Table 5.13. Scenario A used only 10% of the parallel corpora as a baseline. Scenario B incorporated the remaining 90% of pseudo-monolingual portion in addition to scenario A without constraint, i.e. the topic ratios $r_{*/*}(k|C)$ in equation 5.20 were set to 1 meaning that parallel clusters were not applied. Scenario C had similar settings as scenario B except that the parallel clusters were applied. Scenario D resembled scenario B except using the real

Table 5.13: Bilingual LSA training scenarios with pseudo monolingual (p-mono) Donga news and real monolingual Xinhua news 2004 corpora.

Scenario	# Chinese doc	# English doc
A. 10% // (baseline)	2.8k	2.8k
B. + p-mono (blind)	+14.5k	+13.7k
C. + p-mono (// cluster)	+14.5k	+13.7k
D. + real mono (blind)	+18.4k	+19.2k
E. + real mono (// cluster)	+18.4k	+19.2k
F. 100% // (golden-line)	28k	28k

monolingual non-parallel corpora from the Chinese and English Xinhua news 2004 corpora. Scenario E shared the same rationale as scenario C but using the real monolingual non-parallel corpora. Scenario F served as an ideal case where 100% parallel corpora were available. We compared these scenarios for crosslingual language model adaptation at the story level using the manual transcription of the RT04 test set of the source language comprising CCTV, NTDTV and RFA shows. Performance metrics were target word perplexity and BLEU.

5.6.4 Results

Table 5.14 shows the top *new* words discovered by bilingual LSA from the pseudo-monolingual corpora after filtering out words which were already covered in the parallel seed documents. The new words tended to be crosslingual word triggers suggesting that our approach worked well in the pseudo-monolingual case. Table 5.15 shows the results in target word perplexity and BLEU after crosslingual language model adaptation via marginal adaptation. The baseline bilingual LSA in scenario A showed reduction in perplexity compared to the unadapted language model which was surprisingly decent given the small amount of parallel training data. Incorporating pseudo-monolingual documents further reduced perplexity in scenario C compared to scenario A, and approached to the ideal case in scenario F using the full parallel corpora. Given that scenario A and F set the over-

Table 5.14: New topical words which are not covered by the parallel corpora are extracted by bilingual LSA using pseudo-monolingual corpora. Words on the Chinese side are translated into English for illustration purpose.

Topics	Top *new* words sorted by $p(w\|k)$
"CH (Art)"	film reward, **ballet**, art festival, **ballet** club,
	edinburgh, orchestra, rock-n-roll, **spartacus**
"EN (Art)"	**ballet**, ballads, **edinburgh**, pianist,
	hop, hip, boa, **spartacus**, wax, sf, swan, beast, oscar,
"CH (Economy)"	export rate, life condition, 2nd season
	international oil, **durable**, **greenspan**, trade deficit,
"EN (Economy)"	diesel, **greenspan**, **durable**, dived,
	revalation, bottomed, recessions, nonferrous, iea
"CH (Electronics)"	router, broadband service, album,
	connector, bundled with, coupon, broadcast
"EN (Electronics)"	3g, bro, pixel, copying, piracy, sw,
	fingerprint, telephony, cartoonists, sos

all upper-bound and lower-bound perplexity of 117 and 111 respectively, our approach was reasonable with the overall perplexity of 113. On the other hand, folding in monolingual corpora without parallel clusters as constraints in scenario B degraded perplexity compared to scenario A. This indicates that using parallel clusters as constraints are crucial in incorporating monolingual non-parallel documents. We observed a similar trend in perplexity performance when the real monolingual corpora were employed, reducing perplexity in scenario E, but deteriorating perplexity in scenario D even further compared to scenario B. This implies that our approach becomes critical for incorporating real monolingual documents. Regarding the translation performance, consistent improvement in BLEU was indicated in scenario C and E similar to their perplexity performance although the gain was not significant. But since the difference in BLEU between the best scenario F and the baseline scenario A was only 0.21%, the gain after incorporating monolingual corpora using our approach was reasonable with 0.19% and 0.15% improvement in scenario

Scenario	CCTV	NTDTV	RFA	OVERALL
BG EN 4-gram	16.12% (85)	14.04 (127)	8.83 (189)	13.22 (126)
A. 10% // (baseline)	16.26 (78)	14.09 (115)	8.90 (181)	13.28 (117)
B. + p-mono (blind)	16.46 (81)	14.29 (116)	8.68 (189)	13.36 (121)
C. + p-mono (// cluster)	**16.52** (75)	**14.31** (109)	**8.95** (178)	**13.47** (113)
D. + real mono (blind)	15.66 (91)	14.28 (135)	8.87 (192)	13.12 (133)
E. + real mono (// cluster)	**16.30** (76)	**14.40** (114)	**9.04** (178)	**13.44** (115)
F. 100% // (golden line)	16.44 (74)	14.38 (107)	9.06 (172)	13.49 (111)

Table 5.15: Crosslingual language model adaptation performance in BLEU (target perplexity) on different training scenarios for bilingual LSA.

C and E respectively using a single target reference for scoring.

5.7 Summary

We have proposed the bilingual N-gram LSA for crosslingual adaptation for statistical machine translation. Our approach is based on bilingual LSA which enables latent topic distribution to be efficiently transferred from a source language to a target language by enforcing a one-to-one topic correspondence between the source and target LSA. During testing, crosslingual adaptation can be performed simultaneously on SMT models by inferring the topic distribution of a source text and then applying the inferred distribution to the target language. Since adaptation is performed before translation, it does not require the adaptation text to be pre-translated as in monolingual adaptation. Therefore, an immediate impact on the translation output is achieved in a single decoding pass. Rapid LSA bootstrapping for a new language can be performed from a well-trained LSA of another language. Results show that our approach has reduced the word perplexity of the target language model significantly. When the adapted language model or lexicon is applied separately, improvement in BLEU and NIST scores has been observed. When both models are applied simultaneously, the gain is additive. Trigram LSA has improved the

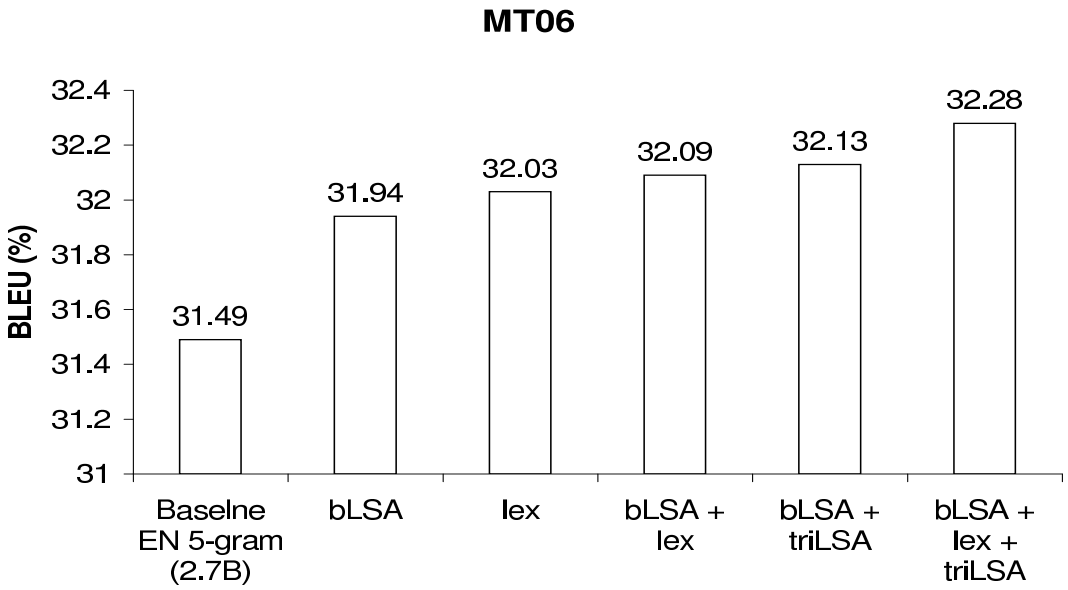

Figure 5.7: Overall translation performance summary after crosslingual adaptation using the GALE-P2.5 system for text translation.

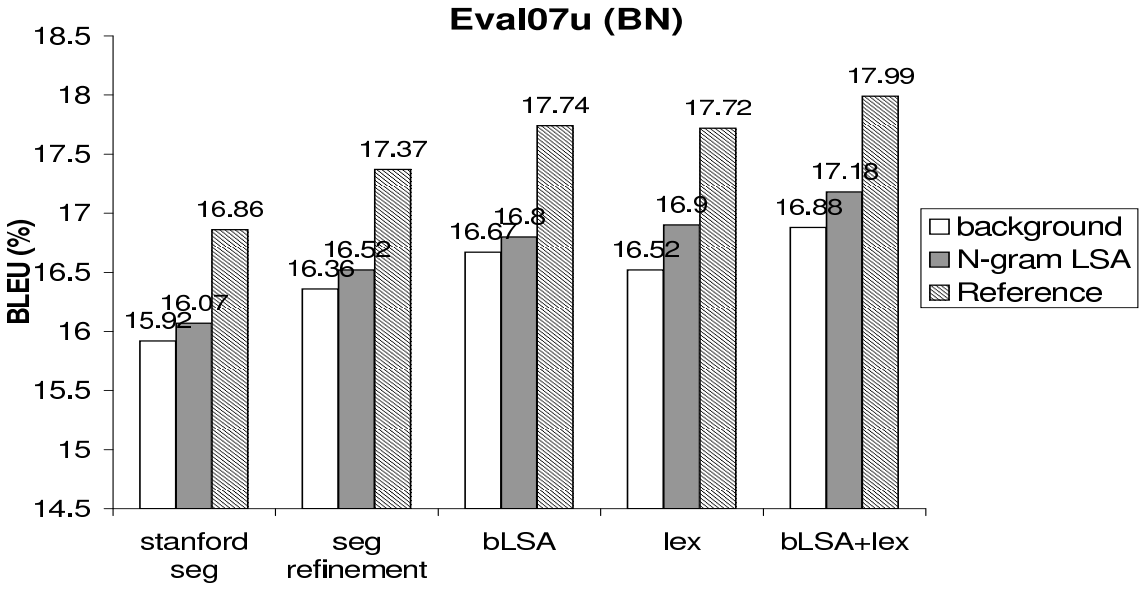

Figure 5.8: Overall translation performance summary after crosslingual adaptation using the GALE-P2.5 system for end-to-end translation.

translation performance further compared to LSA. On the medium-scale SMT system, the improvement is statistically significant at a 95% confidence interval with respect to the unadapted baseline. Effective language model adaptation improves word reordering while a better translation lexicon leads to a better phrase table. Our approach works well on a large-scale evaluation with consistent improvement using the GALE-P2.5 SMT system. The improvement in the NIST score is statistically significant at a 95% confidence interval. For end-to-end translation, the gain from improved recognition accuracy and sharper translation models after topic adaptation are additive. Figure 5.7 and Figure 5.8 summarize our contributions towards better translation performance using our GALE-P2.5 system on text translation and end-to-end translation.

The limitation of bilingual LSA training using parallel documents is relaxed via incorporating monolingual non-parallel documents using a semi-supervised approach. We have enforced a one-to-one topic correspondence between the parallel clusters populated with monolingual non-parallel documents. The proposed bilingual LSA training is based on variational Expectation-Maximization algorithm and the Lagrangian theory. Our approach have incorporated monolingual corpora successfully and has yielded slightly better crosslingual language model adaptation performance compared to the baseline without the monolingual non-parallel corpora. Incorporating monolingual corpora without the parallel clusters can lead to severe performance degradation, implying that a one-to-one topic correspondence between the parallel clusters is crucial. This approach has potential in building a crosslingual word trigger model with enhanced vocabulary coverage in a resource deficient scenario where only a small amount of parallel documents are available.

Language model adaptation, translation lexicon adaptation and incorporating monolingual non-parallel corpora address different aspects of statistical machine translation. Intuitively, language model adaptation improves fluency via better word reordering. Translation lexicon adaptation reduces ambiguity of word translation options. Incorporating monolingual non-parallel corpora potentially addresses the out-of-vocabulary issue. We speculate that language model adaptation is more important than translation lexicon adaptation, which is in turn more important than incorporating monolingual non-parallel corpora. It is widely accepted that the average length of a phrase match between source text

and a phrase table governs the quality of local word reordering, and thus the quality of translation. When the length of a matched source phrase gets large, the number of translation options of the source phrase decreases. Therefore, the effectiveness of translation lexicon adaptation may be reduced. The impact of incorporating monolingual non-parallel corpora may be the least, especially when the bilingual resources are rich since the out-of-vocabulary issue is less severe. However, its potential for future research is expected to be the largest among language model adaptation and translation lexicon adaptation.

Chapter 6

Conclusions

Adaptation is an indispensable part of speech and natural language applications due to mismatch between a background model and a test domain. Our unified topic adaptation framework via latent semantic analysis reduces this mismatch within and across languages.

6.1 Contributions

We list out the contributions as follows:

- We have shown that latent Dirichlet allocation for Bayesian latent semantic analysis (LSA) can be applied for unsupervised language model adaptation via topic caching. Topic caching is more robust against speech recognition errors compared to word caching. As a cache model, adaptation can be performed rapidly in terms of a small amount of adaptation text. Our results have shown improvement in recognition performance (Section 4.1.2).
- We have proposed latent Dirichlet-Tree allocation to model topic correlation to generalize latent Dirichlet allocation. Our model can be trained using an efficient variational Expectation-Maximization algorithm. Our approach addresses the model initialization issue via a structured topic prior so that our model training converges

faster than latent Dirichlet allocation, which is crucial for training on a large volume of data (Section 4.2.2).

- We have proposed incremental marginal adaptation for lattice rescoring that has yielded additive improvement after topic caching (Section 4.3.2).

- We have employed N-gram LSA to relax the "bag-of-word" assumption. Efficient model training and smoothing are the major problems in N-gram LSA for large-scale application. We have addressed the model training issue via a bootstrapping algorithm and a variational Expectation-Maximization algorithm. We have investigated a fractional Kneser-Ney approach to smooth N-gram LSA that employs fractional counts. The smoothing algorithm generalizes the original formulation, generating a more compact model compared to the Witten-Bell smoothing. In addition, the smoothing algorithm applies to other higher-order language model, including a factored language model. Our results have shown that N-gram LSA yields significant improvement in recognition performance compared to LSA for large-scale GALE evaluations on two languages: Mandarin and Arabic (Section 4.4.6).

- We have extended our monolingual topic adaptation to crosslingual adaptation via bilingual LSA. Since bilingual LSA captures a one-to-one correspondence between a source and target language, adaptation on the target language can be performed using information from the source language. As a consequence, pre-translation of a source text is not required to adapt the target models in statistical machine translation, giving immediate impact before translation. We have shown that adapting a language model and translation lexicon together have yielded additive improvement in translation quality. Applying N-gram LSA on the target language as an additional language model feature function further improves translation quality for text translation (Section 5.4). For end-to-end translation, we have shown that the gain from better recognition accuracy and sharper translation models after topic adaptation are additive (Section 5.5).

- We have relaxed the requirement of using parallel corpora for bilingual LSA training via incorporating monolingual non-parallel corpora in a semi-supervised fashion.

Our approach improves the performance of crosslingual language model adaptation compared to blind incorporation of monolingual non-parallel corpora. In addition, our approach has a potential in building a crosslingual word trigger with an enhanced vocabulary coverage in a resource deficient scenario where parallel resources are scarce (Section 5.6.4).

6.2 Summary of Results

Using our baseline systems trained on sufficiently large amount of training data, we have achieved the following results:

For automatic speech recognition, we have achieved significant relative reduction in recognition error rates in the range of 5–7% and 3% respectively on the Mandarin and Arabic test sets after applying the proposed unsupervised language model adaptation for the GALE-P3 evaluation. The reductions are statistically significant at a 0.1% significance level compared to an unadapted baseline that employs state-of-the-art techniques such as discriminative training and acoustic model adaptation.

For statistical machine translation, we have yielded significant relative improvement in BLEU and NIST by 4.8% and 4.2% respectively on MT06 after topic adaptation using the GALE Chinese-English development system (500M). The improvement is statistically significant at a 95% confidence interval compared to an unadapted baseline. In addition, we have achieved significant improvement in NIST using the GALE-P2.5 Chinese-English system compared to an unadapted baseline. For end-to-end translation, we have yielded 1.8% and 1.1% relative improvement in BLEU and NIST respectively after end-to-end topic adaptation compared to the unadapted baseline.

In summary, topic adaptation for language models encourages better recognition accuracy on topical words in automatic speech recognition. The results are important for applications such as spoken language understanding and question answering systems since topical words usually carry the most important information of an utterance compared to stopwords which contribute much less to the automatic understanding process.

6.3 Future Challenges and Potentials

Although our topic adaptation approach is beneficial to overall recognition performance, less gain is observed on broadcast conversation compared to broadcast news. A major challenge is the mismatch between our training corpora, which are mostly newspaper text with relatively few audio transcript. In addition, broadcast conversation is spontaneous in speaking style with much disfluency, including hesitation and repetition. The spontaneous events in broadcast conversation are usually independent of topic context. Thus we believe that N-gram latent semantic analysis is not a good model for broadcast conversation. An improved language modeling approach for broadcast conversation deserves much attention for future research.

Statistical machine translation usually requires parallel sentences so that a phrase table can be built for translation. A major hurdle for this approach is the inability to translate out-of-vocabulary terms that are not covered in the phrase table. As a consequence, the benefit of crosslingual language model adaptation may be reduced since the N-gram entries containing the out-of-vocabulary terms are never used during decoding. This problem is severe on minority language pairs that parallel resources are deficient, producing significant amount of out-of-vocabulary terms. Therefore, leveraging monolingual non-parallel corpora is a research direction to discover potential translation candidates for out-of-vocabulary terms.

We envision that our approaches can be extended to a different application, such as crosslingual semantic search. Semantic search is an interesting application where the goal is to deliver information queried by a user rather than having a user sort through a list of loosely related keyword results. Bilingual LSA lends itself well to the search application since it can work as a crosslingual trigger between an input user query and output objects without machine translation or dictionary lookup. In addition, bilingual LSA can also be extended to other multimedia such as images and videos. For instance, text annotation on images has been shown feasible using latent Dirichlet allocation (Blei and Jordan, 2003).

Many objects in the Internet including images and videos are not annotated and indexed. Without proper indexing, there seems no hope to retrieve these objects given a user

query. On the other hand, a user query and the clicked output objects can be treated as "parallel" data. Our semi-supervised approach for bilingual LSA can be applied similarly to this scenario via the notion of parallel clusters that incorporate unannotated objects into the clusters of a target side. Since our approach has shown potential to discover novel crosslingual word triggers that are not covered in parallel corpora, we speculate that the concept of crosslingual triggering may be applied to crosslingual semantic search to trigger unannotated objects given an input query.

Appendix A

Gibbs Sampling

Gibbs sampling is another useful approach for posterior inference based on Monte Carlo methods. It has a nice property that the Gibbs sampler will converge to the true distribution with sufficiently large number of sampling iterations. However, it is usually slow compared to variational Bayes. Nevertheless, Gibbs sampling has been applied to latent Dirichlet allocation as a convenient alternative to variational Bayes.

A.1 Latent Dirichlet Allocation

The basic principle of Gibbs sampling is to draw samples from a probability distribution conditioned on other variables with some fixed values. In latent Dirichlet allocation, the latent variables are the topic mixture weights θ and the topic index z_i for each word in a document w_1^N. In principle, the conditional distributions to consider are $p(\theta|z_1^N, w_1^N)$ and $p(z_i|z_{-i}, \theta, w_1^N)$ where z_{-i} denotes the topic sequence of w_1^N excluding z_i. However, using the conjugate property between a Dirichlet and a multinomial distribution, we can marginalize out θ easily so that we only need to draw samples for z_i. This approach is known as the collapsed Gibbs sampling (Griffiths and Steyvers, 2004). Assume that the topic-dependent unigram language model $\{p(w|k)\}$ is an unknown variable, we can derive

the conditional distribution for z_i in document d as follows:

$$
\begin{aligned}
p(z_i|w_i, z_{-i}, w_{-i}) &\propto p(w_i|z_i, z_{-i}, w_{-i}) \cdot p(z_i|z_{-i}, w_{-i}) && \text{(A.1)}\\
&\propto p(w_i|z_i) \cdot p(z_i|z_{-i}) && \text{(A.2)}\\
\text{where } p(z_i|z_{-i}) &= \int_\theta p(z_i|\theta) \cdot p(\theta|z_{-i})\ d\theta && \text{(A.3)}\\
&= \int_\theta \theta_{z_i} \cdot p(\theta|z_{-i})\ d\theta && \text{(A.4)}\\
&= E[\theta_{z_i}|z_{-i}] && \text{(A.5)}\\
&= \frac{\alpha_k + C_d(k)_{-i}}{\sum_{k'=1}^{K} \alpha_{k'} + C_d(k')_{-i}} && \text{(A.6)}\\
\text{and } p(w_i = w|z_i = k) &= \frac{C(w, k)_{-i} + \xi}{C(k)_{-i} + V \cdot \xi} && \text{(A.7)}
\end{aligned}
$$

where $C_d(k)_{-i}$ denotes the total integral counts of topic k according to the current values of other topic samples $z_{j\neq i}$ in document d, i.e. $C_d(k)_{-i} = \sum_{j\neq i}^{N} \delta(k, z_j)$. Similarly, $C(w, k)_{-i}$ denotes the total integral counts of word w assigned to topic k in the corpus except the count from word w_i. A small count ξ is applied for simple Laplace smoothing to avoid zero probability where V denotes the size of vocabulary. Informally, Gibbs sampling performs a leave-one-out maximum likelihood estimation of the probability distributions by holding out the i-th token in equation A.6 and equation A.7.

A.2 Bigram Topic Model

Similarly, the Gibbs sampling formula for bigram topic model (Wallach, 2006) is as follows:

$$
\begin{aligned}
p(z_i|w_i, z_{-i}, w_{-i}) &\propto p(w_i|w_{i-1}, z_i) \cdot p(z_i|z_{-i}) && \text{(A.8)}\\
&\propto p(w_i = v|w_{i-1} = u, z_i = k) \cdot p(z_i|z_{-i}) && \text{(A.9)}\\
&\propto \frac{C(u, v, k)_{-i} + \xi}{C(u, k)_{-i} + V \cdot \xi} \cdot \frac{\alpha_k + C_d(k)_{-i}}{\sum_{k'=1}^{K} \alpha_{k'} + C_d(k')_{-i}} && \text{(A.10)}
\end{aligned}
$$

where $C(u, v, k)_{-i}$ denotes the bigram count of (u, v) assigned to topic k excluding the i-th word token.

Appendix B

Latent Dirichlet-Tree Allocation

Similar to latent Dirichlet allocation, we apply variational expectation-maximization algorithm for model training. We define the following notations for the derivation:

$\alpha_{j\cdot}$	A Dirichlet parameter of the j-th node in a Dirichlet-Tree.
$b_{j\cdot}$	A probability vector over the branches sampled from the j-th node so that $\sum_c b_{jc} = 1$ where c is a branch index.
$\delta_{jc}(k)$	A 0-1 indicator which sets to one when a path from a root node to the k-th leaf node passes through the c-th branch of the j-th Dirichlet node; and zero otherwise.
$\{\beta_{vk}\}$	A shorthand for $p(v\|k)$ where v is the vocabulary index and k is the topic index.
Λ	The model parameters containing $\{\alpha_{j\cdot}\}$ and $\{\beta_{vk}\}$.
w_1^N	A word sequence of an input document.
M	The number of documents in the training corpora.
ϕ_{nk}	A shorthand for variational multinomial posterior $q(z_n = k)$ of the n-th word assigned to topic k in a document.

The latent variables are the topic assignment $z_1^N = z_1 z_2 ... z_N$ and the branching variables $b_1^J = b_1 b_2 ... b_J$ where J denotes the number of nodes in the Dirichlet-Tree. Using

variational Bayes, the auxiliary function to maximize is:

$$Q(w_1^N;\Lambda,\Gamma) = E_q[\log\frac{p(w_1^N,z_1^N,b_1^J;\Lambda)}{q(z_1^N,b_1^J;\Gamma)}] \quad \text{(B.1)}$$

$$= E_q[\log p(w_1^N|z_1^N)] + E_q[\log\frac{p(z_1^N|b_1^J)}{q(z_1^N)}] + E_q[\log\frac{p(b_1^J)}{q(b_1^J)}] \quad \text{(B.2)}$$

The first term in equation B.2 is computed as follows:

$$E_q[\log p(w_1^N|z_1^N)] = E_q[\log\prod_{n=1}^{N}\beta_{w_n z_n}] = \sum_{n=1}^{N} E_q[\log\beta_{w_n z_n}] \quad \text{(B.3)}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\log\beta_{w_n k} \quad \text{(B.4)}$$

Using the following relation:

$$p(z_n=k|b_1^J) = \prod_{jc} b_{jc}^{\delta_{jc}(k)} \quad \text{(B.5)}$$

$$\Longrightarrow \log P(z_n=k|b_1^J) = \sum_{jc}\delta_{jc}(k)\cdot\log b_{jc} \quad \text{(B.6)}$$

The second term in equation B.2 is computed as follows:

$$E_q[\log\frac{p(z_1^N|b_1^J)}{q(z_1^N)}] = E_q[\log\prod_{n=1}^{N}\frac{p(z_n|b_1^J)}{q(z_n)}] \quad \text{(B.7)}$$

$$= \sum_{n=1}^{N} E_q[\log p(z_n|b_1^J)] - E_q[\log q(z_n)] \quad \text{(B.8)}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\left(\sum_{jc}\delta_{jc}(k)\cdot E_q[\log b_{jc}] - \log\phi_{nk}\right) \quad \text{(B.9)}$$

Using the following relation:

$$p(b_1^J) = \prod_{j=1}^{J}\text{Dirichlet}(b_j;\alpha_j) = \prod_{j=1}^{J}\left(\prod_c\frac{\Gamma(\sum_c\alpha_{jc})}{\prod_c\Gamma(\alpha_{jc})}\cdot b_{jc}^{\alpha_{jc}-1}\right) \quad \text{(B.10)}$$

The third term in equation B.2 is computed as follows:

$$E_q[\log \frac{p(b_1^J)}{q(b_1^J)}] = \sum_{j=1}^{J} \left(\log \Gamma(\sum_c \alpha_{jc}) - \sum_c \log \Gamma(\alpha_{jc}) + \sum_c (\alpha_{jc} - 1)(E_q[\log b_{jc}]) \right) - \sum_{j=1}^{J} \left(\log \Gamma(\sum_c \gamma_{jc}) - \sum_c \log \Gamma(\gamma_{jc}) + \sum_c (\gamma_{jc} - 1)(E_q[\log b_{jc}]) \right)$$

B.1 Variational Multinomials

We isolate terms of the auxiliary function in equation B.2 which depends on ϕ_{nk}. Then we introduce Lagrange multipliers λ_n for each n-th position to ensure that $\sum_{k=1}^{K} \phi_{nk} = 1$.

$$Q_{[\phi_{nk}]} = \phi_{nk} \log \beta_{w_n k} + \phi_{nk} \left(\sum_{jc} \delta_{jc}(k) \cdot E_q[\log b_{jc}] - \log \phi_{nk} \right) + \lambda_n \phi_{nk} \quad \text{(B.11)}$$

By computing the partial derivative of $Q_{[\phi_{nk}]}$ with respect to ϕ_{nk}, we have:

$$\frac{\partial Q_{[\phi_{nk}]}}{\partial \phi_{nk}} = \log \beta_{w_n k} + \sum_{jc} \delta_{jc}(k) \cdot E_q[\log b_{jc}] - \log \phi_{nk} - 1 + \lambda_n = 0 \quad \text{(B.12)}$$

$$\implies \phi_{nk} \propto \beta_{w_n k} \cdot e^{\sum_{jc} \delta_{jc}(k) \cdot E_q[\log b_{jc}]} \quad \text{(B.13)}$$

B.2 Variational Dirichlet

We isolate terms of the auxiliary function in equation B.2 which depends on γ_{jc}.

$$\begin{aligned}
Q_{[\gamma_{jc}]} &= \left(\sum_c(\alpha_{jc} + \sum_{nk}\phi_{nk}\cdot\delta_{jc}(k) - \gamma_{jc})\cdot(\Psi(\gamma_{jc}) - \Psi(\sum_c \gamma_{jc}))\right) \\
&\quad - \log\Gamma(\sum_c \gamma_{jc}) + \log\Gamma(\gamma_{jc}) \\
\frac{\partial Q_{[\gamma_{jc}]}}{\partial \gamma_{jc}} &= \Psi'(\gamma_{jc})\cdot(\alpha_{jc} + \sum_{nk}\phi_{nk}\cdot\delta_{jc}(k) - \gamma_{jc}) \\
&\quad -\Psi'(\sum_c \gamma_{jc})\cdot\sum_c(\alpha_{jc} + \sum_{nk}\phi_{nk}\cdot\delta_{jc}(k) - \gamma_{jc}) \\
&= 0 \\
\Longrightarrow\ \gamma_{jc} &= \alpha_{jc} + \sum_{nk}\phi_{nk}\cdot\delta_{jc}(k)
\end{aligned}$$

B.3 Conditional Multinomials

We isolate terms of the auxiliary function in equation B.2 which only depends on β_{vk}. By considering all training documents indexed by d and introducing Lagrange multipliers λ_k to enforce $\sum_{v=1}^{V}\beta_{vk}$ = 1 for all k, we get:

$$Q_{[\beta]} = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\phi_{dnk}\log\beta_{w_{dn}k} + \sum_{k=1}^{K}\lambda_k(\sum_{v=1}^{V}\beta_{vk} - 1) \tag{B.14}$$

We take the derivative with respect to β_{vk} and set it to zero, we get:

$$\beta_{vk} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dnk}\cdot\delta(w_{dn}, v) \tag{B.15}$$

B.4 Dirichlet Node in a Dirichlet-Tree

The terms which contain α_j of the j-th node in a Dirichlet-Tree are:

$$Q_{[\alpha_j]} = \sum_{d=1}^{M} \left(\log\Gamma(\sum_c \alpha_{jc}) - \sum_c \log\Gamma(\alpha_{jc}) + \sum_c (\alpha_{jc}-1)\cdot(\Psi(\gamma_{djc}) - \Psi(\sum_c \gamma_{djc})) \right)$$

Taking the derivative with respect to α_{jc} gives:

$$\frac{\partial Q_{[\alpha_j]}}{\partial \alpha_{jc}} = -M \cdot \left(\Psi(\alpha_{jc}) - \Psi(\sum_c \alpha_{jc}) \right) + \sum_{d=1}^{M} \left(\Psi(\gamma_{djc}) - \Psi(\sum_c \gamma_{djc}) \right)$$

Simple constrained gradient ascent can be applied to ensure that $\alpha_{jc} > 0$ via parameter transformation: $\log(.)$: $\tilde{\alpha}_{jc} = \log \alpha_{jc}$.

B.5 Alternative Proof

We can use the generic solution of the variational E-steps to derive the formula for latent Dirichlet-Tree allocation. Recall that the generic solution has the following form:

$$q(z_j) \propto e^{E_q[\log p(X,Z;\Lambda)]_{\setminus z_j}} \tag{B.16}$$

where the expectation is taken over all other latent variables $\{z_i\}$ excluding z_j. Instead of considering the full joint distribution for expectation in equation B.16, only a subset of local conditional distributions involving z_j are actually involved, that is $p(z_j|\cdot)$ or $p(\cdot|\cdot z_j \cdot)$. The remaining conditional distributions which do not contain z_j are cancelled out due to probability normalization.

In latent Dirichlet-Tree allocation, the observation is the word sequence w_1^N, and the latent variables are the topic assignments z_1^N and the branching probabilities of each node in the Dirichlet tree b_1^J. Using the relation $p(z|b_1^J) = \prod_{jc} b_{jc}^{\delta_{jc}(z)}$ and applying equation B.16,

the variational posterior of a Dirichlet node j is:

$$
\begin{aligned}
q(b_j) &\propto e^{E_q[\log p(b_j)]_{\setminus b_j} + \sum_{i=1}^{N} E_q[\log p(z_i|b_1^J)]_{\setminus b_j}} && \text{(B.17)} \\
&= e^{\log p(b_j) + \sum_{i=1}^{N} \sum_c E_q[\delta_{jc}(k)] \log b_{jc}} \cdot \\
&\quad \underbrace{e^{\sum_{i=1}^{N} \sum_{j' \neq j}^{J} \sum_c E_q[\delta_{j'c}(k)] E_q[\log b_{j'c}]}}_{\text{independent of } b_j} && \text{(B.18)} \\
&\propto p(b_j) e^{\sum_{i=1}^{N} \sum_c E_q[\delta_{jc}(k)] \log b_{jc}} && \text{(B.19)} \\
&= \text{Dirichlet}(b_j) \prod_c b_{jc}^{\sum_{i=1}^{N} E_q[\delta_{jc}(k)]} && \text{(B.20)} \\
&\propto \prod_c b_{jc}^{\alpha_{jc}-1} \prod_c b_{jc}^{\sum_{i=1}^{N} E_q[\delta_{jc}(k)]} && \text{(B.21)} \\
&= \prod_c b_{jc}^{\alpha_{jc} + \sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i=k) \delta_{jc}(k) - 1} && \text{(B.22)} \\
&= \text{Dirichlet}(\{\gamma_{jc}\}) && \text{(B.23)}
\end{aligned}
$$

where $\gamma_{jc} = \alpha_{jc} + \sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k) \delta_{jc}(k)$. Similarly, the variational posterior of a topic assignment z_i for word w_i is:

$$
\begin{aligned}
q(z_i) &\propto e^{E_q[\log p(z_i|b_1^J)]_{\setminus z_i} + E_q[\log p(w_i|z_i)]_{\setminus z_i}} && \text{(B.24)} \\
&= e^{\sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}] + \log p(w_i|z_i)} && \text{(B.25)} \\
&\propto p(w_i|z_i) e^{\sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}]} && \text{(B.26)}
\end{aligned}
$$

Bibliography

Y. Akita and T. Kawahara. Language model adaptation based on PLSA of topics and speakers. In *Proceedings of International Conference on Spoken Language Processing*, pages 1045–1048, Jeju Island, Korea, October 2004.

J. Bellegarda. Latent semantic mapping: Dimensionality reduction via globally optimal continuous parameter modeling. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 127–132, San Juan, Puerto Rico, November 2005.

J. Bellegarda. Exploiting latent semantic information in statistical language modeling. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 88(8):63–75, August 2000.

Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50:434–451, 2008.

C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems*, volume 15, pages 777–784, 2003.

D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, Vancouver, Britich Columbia, Canada, December 2005.

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

P. F. Brown, V. J. D Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based N-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

P. F. Brown, S. D. Pietra, V. J. D Pietra, and R. L. Mercer. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1994.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

S. Chen, K. Seymore, and R. Rosenfeld. Topic adaptation for language modeling using unnormalized exponential models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 681–684, Seattle, Washington, USA, May 1998.

J. T. Chien and C. H. Chueh. Latent Dirichlet language model for speech recognition. In *IEEE Workshop on Spoken Language Technology*, pages 201–204, Goa, India, December 2008.

P. Clarkson and T. Robinson. The applicability of adaptive language modeling for the broadcast news task. In *Proceedings of International Conference on Spoken Language Processing*, pages 233–236, Sydney, Australia, November–December 1998.

P. R. Clarkson and A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 799–802, Munich , Bavaria, Germany, April 1997.

N. Coccaro and D. Jurafsky. Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, November–December 1998.

J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

G. Doddington. Automatic evaluation of MT quality using N-gram co-occurrence statistics. In *Proceedings of human language technology conference*, pages 138–145, San Diego, California, USA, March 2002.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., second edition, 2001.

M. Eck, S. Vogel, and A. Waibel. Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of LREC*, Lisbon, Portugal, May 2004.

M. Federico. Language model adaptation through topic decomposition and MDI estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 773–776, Orlando, Florida, USA, May 2002.

M. Finke, J. Fritsch, P. Geutner, K. Ries, T. Zeppenfeld, and A. Waibel. The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proceedings of the LVCSR Hub5-e Workshop*, 1997.

M. J. F Gales. Maximum likelihood linear transformation for HMM-based speech recognition. Technical Report CUED/F-INFENG/TR291, Cambridge University, 1997.

M. J. F Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 7(3):272–281, May 1999.

D. Gildea and T. Hofmann. Topic-based language models using EM. In *Proceedings of European Conference on Speech Communication and Technology*, pages 2167–2170, Budapest, Hungary, September 1999.

L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 532–535, Glasgow, Scotland, May 1989.

T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of National Academy of Science*, volume 101(1), pages 5228–5235, 2004.

A. Heidel, H. Chang, and L. Lee. Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm. In *Proceedings of Interspeech*, pages 2361–2364, Antwerp, Belgium, August 2007.

A. Hildebrand, K. Rottmann, T. Notari, Q. Gao, S. Hewavitharana, N. Bach, and S. Vogel. Recent improvements in the CMU large-scale Chinese-English SMT system. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–80, The Ohio State University, Columbus, Ohio, USA, June 2008.

H. Hsiao and T. Schultz. Generalized discriminative feature transformation for speech recognition. In *Proceedings of Interspeech*, Brighton, UK, September 2009.

H. Hsiao, M. Fuhs, Y. C. Tam, Q. Jin, and T. Schultz. The CMU-InterACT 2008 Mandarin transcription system. In *Proceedings of Interspeech*, pages 1445–1448, Brisbane, Australia, September 2008.

H. Hsiao, Y. C. Tam, and T. Schultz. Generalized baum-welch algorithm for discriminative training on large vocabulary continuous speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3769–3772, Taipei, Taiwan, April 2009.

B. J. Hsu and J. Glass. Style and topic language model adaptation using HMM-LDA. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Pro-*

cessing, pages 373–381, Sydney, Australia, July 2006. Association for Computational Linguistics.

S. Y. Dennis III. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Communications in Statistics – Theory and Methods*, 20(12):4069–4081, 1991.

R. Iyer and M. Ostendorf. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7(1):30–39, Jan 1999.

Qin Jin and Tanja Schultz. Speaker segmentation and clustering in meetings. In *Proceedings of the International Conference on Spoken Language Processing*, pages 597–600, Jeju Island, Korea, October 2004.

M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Machine Learning*, volume 37(2), pages 183–233, 1999.

W. Kim. *Language model adaptation for automatic speech recognition and statistical machine translation*. PhD thesis, Johns Hopkins University, 2004. Distributed by JHU.

W. Kim and S. Khudanpur. Language model adaptation using cross-lingual information. In *Proceedings of European Conference on Speech Communication and Technology*, pages 3129–3132, Geneva, Switzerland, September 2003.

W. Kim and S. Khudanpur. Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *ACM Transactions on Asian Language Information Processing*, 3(2):94–112, June 2004.

R. Kneser and H. Ney. Improved backing-off for M-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, Michigan, USA, May 1995.

R. Kneser, J. Peters, and D. Klakow. Language model adaptation using dynamic marginals. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1971–1974, Rhodes, Greece, September 1997.

P. Koehn, F. J. Och, and D. Marcu. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, pages 48–54, Edmonton, Canada, May 2003.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007: proceedings of demo and poster sessions*, pages 177–180, Prague, Czech Republic, June 2007.

R. Kuhn and R. De Mori. A cache-based natural language model for speech reproduction. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1990.

C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, April 1995.

W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM, 2006.

Y. Liu and F. Liu. Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4921–4924, Las Vegas, Nevada, USA, April 2008.

M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, Washington, May 1999.

U. Manber and E. W. Myers. Suffix arrays: A new method for on-line string searches. In *SIAM Journal on Computing*, pages 935–948, 1993.

T. Minka. The Dirichlet-Tree distribution (Technical Report), 1999.

D. Mrva and P. C. Woodland. A PLSA-based language model for conversational telephone speech. In *Proceedings of International Conference on Spoken Language Processing*, pages 2257–2260, Jeju Island, Korea, October 2004.

M. Noamany, T. Schaaf, and T. Schultz. Advances in the CMU-InterACT Arabic Gale transcription system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 129–132, Rochester, New York, USA, April 2007. Association for Computational Linguistics.

F. J. Och. Minimum error rate training in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, 2003.

F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA, July 2002.

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In Pascale Fung and Joe Zhou, editors, *Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, USA, June 1999.

D. Pallett, W. Fisher, and J. Fiscus. Tools for the analysis of benchmark speech recognition tests. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 97–100, Albuquerque, New Mexico, USA, April 1990.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association*

for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

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

M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel. Speech Translation Enhanced Automatic Speech Recognition. In *Automatic Speech Recognition and Understanding Workshop*, 2005b.

I. Porteous, A. Ascuncion, D. Newman, A. Ihler, P. Smyth, and M. Welling. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 569–577, New York, NY, USA, 2008. ACM.

D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University Engineering Dept, 2003.

D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah. Boosted MMI for model and feature-space discriminative training. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4057–4060, Las Vegas, Nevada, USA, April 2008.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

I. Rogina. Automatic architecture design by likelihood-based context clustering with crossvalidation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1223–1226, Rhodes, Greece, September 1997.

R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, April 1994.

K. Rottmann and S. Vogel. Word reordering in statistical machine translation with a POS-based distortion model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 171–180, Skövde, Sweden, September 2007.

H. Schwenk. Continuous space language models. *Journal of Computer Speech and Language*, 3(21):492–518, 2007.

H. Soltau, F. Metze, C. Fügen, and A. Waibel. A one-pass decoder based on polymorphic linguistic context. In *Automatic Speech Recognition and Understanding Workshop*, 2001.

A. Stolcke. SRILM - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA, September 2002.

Y. C. Tam and T. Schultz. Language model adaptation using variational Bayes inference. In *Proceedings of Interspeech*, pages 5–8, Lisbon, Portugal, September 2005.

Y. C. Tam and T. Schultz. Unsupervised language model adaptation using latent semantic marginals. In *Proceedings of Interspeech*, Pittsburgh, PA, September 2006.

Y. C. Tam and T. Schultz. Bilingual LSA-based translation lexicon adaptation for spoken language translation. In *Proceedings of Interspeech*, pages 2461–2464, Antwerp, Belgium, August 2007a.

Y. C. Tam and T. Schultz. Correlated latent semantic model for unsupervised language model adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 41–44, Honolulu, Hawaii, USA, April 2007b.

Y. C. Tam and T. Schultz. Correlated bigram LSA for unsupervised LM adaptation. In *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2008.

Y. C. Tam and T. Schultz. Incorporating monolingual corpora into bilingual latent semantic analysis for crosslingual LM adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4821–4824, Taipei, Taiwan, April 2009.

Y. C. Tam, I. Lane, and T. Schultz. Bilingual LSA-based LM adaptation for spoken language translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 520–527, Prague, Czech Republic, June 2007a.

Y. C. Tam, I. Lane, and T. Schultz. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207, December 2007b.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young. MMIE training of large vocabulary speech recognition systems. *Speech Communication*, 22:303–314, 1997.

S. Vogel. PESA: Phrase pair extraction as sentence splitting. In *MT Summit X*, pages 251–258, Phuket, Thailand, September 2005.

S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In *The 16th International Conference on Computational Linguistics*, pages 836–841, Center for Sprogteknologi, Copenhagen, August 1996.

S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel. The CMU statistical translation system. In *MT Summit IX*, pages 402–409, New Orleans, USA, September 2003.

H. M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984, New York, NY, USA, 2006. ACM Press.

I. Witten and T. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4): 1085–1094, 1991.

J. Wu and S. Khudanpur. Building a topic-dependent maximum entropy language model for very large corpora. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 777–780, Orlando, Florida, USA, May 2002.

P. Xu, A. Emami, and F. Jelinek. Training connectionist models for the structured language model. In *Proceedings of Empirical Methods on Natural Language Processing*, Sapporo, Japan, July 2003.

H. Yu, Y. C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz. The ISL RT04 Mandarin Broadcast News Evaluation System. In *EARS Rich Transcription Workshop*, 2004.

P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042, Munich , Bavaria, Germany, April 1997.

Y. Zhang and S. Vogel. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation*, pages 85–94, Baltimore, Maryland, USA, October 2004.

Y. Zhang and S. Vogel. Suffix array and its applications in empirical natural language processing. In *Technical Report CMU-LTI-06-010*, Carnegie Mellon University, Pittsburgh, PA, USA, 2006.

B. Zhao and E. P. Xing. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the Coling/ACL*, pages 969–976, Sydney, Australia, July 2006.

B. Zhao and E. P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2007.

B. Zhao, M. Eck, and S. Vogel. Language model adaptation for statistical machine translation via structured query models. In *Proceedings of Coling*, Geneva, Switzerland, August 2004.