BiTAM: Bilingual Topic AdMixture Models for Word Alignment

Bing Zhao[†] and **Eric P. Xing**[†‡]
{bzhao,epxing}@cs.cmu.edu
Language Technologies Institute[†] and Machine Learning Department[‡]
School of Computer Science, Carnegie Mellon University

Abstract

We propose a novel bilingual topical admixture (BiTAM) formalism for word alignment in statistical machine translation. Under this formalism, the parallel sentence-pairs within a document-pair are assumed to constitute a mixture of hidden topics; each word-pair follows a topic-specific bilingual translation model. Three BiTAM models are proposed to capture topic sharing at different levels of linguistic granularity (i.e., at the sentence or word levels). These models enable word-alignment process to leverage topical contents of document-pairs. Efficient variational approximation algorithms are designed for inference and parameter estimation. With the inferred latent topics, BiTAM models facilitate coherent pairing of bilingual linguistic entities that share common topical aspects. Our preliminary experiments show that the proposed models improve word alignment accuracy, and lead to better translation quality.

1 Introduction

Parallel data has been treated as sets of unrelated sentence-pairs in state-of-the-art statistical machine translation (SMT) models. Most current approaches emphasize within-sentence dependencies such as the distortion in (Brown et al., 1993), the dependency of alignment in HMM (Vogel et al., 1996), and syntax mappings in (Yamada and Knight, 2001). Beyond the sentence-level, corpus-level word-correlation and contextual-level topical information may help to disambiguate translation candidates and word-alignment choices. For example, the most frequent source words (e.g., functional words) are likely to be translated into words which are also frequent on the target side; words of the same topic generally bear correlations and similar translations. Extended contextual information is especially useful when translation models are vague due to their reliance solely on word-pair co-occurrence statistics. For example, the word *shot* in "*It was a nice shot.*" should be translated differently depending on the context of the sentence: a *goal* in the context of sports, or a *photo* within the context of sightseeing. Nida (1964) stated that sentence-pairs are tied by the logic-flow in a document-pair; in other words, the document-pair should be word-aligned as one entity instead of being uncorrelated instances. In this paper, we propose a probabilistic admixture model to capture latent topics underlying the context of document-pairs. With such topical information, the translation models are expected to be sharper and the word-alignment process less ambiguous.

Previous works on topical translation models concern mainly explicit logical representations of semantics for machine translation. This include knowledge-based (Nyberg and Mitamura, 1992) and interlingua-based (Dorr and Habash, 2002) approaches. These approaches can be expensive, and they do not emphasize stochastic translation aspects. Recent investigations along this line includes using word-disambiguation schemes (Carpua and Wu, 2005) and non-overlapping bilingual word-clusters (Wang et al., 1996; Och, 1999; Zhao et al., 2005) with particular translation models, which showed various degrees of success. We propose a new statistical formalism: Bilingual Topic AdMixture model, or BiTAM, to facilitate topic-based word alignment in SMT.

Variants of admixture models have appeared in population genetics (Pritchard et al., 2000) and text modeling (Blei et al., 2003). Statistically, an object is said to be derived from an *admixture* if it consists of a bag of elements, each sampled independently or coupled in some way, from a mixture model. In a typical SMT setting, each document-pair corresponds to an object; depending on a chosen modeling granularity, all sentence-pairs or word-pairs in the document-pair correspond to the elements constituting the object. Correspondingly, a latent topic is sampled for each pair from a prior topic distribution to induce topic-specific translations; and the resulting sentence-pairs and word-pairs are marginally dependent. Generatively, this *admixture formalism* enables word translations to be instantiated by topic-specific bilingual models

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 969–976,
Sydney, July 2006. ©2006 Association for Computational Linguistics

and/or monolingual models, depending on their contexts. In this paper we investigate three instances of the BiTAM model, They are data-driven and do not need hand-crafted knowledge engineering.

The remainder of the paper is as follows: in section 2, we introduce notations and baselines; in section 3, we propose the topic admixture models; in section 4, we present the learning and inference algorithms; and in section 5 we show experiments of our models. We conclude with a brief discussion in section 6.

2 Notations and Baseline

In statistical machine translation, one typically uses parallel data to identify entities such as "word-pair", "sentence-pair", and "document-pair". Formally, we define the following terms[1]:

- A *word-pair* (f_j, e_i) is the basic unit for word alignment, where f_j is a French word and e_i is an English word; j and i are the *position indices* in the corresponding French sentence $\mathbf{f}$ and English sentence $\mathbf{e}$.
- A *sentence-pair* $(\mathbf{f}, \mathbf{e})$ contains the *source* sentence $\mathbf{f}$ of a *sentence length* of J; a *target* sentence $\mathbf{e}$ of *length* I. The two sentences $\mathbf{f}$ and $\mathbf{e}$ are translations of each other.
- A *document-pair* $(\mathbf{F}, \mathbf{E})$ refers to two documents which are translations of each other. Assuming sentences are one-to-one correspondent, a document-pair has a sequence of N parallel sentence-pairs $\{(\mathbf{f}_n, \mathbf{e}_n)\}$, where $(\mathbf{f}_n, \mathbf{e}_n)$ is the $n'th$ parallel sentence-pair.
- A *parallel corpus* $\mathbf{C}$ is a collection of M parallel document-pairs: $\{(\mathbf{F}_d, \mathbf{E}_d)\}$.

2.1 Baseline: IBM Model-1

The translation process can be viewed as operations of word substitutions, permutations, and insertions/deletions (Brown et al., 1993) in noisy-channel modeling scheme at parallel sentence-pair level. The translation lexicon $p(f|e)$ is the key component in this generative process. An efficient way to learn $p(f|e)$ is IBM-1:

$$p(\mathbf{f}|\mathbf{e}) = \prod_{j=1}^{J} \sum_{i=1}^{I} p(f_j|e_i) \cdot p(e_i|\mathbf{e}). \quad (1)$$

IBM-1 has global optimum; it is efficient and easily scalable to large training data; it is one of the most informative components for re-ranking translations (Och et al., 2004). We start from IBM-1 as our baseline model, while higher-order alignment models can be embedded similarly within the proposed framework.

3 Bilingual Topic AdMixture Model

Now we describe the BiTAM formalism that captures the latent topical structure and generalizes word alignments and translations beyond sentence-level via topic sharing across sentence-pairs:

$$\mathbf{E}^* = \arg\max_{\{\mathbf{E}\}} p(\mathbf{F}|\mathbf{E})p(\mathbf{E}), \quad (2)$$

where $p(\mathbf{F}|\mathbf{E})$ is a document-level translation model, generating the document $\mathbf{F}$ as one entity. In a BiTAM model, a document-pair $(\mathbf{F}, \mathbf{E})$ is treated as an admixture of topics, which is induced by random draws of a topic, from a pool of topics, for each sentence-pair. A unique normalized and real-valued vector θ, referred to as a *topic-weight vector,* which captures contributions of different topics, are instantiated for each document-pair, so that the sentence-pairs with their alignments are generated from topics mixed according to these common proportions. Marginally, a sentence-pair is word-aligned according to a unique bilingual model governed by the hidden topical assignments. Therefore, the sentence-level translations are coupled, rather than being independent as assumed in the IBM models and their extensions.

Because of this coupling of sentence-pairs (via topic sharing across sentence-pairs according to a common topic-weight vector), BiTAM is likely to improve the coherency of translations by treating the document as a whole entity, instead of uncorrelated segments that have to be independently aligned and then assembled. There are at least two levels at which the hidden topics can be sampled for a document-pair, namely: the *sentence-pair* and the *word-pair* levels. We propose three variants of the BiTAM model to capture the latent topics of bilingual documents at different levels.

3.1 BiTAM-1: The Frameworks

In the first BiTAM model, we assume that topics are sampled at the sentence-level. Each document-pair is represented as a random mixture of latent topics. Each topic, topic-k, is presented by a topic-specific word-translation table: B_k, which is

[1] We follow the notations in (Brown et al., 1993) for English-French, i.e., $\mathbf{e} \leftrightarrow \mathbf{f}$, although our models are tested, in this paper, for English-Chinese. We use the *end-user terminology* for *source* and *target* languages.

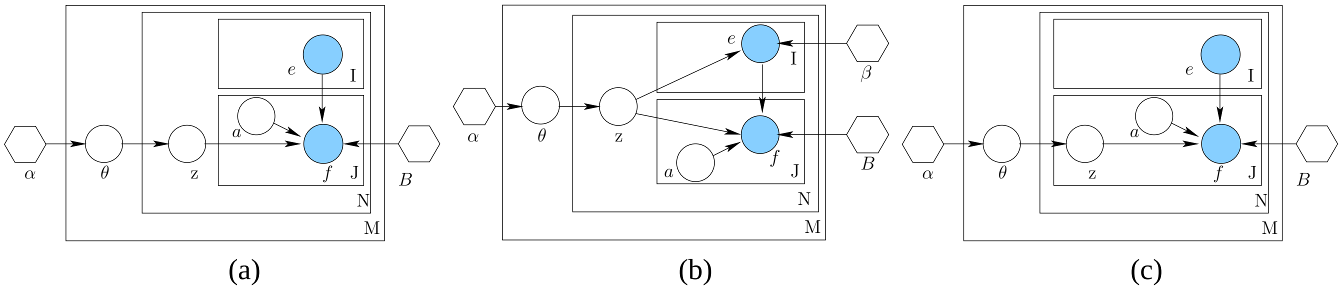

Figure 1: BiTAM models for Bilingual document- and sentence-pairs. A node in the graph represents a random variable, and a hexagon denotes a parameter. Un-shaded nodes are hidden variables. All the plates represent replicates. The outmost plate (M-plate) represents M bilingual document-pairs, while the inner N-plate represents the N repeated choice of topics for each sentence-pairs in the document; the inner J-plate represents J word-pairs within each sentence-pair. (a) BiTAM-1 samples one topic (denoted by z) per sentence-pair; (b) BiTAM-2 utilizes the sentence-level topics for both the translation model (i.e., $p(f|e,z)$) and the monolingual word distribution (i.e., $p(e|z)$); (c) BiTAM-3 samples one topic per word-pair.

a translation lexicon: $B_{i,j,k}{=}p(f{=}f_j|e{=}e_i, z{=}k)$, where z is an indicator variable to denote the choice of a topic. Given a specific *topic-weight vector* θ_d for a document-pair, each sentence-pair draws its conditionally independent topics from a mixture of topics. This generative process, for a document-pair $(\mathbf{F}_d, \mathbf{E}_d)$, is summarized as below:

1. Sample *sentence-number* N from a Poisson(γ).
2. Sample *topic-weight vector* θ_d from a Dirichlet(α).
3. For each sentence-pair $(\mathbf{f}_n, \mathbf{e}_n)$ in the $d'th$ doc-pair ,
   (a) Sample *sentence-length* J_n from Poisson(δ);
   (b) Sample a topic z_{dn} from a Multinomial(θ_d);
   (c) Sample e_j from a monolingual model $p(e_j)$;
   (d) Sample each word alignment link a_j from a uniform model $p(a_j)$ (or an HMM);
   (e) Sample each f_j according to a topic-specific translation lexicon $p(f_j|\mathbf{e}, a_j, z_n, \mathbf{B})$.

We assume that, in our model, there are K possible topics that a document-pair can bear. For each document-pair, a K-dimensional Dirichlet random variable θ_d, referred to as the topic-weight vector of the document, can take values in the $(K{-}1)$-simplex following a probability density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1}, \tag{3}$$

where the hyperparameter α is a K-dimension vector with each component $\alpha_k{>}0$, and $\Gamma(x)$ is the Gamma function. The alignment is represented by a J-dimension vector $\mathbf{a} = \{a_1, a_2, \cdots, a_J\}$; for each French word f_j at the position j, an position variable a_j maps it to an English word e_{a_j} at the position a_j in English sentence. The word level translation lexicon probabilities are topic-specific, and they are parameterized by the matrix $\mathbf{B} = \{B_k\}$.

For simplicity, in our current models we omit the modelings of the *sentence-number* N and the *sentence-length* J_n, and focus only on the bilingual translation model. Figure 1 (a) shows the graphical model representation for the BiTAM generative scheme discussed so far. Note that, the sentence-pairs are now connected by the node θ_d. Therefore, marginally, the sentence-pairs are *not* independent of each other as in traditional SMT models, instead they are *conditionally independent* given the topic-weight vector θ_d. Specifically, BiTAM-1 assumes that each sentence-pair has one single topic. Thus, the word-pairs within this sentence-pair are *conditionally independent* of each other given the hidden topic index z of the sentence-pair.

The last two sub-steps (3.d and 3.e) in the BiTam sampling scheme define a translation model, in which an alignment link a_j is proposed and an observation of f_j is generated according to the proposed distributions. We simplify *alignment model* of $\mathbf{a}$, as in *IBM-1*, by assuming that a_j is sampled uniformly at random. Given the parameters α, B, and the English part $\mathbf{E}$, the joint conditional distribution of the topic-weight vector θ, the topic indicators $\mathbf{z}$, the alignment vectors $\mathbf{A}$, and the document $\mathbf{F}$ can be written as:

$$\begin{aligned} &p(\mathbf{F}, \mathbf{A}, \theta, \mathbf{z}|\mathbf{E}, \alpha, \mathbf{B}) = \\ &\qquad p(\theta \mid \alpha)\prod_{n=1}^{N} p(z_n|\theta)p(\mathbf{f}_n, \mathbf{a}_n|\mathbf{e}_n,\ \alpha,\ B_{z_n}), \end{aligned} \tag{4}$$

where N is the number of the sentence-pair. Marginalizing out θ and $\mathbf{z}$, we can obtain the marginal conditional probability of generating $\mathbf{F}$ from $\mathbf{E}$ for each document-pair:

$$\begin{aligned} &p(\mathbf{F}, \mathbf{A}|\mathbf{E}, \alpha, B_{z_n}) = \\ &\qquad \int p(\theta|\alpha)\Big(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(\mathbf{f}_n, \mathbf{a}_n|\mathbf{e}_n, B_{z_n})\Big)d\theta, \end{aligned} \tag{5}$$

where $p(\mathbf{f}_n, \mathbf{a}_n|\mathbf{e}_n, B_{z_n})$ is a topic-specific sentence-level translation model. For simplicity, we assume that the French words f_j's are conditionally independent of each other; the alignment variables a_j's are independent of other variables and are uniformly distributed *a priori*. Therefore, the distribution for each sentence-pair is:

$$\begin{aligned} p(\mathbf{f}_n, \mathbf{a}_n | \mathbf{e}_n, B_{z_n}) &= p(\mathbf{f}_n | \mathbf{e}_n, \mathbf{a}_n, B_{z_n}) p(\mathbf{a}_n | \mathbf{e}_n, B_{z_n}) \\ &= \frac{1}{I_n^{J_n}} \prod_{j=1}^{J_n} p(f_{nj} | e_{a_{nj}}, B_{z_n}). \end{aligned} \tag{6}$$

Thus, the conditional likelihood for the entire parallel corpus is given by taking the product of the marginal probabilities of each individual document-pair in Eqn. 5.

3.2 BiTAM-2: Monolingual Admixture

In general, the monolingual model for English can also be a rich topic-mixture. This is realized by using the same topic-weight vector θ_d and the same topic indicator z_{dn} sampled according to θ_d, as described in §3.1, to introduce not only topic-dependent translation lexicon, but also topic-dependent monolingual model of the source language, English in this case, for generating each sentence-pair (Figure 1 (b)). Now $\mathbf{e}$ is generated from a topic-based language model β, instead of a uniform distribution in BiTAM-1. We refer to this model as BiTAM-2.

Unlike BiTAM-1, where the information observed in e_i is indirectly passed to z via the node of f_j and the hidden variable a_j, in BiTAM-2, the topics of corresponding English and French sentences are also strictly aligned so that the information observed in e_i can be directly passed to z, in the hope of finding more accurate topics. The topics are inferred more directly from the observed bilingual data, and as a result, improve alignment.

3.3 BiTAM-3: Word-level Admixture

It is straightforward to extend the sentence-level BiTAM-1 to a word-level admixture model, by sampling topic indicator $z_{n,j}$ for each word-pair (f_j, e_{a_j}) in the $n'th$ sentence-pair, rather than once for all (words) in the sentence (Figure 1 (c)). This gives rise to our BiTAM-3. The conditional likelihood functions can be obtained by extending the formulas in §3.1 to move the variable $z_{n,j}$ inside the same loop over each of the $f_{n,j}$.

3.4 Incorporation of Word "Null"

Similar to IBM models, "Null" word is used for the source words which have no translation counterparts in the target language. For example, Chinese words "de" (的) , "ba" (把) and "bei" (被) generally do not have translations in English. "Null" is attached to every target sentence to align the source words which miss their translations. Specifically, the latent Dirichlet allocation (LDA) in (Blei et al., 2003) can be viewed as a special case of the BiTAM-3, in which the target sentence contains only one word: "Null", and the alignment link a is no longer a hidden variable.

4 Learning and Inference

Due to the hybrid nature of the BiTAM models, exact posterior inference of the hidden variables $\mathbf{A}, \mathbf{z}$ and θ is intractable. A variational inference is used to approximate the true posteriors of these hidden variables. The inference scheme is presented for BiTAM-1; the algorithms for BiTAM-2 and BiTAM-3 are straight forward extensions and are omitted.

4.1 Variational Approximation

To approximate: $p(\theta, \mathbf{z}, \mathbf{A} | \mathbf{E}, \mathbf{F}, \alpha, B)$, the joint posterior, we use the fully factorized distribution over the same set of hidden variables:

$$\begin{aligned} & q(\theta, \mathbf{z}, \mathbf{A}) \propto q(\theta | \gamma, \alpha) \cdot \\ & \quad \prod_{n=1}^{N} q(z_n | \phi_n) \prod_{j=1}^{J_n} q(a_{nj}, f_{nj} | \varphi_{nj}, e_n, \mathbf{B}), \end{aligned} \tag{7}$$

where the Dirichlet parameter γ, the multinomial parameters $(\phi_1, \cdots, \phi_n)$, and the parameters $(\varphi_{n1}, \cdots, \varphi_{nJ_n})$ are known as variational parameters, and can be optimized with respect to the Kullback-Leibler divergence from $q(\cdot)$ to the original $p(\cdot)$ via an iterative fixed-point algorithm. It can be shown that the fixed-point equations for the variational parameters in BiTAM-1 are as follows:

$$\gamma_k = \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk} \tag{8}$$

$$\begin{aligned} \phi_{dnk} &\propto \exp\Big(\Psi(\gamma_k) - \Psi(\sum_{k'=1}^{K} \gamma_{k'})\Big) \cdot \\ & \qquad \exp\Big(\sum_{j=1}^{J_{dn}} \sum_{i=1}^{I_{dn}} \varphi_{dnji} \log B_{f_j, e_i, k}\Big) \end{aligned} \tag{9}$$

$$\varphi_{dnji} \propto \exp\Big(\sum_{k=1}^{K} \phi_{dnk} \log B_{f_j, e_i, k}\Big), \tag{10}$$

where $\Psi(\cdot)$ is a digamma function. Note that in the above formulas ϕ_{dnk} is the variational parameter underlying the topic indicator z_{dn} of the n-th sentence-pair in document d, and it can be used to predict the topic distribution of that sentence-pair.

Following a variational EM scheme (Beal and Ghahramani, 2002), we estimate the model parameters α and $\mathbf{B}$ in an unsupervised fashion. Essentially, Eqs. (8-10) above constitute the E-step, where the posterior estimations of the latent variables are obtained. In the M-step, we update α and $\mathbf{B}$ so that they improve a lower bound of the log-likelihood defined bellow:

$$\begin{aligned} L(\gamma,\phi,\varphi;\alpha,\mathbf{B}) = E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{z}|\theta)] \\ + E_q[\log p(\mathbf{a})] + E_q[\log p(\mathbf{f}|\mathbf{z},\mathbf{a},\mathbf{B})] - E_q[\log q(\theta)] \\ - E_q[\log q(\mathbf{z})] - E_q[\log q(\mathbf{a})]. \end{aligned} \quad (11)$$

The close-form iterative updating formula $\mathbf{B}$ is:

$$B_{f,e,k} \propto \sum_{d}^{M} \sum_{n=1}^{N_d} \sum_{j=1}^{J_{dn}} \sum_{i=1}^{I_{dn}} \delta(f,f_j)\delta(e,e_i)\phi_{dnk}\varphi_{dnji} \quad (12)$$

For α, close-form update is not available, and we resort to gradient accent as in (Sjölander et al., 1996) with re-starts to ensure each updated $\alpha_k > 0$.

4.2 Data Sparseness and Smoothing

The translation lexicons $B_{f,e,k}$ have a potential size of V^2K, assuming the vocabulary sizes for both languages are V. The data sparsity (i.e., lack of large volume of document-pairs) poses a more serious problem in estimating $B_{f,e,k}$ than the monolingual case, for instance, in (Blei et al., 2003). To reduce the data sparsity problem, we introduce two remedies in our models. First: *Laplace smoothing*. In this approach, the matrix set $\mathbf{B}$, whose columns correspond to parameters of conditional multinomial distributions, is treated as a collection of random vectors all under a symmetric Dirichlet prior; the posterior expectation of these multinomial parameter vectors can be estimated using Bayesian theory. Second: *interpolation smoothing*. Empirically, we can employ a linear interpolation with IBM-1 to avoid overfitting:

$$B^*_{f,e,k} = \lambda B_{f,e,k} + (1-\lambda)p(f|e). \quad (13)$$

As in Eqn. 1, $p(f|e)$ is learned via IBM-1; λ is estimated via EM on held out data.

4.3 Retrieving Word Alignments

Two word-alignment retrieval schemes are designed for BiTAMs: the *uni-direction* alignment (*UDA*) and the *bi-direction* alignment (*BDA*). Both use the posterior mean of the alignment indicators a_{dnji}, captured by what we call the *posterior alignment matrix* $\varphi \equiv \{\varphi_{dnji}\}$. UDA uses a French word f_{dnj} (at the $j'th$ position of $n'th$ sentence in the $d'th$ document) to query φ to get the best aligned English word (by taking the maximum point in a row of φ):

$$a_{dnj} = \arg\max_{i \in [1, I_{dn}]} \varphi_{dnji}. \quad (14)$$

BDA selects iteratively, for each f, the best aligned e, such that the word-pair (f,e) is the maximum of both row and column, or its neighbors have more aligned pairs than the other combpeting candidates.

A close check of $\{\varphi_{dnji}\}$ in Eqn. 10 reveals that it is essentially an exponential model: weighted log probabilities from individual topic-specific translation lexicons; or it can be viewed as weighted geometric mean of the individual lexicon's strength.

5 Experiments

We evaluate BiTAM models on the *word alignment accuracy* and the *translation quality*. For word alignment accuracy, *F-measure* is reported, i.e., the harmonic mean of precision and recall against a gold-standard reference set; for translation quality, *Bleu* (Papineni et al., 2002) and its variation of NIST scores are reported.

Table 1: Training and Test Data Statistics

Train	#Doc.	#Sent.	#Tokens	
			English	Chinese
Treebank	316	4172	133K	105K
FBIS.BJ	6,111	105K	4.18M	3.54M
Sinorama	2,373	103K	3.81M	3.60M
XinHua	19,140	115K	3.85M	3.93M
Test	95	627	25,500	19,726

We have two training data settings with different sizes (see Table 1). The small one consists of 316 document-pairs from Treebank (*LDC2002E17*). For the large training data setting, we collected additional document-pairs from FBIS (*LDC2003E14*, Beijing part), Sinorama (*LDC2002E58*), and Xinhua News (*LDC2002E18*, document boundaries are kept in our sentence-aligner (Zhao and Vogel, 2002)). There are 27,940 document-pairs, containing 327K sentence-pairs or 12 million (12M) English tokens and 11M Chinese tokens. To evaluate word alignment, we hand-labeled 627 sentence-pairs from 95 document-pairs sampled from TIDES'01 dryrun data. It contains 14,769 alignment-links. To evaluate translation quality, TIDES'02 Eval. test is used as development set, and TIDES'03 Eval. test is used as the unseen test data.

5.1 Model Settings

First, we explore the effects of Null word and smoothing strategies. Empirically, we find that adding "Null" word is always beneficial to all models regardless of number of topics selected.

Topics-Lexicons	Topic-1	Topic-2	Topic-3	Cooc.	IBM-1	HMM	IBM-4
p(ChaoXian (朝鲜)|Korean)	0.0612	0.2138	*0.2254*	38	0.2198	0.2157	0.2104
p(HanGuo (韩国)|Korean)	0.8379	0.6116	*0.0243*	46	0.5619	0.4723	0.4993

Table 2: Topic-specific translation lexicons are learned by a 3-topic BiTAM-1. The *third* lexicon (*Topic-3*) prefers to translate the word *Korean* into *ChaoXian* (朝鲜:North Korean). The co-occurrence (*Cooc*), IBM-1&4 and HMM only prefer to translate into *HanGuo* (韩国:South Korean). The two candidate translations may both fade out in the learned translation lexicons.

Unigram-rank	1	2	3	4	5	6	7	8	9	10
Topic A.	foreign	china	u.s.	development	trade	enterprises	technology	countries	year	economic
Topic B.	chongqing	companies	takeovers	company	city	billion	more	economic	reached	yuan
Topic C.	sports	disabled	team	people	cause	water	national	games	handicapped	members

Table 3: Three most distinctive topics are displayed. The English words for each topic are ranked according to $p(e|z)$ estimated from the topic-specific English sentences weighted by $\{\phi_{dnk}\}$. 33 functional words were removed to highlight the main content of each topic. Topic A is about Us-China economic relationships; Topic B relates to Chinese companies' merging; Topic C shows the sports of handicapped people.

The interpolation smoothing in §4.2 is effective, and it gives slightly better performance than Laplace smoothing over different number of topics for BiTAM-1. However, the interpolation leverages the competing baseline lexicon, and this can blur the evaluations of BiTAM's contributions. Laplace smoothing is chosen to emphasize more on BiTAM's strength. Without any smoothing, F-measure drops very quickly over two topics. In all our following experiments, we use both Null word and Laplace smoothing for the BiTAM models. We train, for comparison, IBM-1&4 and HMM models with 8 iterations of IBM-1, 7 for HMM and 3 for IBM-4 ($1^8h^74^3$) with Null word and a maximum fertility of 3 for Chinese-English.

Choosing the number of topics is a model selection problem. We performed a ten-fold cross-validation, and a setting of three-topic is chosen for both the small and the large training data sets. The overall computation complexity of the BiTAM is linear to the number of hidden topics.

5.2 Variational Inference

Under a non-symmetric Dirichlet prior, hyperparameter α is initialized randomly; $\mathbf{B}$ (K translation lexicons) are initialized uniformly as did in IBM-1. Better initialization of $\mathbf{B}$ can help to avoid local optimal as shown in § 5.5.

With the learned $\mathbf{B}$ and α fixed, the variational parameters to be computed in Eqn. (8-10) are initialized randomly; the fixed-point iterative updates stop when the change of the likelihood is smaller than 10^{-5}. The convergent variational parameters, corresponding to the highest likelihood from 20 random restarts, are used for retrieving the word alignment for unseen document-pairs. To estimate $\mathbf{B}$, β (for BiTAM-2) and α, at most eight variational EM iterations are run on the training data. Figure 2 shows absolute 2$\sim$3% better F-measure over iterations of variational EM using two and three topics of BiTAM-1 comparing with IBM-1.

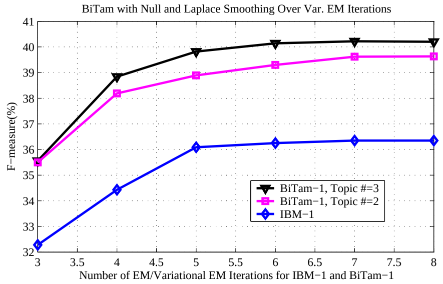

Figure 2: performances over eight Variational EM iterations of BiTAM-1 using both the "Null" word and the laplace smoothing; IBM-1 is shown over eight EM iterations for comparison.

5.3 Topic-Specific Translation Lexicons

The topic-specific lexicons B_k are smaller in size than IBM-1, and, typically, they contain topic trends. For example, in our training data, North *Korean* is usually related to *politics* and translated into "ChaoXian" (朝鲜); South *Korean* occurs more often with *economics* and is translated as "HanGuo"(韩国). BiTAMs discriminate the two by considering the topics of the context. Table 2 shows the lexicon entries for "*Korean*" learned by a 3-topic BiTAM-1. The values are relatively sharper, and each clearly favors one of the candidates. The co-occurrence count, however, only favors "HanGuo", and this can easily dominate the decisions of IBM and HMM models due to their ignorance of the topical context. Monolingual topics learned by BiTAMs are, roughly speaking, fuzzy especially when the number of topics is small. With proper filtering, we find that BiTAMs do capture some topics as illustrated in Table 3.

5.4 Evaluating Word Alignments

We evaluate word alignment accuracies in various settings. Notably, BiTAM allows to test alignments in two directions: English-to-Chinese (EC) and Chinese-to-English (CE). Additional *heuristics* are applied to further improve the accuracies. *Inter* takes the intersection of the two directions and generates high-precision alignments; the

SETTING	IBM-1	HMM	IBM-4	BITAM-1		BITAM-2		BITAM-3	
				UDA	BDA	UDA	BDA	UDA	BDA
CE (%)	36.27	43.00	45.00	40.13	48.26	40.26	48.63	40.47	49.02
EC (%)	32.94	44.26	45.96	36.52	46.61	37.35	46.30	37.54	46.62
REFINED (%)	**41.71**	44.40	48.42	**45.06**	**49.02**	**47.20**	47.61	**47.46**	48.18
UNION (%)	32.18	42.94	43.75	35.87	48.66	36.07	**48.99**	36.26	**49.35**
INTER (%)	39.86	**44.87**	**48.65**	43.65	43.85	44.91	45.18	45.13	45.48
NIST	6.458	6.822	6.926	6.937	6.954	6.904	6.976	6.967	6.962
BLEU	15.70	17.70	18.25	17.93	18.14	18.13	18.05	18.11	18.25

Table 4: Word Alignment Accuracy (F-measure) and Machine Translation Quality for BiTAM Models, comparing with IBM Models, and HMMs with a training scheme of $1^8h^74^3$ on the Treebank data listed in Table 1. For each column, the highlighted alignment (the best one under that model setting) is picked up to further evaluate the translation quality.

Union of two directions gives high-recall; *Refined* grows the intersection with the neighboring word-pairs seen in the union, and yields high-precision and high-recall alignments.

As shown in Table 4, the baseline IBM-1 gives its best performance of 36.27% in the CE direction; the UDA alignments from BiTAM-1∼3 give 40.13%, 40.26%, and 40.47%, respectively, which are significantly better than IBM-1. A close look at the three BiTAMs does not yield significant difference. BiTAM-3 is slightly better in most settings; BiTAM-1 is slightly worse than the other two, because the topics sampled at the sentence level are not very concentrated. The BDA alignments of BiTAM-1∼3 yield 48.26%, 48.63% and 49.02%, which are even better than HMM and IBM-4 — their best performances are at 44.26% and 45.96%, respectively. This is because BDA partially utilizes similar heuristics on the approximated posterior matrix $\{\varphi_{dnji}\}$ instead of direct operations on alignments of two directions in the heuristics of *Refined*. Practically, we also apply BDA together with heuristics for IBM-1, HMM and IBM-4, and the best achieved performances are at 40.56%, 46.52% and 49.18%, respectively. Overall, BiTAM models achieve performances close to or higher than HMM, using only a very simple IBM-1 style alignment model.

Similar improvements over IBM models and HMM are preserved after applying the three kinds of heuristics in the above. As expected, since BDA already encodes some heuristics, it is only slightly improved with the *Union* heuristic; UDA, similar to the viterbi style alignment in IBM and HMM, is improved better by the *Refined* heuristic.

We also test BiTAM-3 on large training data, and similar improvements are observed over those of the baseline models (see Table. 5).

5.5 Boosting BiTAM Models

The translation lexicons of $B_{f,e,k}$ are initialized uniformly in our previous experiments. Better initializations can potentially lead to better performances because it can help to avoid the undesirable local optima in variational EM iterations. We use the lexicons from IBM Model-4 to initialize $B_{f,e,k}$ to boost the BiTAM models. This is one way of applying the proposed BiTAM models into current state-of-the-art SMT systems for further improvement. The boosted alignments are denoted as BUDA and BBDA in Table. 5, corresponding to the uni-direction and bi-direction alignments, respectively. We see an improvement in alignment quality.

5.6 Evaluating Translations

To further evaluate our BiTAM models, word alignments are used in a phrase-based decoder for evaluating translation qualities. Similar to the Pharoah package (Koehn, 2004), we extract phrase-pairs directly from word alignment together with coherence constraints (Fox, 2002) to remove noisy ones. We use TIDES Eval'02 CE test set as development data to tune the decoder parameters; the Eval'03 data (919 sentences) is the unseen data. A trigram language model is built using 180 million English words. Across all the reported comparative settings, the key difference is the bilingual ngram-identity of the phrase-pair, which is collected directly from the underlying word alignment.

Shown in Table 4 are results for the small-data track; the large-data track results are in Table 5. For the small-data track, the baseline Bleu scores for IBM-1, HMM and IBM-4 are 15.70, 17.70 and 18.25, respectively. The UDA alignment of BiTAM-1 gives an improvement over the baseline IBM-1 from 15.70 to 17.93, and it is close to HMM's performance, even though BiTAM doesn't exploit any sequential structures of words. The proposed BiTAM-2 and BiTAM-3 are slightly better than BiTAM-1. Similar improvements are observed for the large-data track (see Table 5). Note that, the boosted BiTAM-3 us-

Setting	IBM-1	HMM	IBM-4	BiTAM-3			
				UDA	BDA	bUDA	bBDA
CE (%)	46.73	49.12	54.17	50.55	56.27	55.80	57.02
EC (%)	44.33	54.56	55.08	51.59	55.18	54.76	58.76
Refined (%)	**54.64**	**56.39**	**58.47**	**56.45**	54.57	**58.26**	56.23
Union (%)	42.47	51.59	52.67	50.23	**57.81**	56.19	**58.66**
Inter (%)	52.24	54.69	57.74	52.44	52.71	54.70	55.35
NIST	7.59	7.77	7.83	7.64	7.68	8.10	8.23
BLEU	19.19	21.99	23.18	21.20	21.43	22.97	24.07

Table 5: Evaluating Word Alignment Accuracies and Machine Translation Qualities for BiTAM Models, IBM Models, HMMs, and boosted BiTAMs using all the training data listed in Table. 1. Other experimental conditions are similar to Table. 4.

ing IBM-4 as the seed lexicon, outperform the *Refined* IBM-4: from 23.18 to 24.07 on Bleu score, and from 7.83 to 8.23 on NIST. This result suggests a straightforward way to leverage BiTAMs to improve statistical machine translations.

6 Conclusion

In this paper, we proposed novel formalism for statistical word alignment based on bilingual admixture (BiTAM) models. Three BiTAM models were proposed and evaluated on word alignment and translation qualities against state-of-the-art translation models. The proposed models significantly improve the alignment accuracy and lead to better translation qualities. Incorporation of within-sentence dependencies such as the alignment-jumps and distortions, and a better treatment of the source monolingual model worth further investigations.

References

M. J. Beal and Zoubin Ghahramani. 2002. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7*.

David Blei, Andrew NG, and M.I. Jordan. 2003. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 1107–1135.

P.F. Brown, Stephen A. Della Pietra, Vincent. J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19(2), pages 263–331.

Marine Carpua and Dekai Wu. 2005. Evaluating the word sense disambiguation performance of statistical machine translation. In *Second International Joint Conference on Natural Language Processing (IJCNLP-2005)*.

Bonnie Dorr and Nizar Habash. 2002. Interlingua approximation: A generation-heavy approach. In *In Proceedings of Workshop on Interlingua Reliability, Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, CA.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 304–311, Philadelphia, PA, July 6-7.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based smt. In *Proceedings of the Conference of the Association for Machine Translation in the Americans (AMTA)*.

Eugene A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles Involved in Bible Translating*. Leiden, Netherlands: E.J. Brill.

Eric Nyberg and Truko Mitamura. 1992. The kant system: Fast, accurate, high-quality translation in practical domains. In *Proceedings of COLING-92*.

Franz J. Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *HLT/NAACL: Human Language Technology Conference*, volume 1:29, pages 161–168.

Franz J. Och. 1999. An efficient method for determining bilingal word classes. In *Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics (EACL'99)*, pages 71–76.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

J. Pritchard, M. Stephens, and P. Donnell. 2000. Inference of population structure using multilocus genotype data. In *Genetics*, volume 155, pages 945–959.

K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler. 1996. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, 12.

S. Vogel, Hermann Ney, and C. Tillmann. 1996. Hmm based word alignment in statistical machine translation. In *Proc. The 16th Int. Conf. on Computational Lingustics, (Coling'96)*, pages 836–841, Copenhagen, Denmark.

Yeyi Wang, John Lafferty, and Alex Waibel. 1996. Word clustering with parallel spoken language corpora. In *proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, pages 2364–2367.

K. Yamada and Kevin. Knight. 2001. Syntax-based statistical translation model. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL-2001)*.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *The 2002 IEEE International Conference on Data Mining*.

Bing Zhao, Eric P. Xing, and Alex Waibel. 2005. Bilingual word spectral clustering for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.