Improving Statistical Machine Translation Using Bayesian Word Alignment and Gibbs Sampling

Coşkun Mermer, Murat Saraçlar, *Member, IEEE*, and Ruhi Sarikaya, *Senior Member, IEEE*

***Abstract*—We present a Bayesian approach to word alignment inference in IBM Models 1 and 2. In the original approach, word translation probabilities (i.e., model parameters) are estimated using the expectation-maximization (EM) algorithm. In the proposed approach, they are random variables with a prior and are integrated out during inference. We use Gibbs sampling to infer the word alignment posteriors. The inferred word alignments are compared against EM and variational Bayes (VB) inference in terms of their end-to-end translation performance on several language pairs and types of corpora up to 15 million sentence pairs. We show that Bayesian inference outperforms both EM and VB in the majority of test cases. Further analysis reveals that the proposed method effectively addresses the high-fertility rare word problem in EM and unaligned rare word problem in VB, achieves higher agreement and vocabulary coverage rates than both, and leads to smaller phrase tables.**

***Index Terms*—Bayesian methods, Gibbs sampling, statistical machine translation (SMT), word alignment.**

I. Introduction

WORD alignment is a crucial early step in the training pipeline of most statistical machine translation (SMT) systems [1]. Whether the employed models are phrase-based or tree-based, they use the estimated word alignments for constraining the set of candidates in phrase or grammar rule extraction [2]–[4]. As such, the coverage and the accuracy of the learned phrase/rule translation models are strongly correlated with those of the word alignment. Given a sentence-aligned parallel corpus, the goal of the word alignment is to identify the mapping between the source and target words in parallel sentences. Since word alignment information is usually not available during corpus generation and human annotation is costly, the task of word alignment is considered as an unsupervised learning problem.

Manuscript received June 04, 2012; revised November 05, 2012; accepted January 02, 2013. Date of publication February 01, 2013; date of current version February 25, 2013. The work of M. Saraclar was supported by a TÜBA-GEBİP award. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

C. Mermer is with TÜBİTAK BİLGEM, Kocaeli 41470, Turkey, and also with the Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul 34342, Turkey (e-mail: coskun.mermer@tubitak.gov.tr).

M. Saraçlar is with the Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul 34342, Turkey (e-mail: murat.saraclar@boun.edu.tr).

R. Sarikaya is with Microsoft, Redmond, WA 98052 USA (e-mail: ruhi.sarikaya@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2244087

State-of-the-art word alignment models, such as IBM Models [5], hidden Markov model (HMM) [6], and the jointly-trained symmetric HMM [7], contain a large number of parameters (such as word translation, transition, and fertility probabilities) that need be estimated in addition to the desired alignment variables. The common method of inference in such models is expectation-maximization (EM) [8] or an approximation to EM when exact EM is intractable. The EM algorithm finds the value of parameters that maximizes the likelihood of the observed variables. However, with many parameters to be estimated without any prior, EM tends to explain the training data by overfitting the parameters. A well-documented example of overfitting in EM-estimated word alignments is the case of rare words, where some rare words act as "garbage collectors" aligning to excessively many words on the other side of the sentence pair [9]–[11]. Moreover, EM is generally prone to getting stuck in a local maximum of the likelihood. Finally, EM is based on the assumption that there is one fixed value of parameters that explains the data, i.e., EM gives a point estimate.

We propose[1] a Bayesian approach in which we utilize a prior distribution on the parameters. The alignment probabilities are inferred by integrating over all possible parameter values. We treat the word translation probabilities as multinomial-distributed random variables with a sparse Dirichlet prior. Inference is performed via Gibbs sampling, which samples the posterior alignment distribution. We compare the EM and Bayesian alignments on the case of IBM Models 1 and 2. The inferred alignments are evaluated in terms of end-to-end translation performance on various language pairs and corpora.

The remainder of this paper is organized as follows: The related literature is reviewed in Section II. The proposed model and the inference algorithm are presented in Section III. The experiments are described and their results are presented in Section IV. A detailed analysis of the results and various aspects of the proposed method are provided in Section V, followed by the conclusions in Section VI.

II. Related Work

Problems with the standard EM estimation of IBM Model 1 were pointed out by Moore [11]. A number of heuristic changes to the estimation procedure, such as smoothing the parameter estimates, were shown to reduce the alignment error rate, but the effects on translation performance were not reported. Zhao and Xing [13] address the data sparsity issue using symmetric Dirichlet priors in parameter estimation and they use variational EM to find the maximum *a posteriori* (MAP) solution. Vaswani

[1]Part of this work was presented at a conference [12].

et al. [14] encourage sparsity in the translation model by placing an ℓ_0 prior on the parameters and then optimize for the MAP objective.

Zhao and Gildea [15] use sampling in their proposed fertility extensions to IBM Model 1 and HMM, but they do not place any prior on the parameters. Their inference method is stochastic EM (also known as Monte Carlo EM), a maximum-likelihood technique in which sampling is used to approximate the expected counts in the E-step. Even though they report substantial reductions in the alignment error rate, the translation performance measured in BLEU does not improve.

Bayesian modeling and inference have recently been applied to several unsupervised learning problems in natural language processing such as part-of-speech tagging [16], [17], word segmentation [18], [19], grammar extraction [20] and finite-state transducer training [21] as well as other tasks in SMT such as synchronous grammar induction [22] and learning phrase alignments directly [23].

Word alignment learning problem was addressed jointly with segmentation learning by Xu *et al.* [24], Nguyen *et al.* [25], and Chung and Gildea [26]. As in this paper, they treat word translation probabilities as random variables (with an associated prior distribution). Both [24] and [25] place *nonparametric* priors (also known as cache models) on the parameters. Similar to our work, this enables integration over the prior distribution. In [24], a Dirichlet Process prior is placed on IBM Model 1 word translation probabilities. In [25], a Pitman-Yor Process prior is placed on word translation probabilities in a proposed bag-of-words translation model that is similar to IBM Model 1. Both studies utilize Gibbs sampling for inference. However, alignment distributions are not sampled from the true posteriors but instead are updated either by running GIZA++ [24] or using a "local-best" maximization search [25]. On the other hand, a sparse Dirichlet prior on the multinomial parameters is used in [26] to prevent overfitting.

Bayesian word alignment with Dirichlet priors was also investigated in a recent study using variational Bayes (VB) [27]. VB is a Bayesian inference method which is sometimes preferred over Gibbs sampling due to its relatively lower computational cost and scalability. However, VB inference approximates the model by assuming independence between the hidden variables and the parameters. To evaluate the effect of this approximation, we also present and analyze the experimental results obtained using VB (Sections IV-C and V-A).

This paper extends the initial work in [12] in several aspects: 1) Extension to inference in Bayesian IBM Model 2, 2) complete derivation of the Gibbs sampler for both Models 1 and 2, 3) performance comparison with VB, 4) improved performance for both baseline and proposed systems via alignment combination, 5) reporting the average and standard deviation over 10 MERT runs for each BLEU score, 6) experimental results on two to three orders of magnitude larger training sets, 7) results with morphologically-segmented corpora, 8) several new metrics, including AER, for intrinsic and extrinsic evaluation of alignments obtained using different methods, 9) analysis of the effect of sampling settings, and 10) convergence behavior of both EM and Gibbs sampling.

III. Bayesian Inference of Word Alignments

We first recap the IBM Model 1 presented in [5] and establish the notation used in this paper. Given a parallel corpus $(\mathbf{E}, \mathbf{F})$ of S sentence pairs, let $\mathbf{e}(\mathbf{f})$ denote the s-th sentence in $\mathbf{E}(\mathbf{F})$, and let $e_i(f_j)$ denote the i-th (j-th) word among a total of $I(J)$ words in $\mathbf{e}(\mathbf{f})$[2]. We also hypothesize an imaginary "null" word e_0 to account for any unaligned words in $\mathbf{f}$. Also let V_E and V_F denote the size of the respective vocabularies.

We associate with each f_j a hidden *alignment* variable a_j whose value ranges over $[0, I]$. The set of alignments for a sentence (corpus) is denoted by $\mathbf{a}(\mathbf{A})$. The model parameters consist of a $V_E \times V_F$ table $\mathbf{T}$ of word translation probabilities such that $t_{e,f} = P(f|e)$. Since f is conditioned on e, we refer to $e(\mathbf{e})$ as the "source" word (sentence) and $f(\mathbf{f})$ as the "target" word (sentence)[3].

The conditional distribution of the Model 1 variables given parameters $\mathbf{T}$ is expressed by the following generative model:

$$a_j|\mathbf{e} \sim \text{Uniform}(a_j; I+1)$$

$$P(\mathbf{F}, \mathbf{A}|\mathbf{E}; \mathbf{T}) = \prod_s P(\mathbf{a}|\mathbf{e}) P(\mathbf{f}|\mathbf{a}, \mathbf{e}; \mathbf{T}) \quad (1)$$

$$= \prod_s \frac{1}{(I+1)^J} \prod_{j=1}^{J} t_{e_{a_j}, f_j}. \quad (2)$$

The two unknowns $\mathbf{A}$ and $\mathbf{T}$ are estimated using the EM algorithm, which finds the value of $\mathbf{T}$ that maximizes the likelihood of the observed variables $\mathbf{E}$ and $\mathbf{F}$ according to the model. Once the value of $\mathbf{T}$ is known, the probability of any alignment becomes straightforward to compute.

In the following derivation of our proposed model, we treat the unknown $\mathbf{T}$ as a random variable. Following the Bayesian approach, we assume a prior distribution on $\mathbf{T}$ and infer the distribution of $\mathbf{A}$ by integrating over all values of $\mathbf{T}$.

A. Canonical Representation of Model 1

We first convert the token-based expression in (2) into a type-based one as (with $\mathbf{T}$ now a random variable):

$$P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T}) = \prod_s \frac{1}{(I+1)^J} \prod_{e=1}^{V_E} \prod_{f=1}^{V_F} (t_{e,f})^{n_{e,f,s}} \quad (3)$$

$$= \prod_{e=1}^{V_E} \prod_{f=1}^{V_F} (t_{e,f})^{N_{e,f}} \cdot \prod_s \frac{1}{(I+1)^J}, \quad (4)$$

where in (3) the count variable $n_{e,f,s}$ denotes the number of times the source word type e is aligned to the target word type f in the sentence pair s, and in (4) $N_{e,f} = \sum_s n_{e,f,s}$.

This formulation exposes two properties of IBM Model 1 that facilitates the derivation of a Bayesian inference algorithm. First, the parametrization on $\mathbf{T}$ is in the canonical form of an

[2]Keeping in mind that $\mathbf{e}$, $\mathbf{f}$, I, J (and $\mathbf{a}$ introduced later) are defined with respect to the s-th sentence, we drop the subscript s for notational simplicity.

[3]Historically, the source and target designations were based on the translation task, when the word alignment direction was dictated by the "noisy channel model" to be the inverse of the translation direction. Today almost all SMT systems using IBM models train alignments in both directions, decoupling the alignment direction from that of translation and nullifying the justification of the early nomenclature.

exponential family distribution (as the inner-product of parameters $\log t_{e,f}$ and sufficient statistics $N_{e,f}$), which implies the existence of a *conjugate prior* that simplifies calculation of the posterior.

Second, the distribution in (4) depends on the variables $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{A}$ only through a set of count variables $N_{e,f}$. In other words, the order of words within a sentence has no effect on the likelihood, which is called *exchangeability* or a "bag of words" model. This results in simplification of the terms when deriving the Gibbs sampler.

B. Prior on Word Translation Probabilities

For each source word type e, by definition $\mathbf{t}_e = t_{e,1} \cdots t_{e,V_F}$ form the parameters of a multinomial distribution that governs the distribution of the target words aligned to e. Hence, the conditional distribution of the j-th target word in a sentence pair is defined by:

$$f_j|\mathbf{a}, \mathbf{e}, \mathbf{T} \sim \text{Multinomial}\left(f_j; \mathbf{t}_{e_{a_j}}\right).$$

Since the conjugate prior of multinomial is the Dirichlet distribution, we choose:

$$\mathbf{t}_e|\boldsymbol{\Theta} \sim \text{Dirichlet}(\mathbf{t}_e; \boldsymbol{\Theta}_e),$$

where $\boldsymbol{\Theta}_e = \theta_{e,1} \cdots \theta_{e,V_F}$. Overall, $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \cdots \boldsymbol{\Theta}_{V_E}$ are the hyperparameters of the model. The mathematical expression for the prior $P(\mathbf{T}; \boldsymbol{\Theta})$ is provided in (13) in the Appendix.

We can encode our prior expectations for $\mathbf{t}_e$ into the model by suitably setting the values of $\boldsymbol{\Theta}_e$. For example, we generally expect the translation probability distribution of a given source word type e to be concentrated on one or a few target word types. Setting $\theta_{e,f} \ll 1, \forall f$ allocates more prior weight to such sparse distributions.

C. Inference by Gibbs Sampling

To infer the posterior distribution of the alignments $P(\mathbf{A}|\mathbf{E}, \mathbf{F}; \boldsymbol{\Theta})$, we use Gibbs sampling [28], a stochastic inference technique that produces random samples that converge in distribution to the desired posterior. In general, for a set of random variables $\mathbf{z} = \{z_j\}$, a Gibbs sampler iteratively updates the variables z_j one at a time by sampling its value from the distribution $P(z_j|\mathbf{z}^{\neg j})$, where the superscript $\neg j$ denotes the exclusion of the variable being sampled.

Before applying Gibbs sampling to our model in (4), since we are only after $\mathbf{A}$, we integrate out the unknown $\mathbf{T}$ using:

$$P(\mathbf{F}, \mathbf{A}|\mathbf{E}; \boldsymbol{\Theta}) = \int_{\mathbf{T}} P(\mathbf{T}; \boldsymbol{\Theta}) P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T}). \quad (5)$$

The remaining set of variables is $\mathbf{z} = \{\mathbf{E}, \mathbf{F}, \mathbf{A}\}$, of which only $\mathbf{A}$ is unknown.

Starting from (5), the Gibbs sampling formula is found as (the derivation steps are outlined in the Appendix):

$$P(a_j = i|\mathbf{E}, \mathbf{F}, \mathbf{A}^{\neg j}; \boldsymbol{\Theta}) \propto \frac{N^{\neg j}_{e_i,f_j} + \theta_{e_i,f_j}}{\sum_{f=1}^{V_F} N^{\neg j}_{e_i,f} + \sum_{f=1}^{V_F} \theta_{e_i,f}}. \quad (6)$$

Here, $N^{\neg j}_{e_i,f_j}$ denotes the number of times the source word type e_i is aligned to the target word type f_j in $\mathbf{A}$, not counting the current alignment link between f_j and e_{a_j}. We can also observe the effect of the prior, where the hyperparameters act as *pseudo-counts* added to N_{e_i,f_j}. Table I describes the complete inference algorithm. In Step 1, $\mathbf{A}$ can be initialized arbitrarily. However, informed initializations, e.g., EM-estimated alignments, can be used for faster convergence. Once the Gibbs sampler is deemed to have converged after B burn-in iterations, we collect M samples of $\mathbf{A}$ to estimate the underlying distribution $P(\mathbf{A}|\mathbf{E}, \mathbf{F})$. To reduce correlation between these M samples, a lag of L iterations is introduced in-between. Thus the algorithm is run for a total of $K = B + M \times L$ iterations.

TABLE I
ALIGNMENT INFERENCE ALGORITHM FOR BAYESIAN IBM MODEL 1 USING GIBBS SAMPLING

```
Input: E, F;  Output: K samples of A
1    Initialize A
2    for k = 1 to K do
3        for each sentence pair s in (E, F) do
4            for j = 1 to J do
5                for i = 0 to I do
6                    Calculate P(a_j = i|...) according to (6)
7                Sample a new value for a_j
```

The phrase/rule extraction step requires as its input the most probable alignment $\mathbf{A}^* = \arg\max_{\mathbf{A}} P(\mathbf{A}|\mathbf{E}, \mathbf{F})$, which is also called the *Viterbi* alignment. Since $\mathbf{A}$ is a vector with a large number of elements, we make the assumption that the most frequent value for the vector $\mathbf{A}$ can be approximated by the vector consisting of the most frequent values for each element a_j. Hence, we select for each a_j its most frequent value in the M collected samples as the Viterbi alignment.

D. Extension to IBM Model 2

IBM Model 1 assumes that all alignments are equally probable, i.e., $P(a_j = i) = (I + 1)^{-1}$. In IBM Model 2 [5], the alignment probability distribution $P(a_j)$ for a given target word at position j depends on the quadruple (i, j, I, J). This dependency is parametrized by a distortion parameter d for each quadruple such that

$$P(a_j = i|j, I, J) = d_{i,j,I,J}. \quad (7)$$

Note that Model 1 is a special case of Model 2 in which the parameters $d_{i,j,I,J}$ are fixed at $(I + 1)^{-1}$.

Different variants of Model 2 have been proposed to reduce the number of parameters, e.g., by dropping dependence on J ($d_{i,j,I}$ [10]) or using relative distortion (d_r where $r = i - \lfloor j(I/J) \rfloor$ [6], also called "diagonal-oriented Model 2" [29]). In the following, we used the latter parametrization; the derivation for inference in the other variants would be similar.

Bayesian inference in Model 2 can be derived in an analogous manner to Model 1. Treating the set of distortion parameters, denoted by $\mathbf{d} = d_{-\max_s I} \cdots d_{\max_s I}$, as a new random variable, equations (2) and (4) can be adapted to Model 2 as:

$$P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T}, \mathbf{d}) = \prod_s \prod_{j=1}^{J} \left(t_{e_{a_j},f_j} \cdot d_{a_j - \lfloor j\frac{I}{J} \rfloor}\right) \quad (8)$$

$$= \prod_{e=1}^{V_E} \prod_{f=1}^{V_F} (t_{e,f})^{N_{e,f}} \cdot \prod_{r=-\max_s I}^{\max_s I} (d_r)^{C_r}, \quad (9)$$

where in (9) the count variable C_r stores the number of times a particular relative distortion r occurs in $\mathbf{A}$.

Since $\mathbf{d}$ form the parameters of a multinomial distribution on a_j (see (7)), we choose a Dirichlet prior on $\mathbf{d}$:

$$a_j|\mathbf{d} \sim \text{Multinomial}(a_j; \mathbf{d})$$
$$\mathbf{d}|\mathbf{\Phi} \sim \text{Dirichlet}(\mathbf{d}; \mathbf{\Phi}),$$

where $\mathbf{\Phi} = \phi_{-\max_s I} \cdots \phi_{\max_s I}$ are the distortion hyperparameters. Integrating out the parameters $\mathbf{T}$ and $\mathbf{d}$ results in the following Gibbs sampling formula for Bayesian IBM Model 2:

$$P(a_j = i|\mathbf{E}, \mathbf{F}, \mathbf{A}^{\neg j}; \mathbf{\Theta}, \mathbf{\Phi}) \propto \frac{N^{\neg j}_{e_i,f_j} + \theta_{e_i,f_j}}{\sum_{f=1}^{V_F} N^{\neg j}_{e_i,f} + \sum_{f=1}^{V_F} \theta_{e_i,f}} \cdot (C^{\neg j}_r + \phi_r), \quad (10)$$

where $r = i - \lfloor j(I/J) \rfloor$. A complete derivation is presented in the Appendix. To infer the alignments under Model 2, the only change needed in Table I is the use of (10) instead of (6) in step 6.

IV. Experimental Results

A. Setup

We evaluated the performance of the Bayesian word alignment via bi-directional translation experiments. We performed the initial experiments and analyses on small data, then tested the best performing baseline and proposed methods on large data. Furthermore, we performed some of the side investigations and compute-intensive experiments such as those concerning the alignment combination schemes, morphological segmentation, convergence and the effect of sampling settings only on the smallest of the datasets (Turkish $\leftrightarrow$ English).

For Turkish $\leftrightarrow$ English ($\text{T} \leftrightarrow \text{E}$) experiments, we used the travel domain BTEC dataset [30] from the annual IWSLT evaluations [31] for training, the CSTAR 2003 test set for tuning, and the IWSLT 2004 test set for testing. For Arabic $\leftrightarrow$ English ($\text{A} \leftrightarrow \text{E}$), we used LDC2004T18 (news from years 2001-2004) for training, subsets of the AFP portion of LDC2004T17 (news from year 1998) for tuning and testing, and the AFP and Xinhua subsets of the respective Gigaword corpora (LDC2007T07 and LDC2007T40) for additional LM training. We filtered out sentence pairs where either side contains more than 70 words for Arabic $\leftrightarrow$ English. All language models are 4-gram in the travel domain experiments and 5-gram in the news domain experiments with modified Kneser-Ney smoothing [32] and interpolation. Table II shows the statistics of the data sets used in the small-data experiments.

For each language pair, we obtained maximum-likelihood word alignments using the EM implementation of GIZA++ [10] and Bayesian alignments using the publicly available Gibbs sampling (GS) implementation [33]. As sampling settings (Section III-C), we used $M = 100$; $L = 10$; and $B = 400$ for $\text{T} \leftrightarrow \text{E}$ and 8000 for $\text{A} \leftrightarrow \text{E}$. We chose identical symmetric Dirichlet priors for all source words e with $\theta_{e,f} = \theta = 0.0001$ to obtain a sparse Dirichlet prior.

After alignments were obtained in both translation directions, standard phrase-based SMT systems were trained in both directions using Moses [34], SRILM [35], and ZMERT [36] tools. The translations were evaluated using the single-reference BLEU [37] metric. Alignments in both directions were symmetrized using the default heuristic in Moses ("grow-diag-final-and"). To account for the random variability in minimum error-rate training (MERT) [38], we report the mean and standard deviation of 10 MERT runs for each evaluation.

TABLE II
Corpus Statistics for Each Language Pair in the Small-Data Experiments. T: Turkish, E: English, A: Arabic

	T / E	A / E
Training set:		
Sentences	20k	56k
Tokens	140k / 183k	1.5M / 1.8M
Tokens/sentence	7.0 / 9.1	27 / 33
Types	18k / 7.3k	80k / 35k
Singletons	10k / 3.2k	35k / 14k
Additional LM tokens	-	215M / 298M
Tuning set sentences	506	873
Test set sentences	500	879

We also investigated alignment combination, both within and across alignment methods, to obtain the best possible performance. For this purpose, we obtained three alignment samples from each inference method while trying to capture as much diversity as possible. For EM, we obtained alignments after 5, 20, and 80 iterations (denoted by EM-5, EM-20, and EM-80, respectively). For GS, we ran three separate chains, two initialized with the EM alignments (denoted by GS-5 and GS-80, respectively), and to provide even more diversity, a third initialized based on co-occurrence (denoted by GS-N): Each target word was initially aligned to the source candidate that it co-occurred with the most number of times in the entire parallel corpus.

B. Performance Comparison of EM and GS

Fig. 1 compares the BLEU scores of SMT systems trained with individual EM- and GS-inferred alignments. In all cases, using GS alignments that are initialized with the alignments from EM leads to higher BLEU scores on average than using the EM alignments directly. In Section V-A, we investigate the intrinsic differences between the EM- and GS-inferred alignments that lead to the improved translation performance.

Alignment combination across methods (heterogeneous combination) has been previously shown [39], [40] to improve the translation performance over individual alignments. Moreover, alignment combination within a method (homogeneous combination) can also cope with random variation (in GS) or overfitting (in EM).

We implemented alignment combination by concatenating the individual sets of alignments, meanwhile replicating the training corpus, and training the SMT system otherwise the same way. We experimented with various alignment combination schemes and found that combining the EM alignments from 5, 20, and 80 iterations is in general better than the individual alignments, with a similar conclusion for combining the three GS alignments described in Section IV-A. Further combination of these two combinations for a total of six alignments sometimes improved the performance even more. So we present the results in this section using these three combination

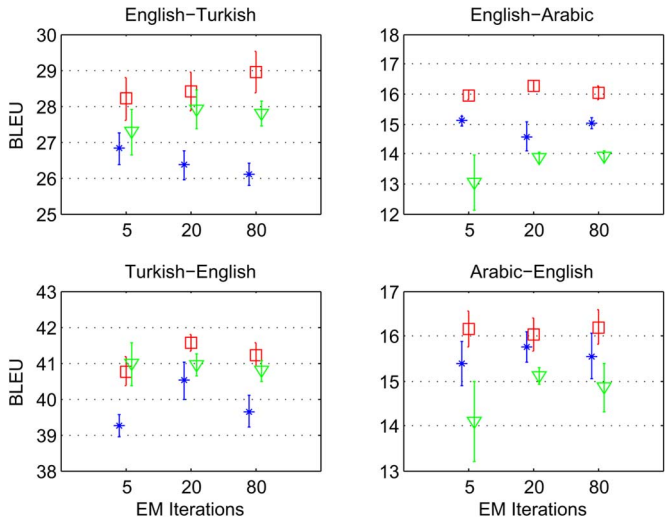

Fig. 1. Translation performance of word alignments obtained by expectation-maximization (EM), Gibbs sampling initialized with EM (GS) and variational Bayes (VB): ∗ EM, □ GS, ∇ VB.

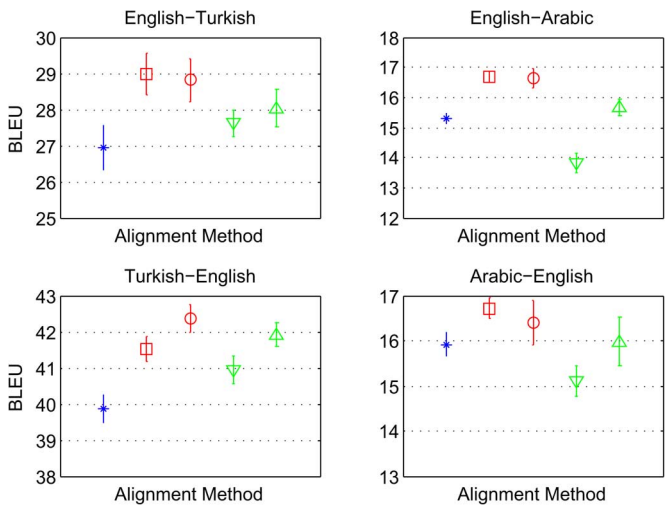

Fig. 2. Translation performance of EM, Gibbs sampling, and variational Bayes after applying alignment combination within and across methods: ∗ EM(Co), □ GS(Co), ∘ EM(Co)+GS(Co), ∇ VB(Co), and △ EM(Co)+VB(Co). The same BLEU scale is used as in Fig. 1.

schemes (denoted by EM(Co), GS(Co), and EM(Co)+GS(Co), respectively, in Fig. 2).

We observe from Fig. 2 that GS(Co) outperforms EM(Co) on average, both by itself and in combination with EM(Co), in most cases by a significant margin. However, which scheme (GS(Co) or EM(Co)+GS(Co)) is the best seems to depend on the language pair and/or dataset.

C. Comparison with Variational Bayes

Using the publicly available software [41], we experimented with variational Bayes (VB) inference using similar alignment combination schemes: combination of three VB-inferred alignments after 5, 20, and 80 Model 1 iterations; and further combination of it with the three EM-inferred alignments above (denoted by VB(Co) and EM(Co)+VB(Co), respectively).

The translation performance of the individual VB alignments in Fig. 1 shows that, compared to EM, VB achieves higher

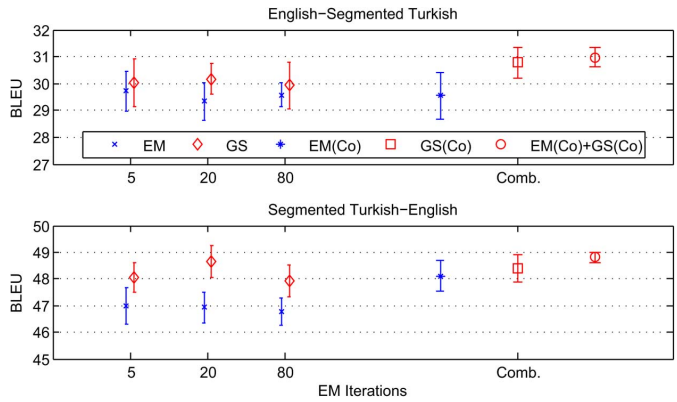

Fig. 3. Results for the morphologically-segmented Turkish-English corpus. All BLEU scores are computed at the word level.

BLEU scores in T ↔ E but lower scores in A ↔ E. On the other hand, GS outperforms VB in all cases but one in Fig. 1. As for the performance after alignment combination, Fig. 2 shows that, for all translation directions GS(Co) leads to higher average BLEU scores compared to VB(Co), both with and without further combination with EM(Co). The performance of VB(Co) relative to EM(Co) is similar to the case for individual alignments (better in T ↔ E, worse in A ↔ E). However, EM(Co)+VB(Co) outperforms or performs as good as EM in all cases, demonstrating that Bayesian word alignment can be beneficial even with a fast, yet approximate inference method.

To explain the particularly low performance of VB in Arabic ↔ English, we inspected the alignments inferred by EM, GS, and VB. We found that while VB with sparse Dirichlet prior avoids excessive alignment fertilities, it leaves many rare source words unaligned. For example, the percentage of unaligned source singletons for EM-5, GS-5, and VB-5 in the English → Arabic (Arabic → English) alignments are 27%, 16%, and 69% (44%, 34%, and 71%), respectively. We believe the higher rate of unaligned singletons can lead to poorer training set coverage and lower translation performance (Section V-A).

D. Experiments With Morphologically Segmented Corpus

Morphological preprocessing is a common practice in modern SMT systems dealing with morphologically unmatched language pairs. Thus, as a side investigation, we also experimented with morphological segmentation in the T ↔ E corpus to see its effect on the performance of our proposed method (morphological segmentation is also applied in the large-data A ↔ E experiments presented in Section IV-E). We used the morphological analyzer by Oflazer [42] to segment the Turkish words into lexical morphemes. As a result, the vocabulary size decreased to 5.6k (from 18k, cf. Table II), with 2.4k of them singletons. The out-of-vocabulary rate in the Turkish tuning and test sets decreased from 5.2% and 6.1% to 0.9% and 0.8%, respectively. The BLEU scores were still computed at the word level in the case of English → Turkish translation by joining the morphemes in the output.

The results in Fig. 3 show that the advantage of GS over EM still holds in the morphologically-segmented condition in both translation directions, both individually and with combination. In addition, comparing the BLEU scores with those in Figs. 1

TABLE III
CORPUS STATISTICS FOR EACH LANGUAGE PAIR IN THE LARGE-DATA EXPERIMENTS. A: ARABIC, E: ENGLISH, C: CZECH, G: GERMAN

	A / E	C / E	G / E
Training:			
Sentences	7.6M	15.4M	2.0M
Tokens	202M / 203M	203M / 230M	50M / 53M
Types	355k / 342k	1.53M / 1.00M	420k / 139k
LM tokens	- / 241M	265M / 1.05G	477M / 1.05G
Tuning sentences	1000	3003	3003
Test sentences	2000	3003	3003

and 2 confirms the previous studies that applying morphological segmentation improves the translation performance significantly, especially in the morphologically poorer direction (i.e., Turkish → English).

E. Experiments on Larger Datasets

The scalability of the alignment inference methods was also tested on publicly available large datasets (Table III). We used the 8-million sentence Multi-UN corpus [43] for Arabic→English translation experiments. As is common in most state-of-the-art systems for this language pair, we performed morphological segmentation on the Arabic side for the best performance (we used the MADA+TOKAN tool [44]). Note that after morphological segmentation, Arabic no longer exhibits the vocabulary characteristics of a morphologically-rich language (Table III). We set aside the last 100k sentences of the corpus and randomly extracted the tuning and test sets from this subset. The English side of the parallel corpus was used for language model training.

We used the WMT 2012 [45] datasets for Czech ↔ English (C ↔ E) and German ↔ English (G ↔ E) translation experiments. The C ↔ E training data consisted of the Europarl, news commentary, and the 15-million sentence CzEng 1.0 [46] corpora while the G ↔ E training data consisted of only the Europarl and news commentary corpora. WMT 2011 and 2012 news testsets were used for tuning and testing, respectively. The WMT 2012 monolingual news corpora covering years 2007–2011 were used for language model training.

In all large-data experiments, sentences longer than 70 words were excluded from translation model training. Gibbs sampling settings of (B, M, L)=(1000, 100, 1) were used. All language models were 4-gram. To obtain the best possible baseline, we also utilized techniques that we had previously observed to improve performance on similar corpora, such as lattice sampling [47] and search in random directions [48] during MERT and minimum Bayes risk decoding [49]. All other experimental settings (e.g., 10 MERT runs etc.) were identical to the small-data experiments (Section IV-A).

To conform with the majority of previous research and evaluations in these language pairs, we trained SMT systems in both directions for the WMT 2012 language pairs and in the Arabic → English direction for the Multi-UN task. For the two largest datasets (C ↔ E and A → E), we also experimented with 1-million sentence versions for faster development experiments and to provide an intermediate data size setting.

The results are presented in Figs. 4–6. For translation *to* English, Gibbs sampling improves over EM for all five corpora, the largest improvement achieved by GS(Co)+EM(Co) in A → E (0.5 to 0.7 BLEU mean difference) and by GS(Co) in C → E and G → E (0.3 to 0.5 BLUE mean difference). However, for translation *from* English (E → C and E → G), we do not observe a consistent improvement over EM.

For the 1-million sentence A → E task, we also report the translation error rates (TERs) [50] (bottom row of Fig. 4). Except for the comparison between GS(Co) and EM(Co)+GS(Co) in the 1M-sentence setting, in all possible pair-wise comparisons between the alignment methods in both corpus settings,

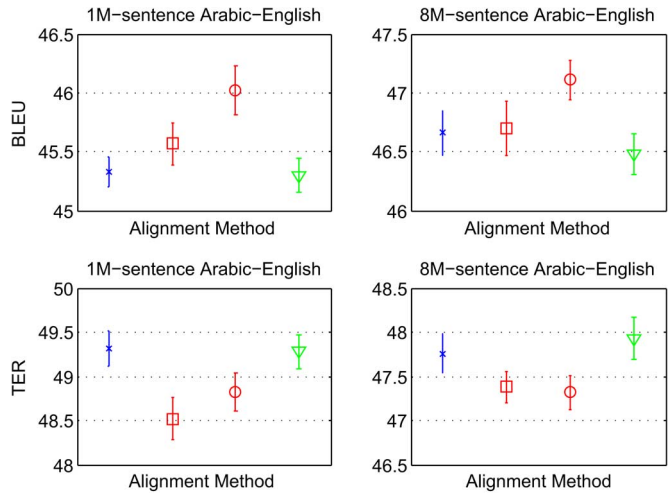

Fig. 4. Arabic → English BLEU and TER scores of various alignment methods: ∗ EM(Co), □ GS(Co), ∘ EM(Co)+GS(Co), and ∇ VB(Co).

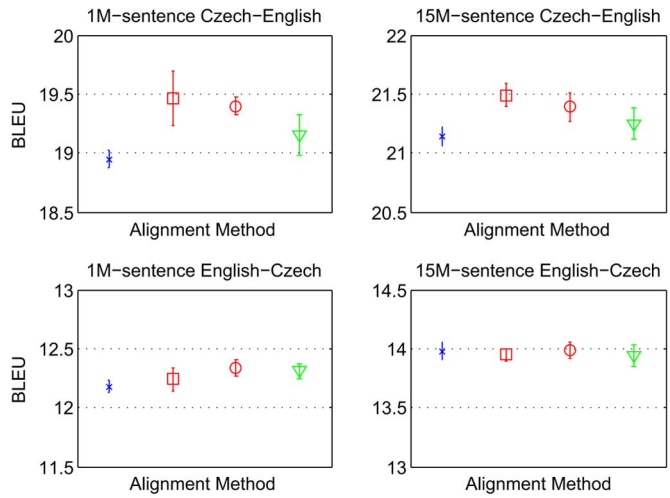

Fig. 5. Czech ↔ English BLEU scores of various alignment methods: ∗ EM(Co), □ GS(Co), ∘ EM(Co)+GS(Co), and ∇ VB(Co).

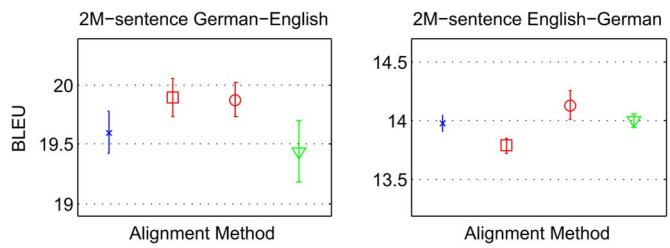

Fig. 6. German ↔ English BLEU scores of various alignment methods: ∗ EM(Co), □ GS(Co), ∘ EM(Co)+GS(Co), and ∇ VB(Co).

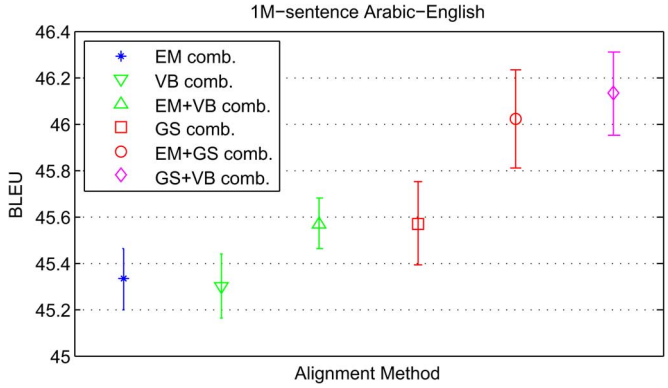

Fig. 7. Arabic → English BLEU scores of various alignment combination schemes in the 1M-sentence translation task.

TABLE IV
BLEU SCORES OF IBM MODEL 2 ALIGNMENT INFERENCE METHODS ON THE 1M-SENTENCE ARABIC → ENGLISH TRANSLATION

Method	Model 2 EM	Model 2 GS
BLEU	46.97 ±*0.15*	47.17 ±*0.14*

the method with the higher mean BLEU score also has the lower mean TER score[4].

In addition, we compared the performance of some of the many possible alignment combination schemes (Fig. 7). Not surprisingly, combination with EM(Co) helps both GS(Co) and VB(Co), and the relative ranking of the latter two does not change after combination with EM(Co). Furthermore, combination of GS(Co)+VB(Co) improves the performance slightly over EM(Co)+GS(Co).

F. Bayesian Model 2 Results

We tested the IBM Model 2 Gibbs sampling algorithm on the 1M-sentence subset of the Arabic-English Multi-UN corpus. Unlike the case of translation parameters $\mathbf{T}$, there is no clear language- and domain-independent knowledge of how the distortion parameters $\mathbf{d}$ (the distribution of a_j) should look like. Therefore, we assumed that all distortion distributions are *a priori* equally probable, which corresponds to setting the distortion hyperparameters $\phi_r = 1$ for all r. We also collapsed the counts for distortions larger in magnitude than 5, resulting in 11 total distortion count variables $N_{r\leq -5}, N_{-4}, \cdots, N_4, N_{r\geq 5}$, as done in [7].

We compared the translation performance of the EM- and GS-inferred Model 2 alignments. Both methods are initialized with the same EM-5 alignments (i.e., 5 iterations of Model 1 EM). Model 2 EM is run for 5 iterations. Model 2 GS is estimated with $B = 1000$, $M = 100$ and $L = 1$. The results are shown in Table IV. Bayesian inference improves the mean BLUE score by 0.2 BLEU. Further improvement could be possible by alignment combination within and across methods, as done in Section IV-B.

V. ANALYSIS AND DISCUSSION

A. Analysis of Inferred Alignments

In order to explain the BLEU score improvements achieved by the Bayesian alignment approach and to characterize the differences between the alignments obtained by various methods, we analyzed the alignments in Fig. 1 using several intrinsic and extrinsic evaluation metrics. As representative alignments from each method, we selected EM-5, VB-5, and GS-5.

[4]BLEU was used as the error metric for optimization in MERT.

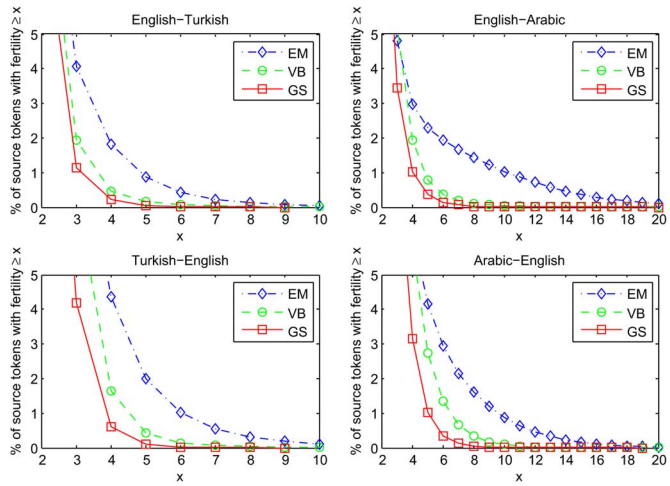

Fig. 8. Distribution of alignment fertilities for source language tokens.

1) Fertility Distributions: Fertility of a source word is defined as the number of target words aligned to it. In general, we expect the fertility values close to the word token ratio between the languages to be the most frequent and high fertility values to be rare. Fig. 8 shows the fertility distributions in alignments obtained from different methods. We can observe the "garbage collecting" effect in the long tails of the EM-estimated alignments. For example, in English-Arabic Model 1 alignment using EM, 1.2% of the English source tokens are aligned with *nine or more* Arabic target words, corresponding to 22.3k total occurrences or about 0.4 occurrence per sentence. In all alignment tasks, both Bayesian methods result in fewer high-fertility alignments compared to EM. Among Bayesian inference techniques, GS is more effective than VB in avoiding high fertilities.

2) Alignment Dictionary Size: Reducing the number of unique alignment pairs has been proposed as an objective for word alignment [51], [52]: it was observed during manual alignment experiments that humans try to find the alignment with the most compact "alignment dictionary" (a vocabulary of unique source-target word pairs) as possible. Fig. 9(a) shows that both GS and VB explain the training data using a significantly smaller alignment-pair vocabulary compared to EM.

3) Singleton Fertilities: The average alignment fertility of source singletons was proposed as an intrinsic evaluation metric in [40]. We expect lower values to correlate with better alignments. However, a value of zero could be achieved by leaving all singletons unaligned, which is clearly not desirable. Therefore, we refine the definition of this metric to calculate the average over *aligned* singletons only. The minimum value thus attainable is one. Fig. 9(b) shows that both Bayesian methods significantly reduce singleton fertilities.

The average fertility of aligned singletons by itself is not sufficient to accurately assess an alignment since unaligned singletons are not represented. Hence, we also report the percentage of unaligned singletons in Fig. 9(c). GS has the lowest unaligned singleton rate among Model 1 inference methods. An interesting

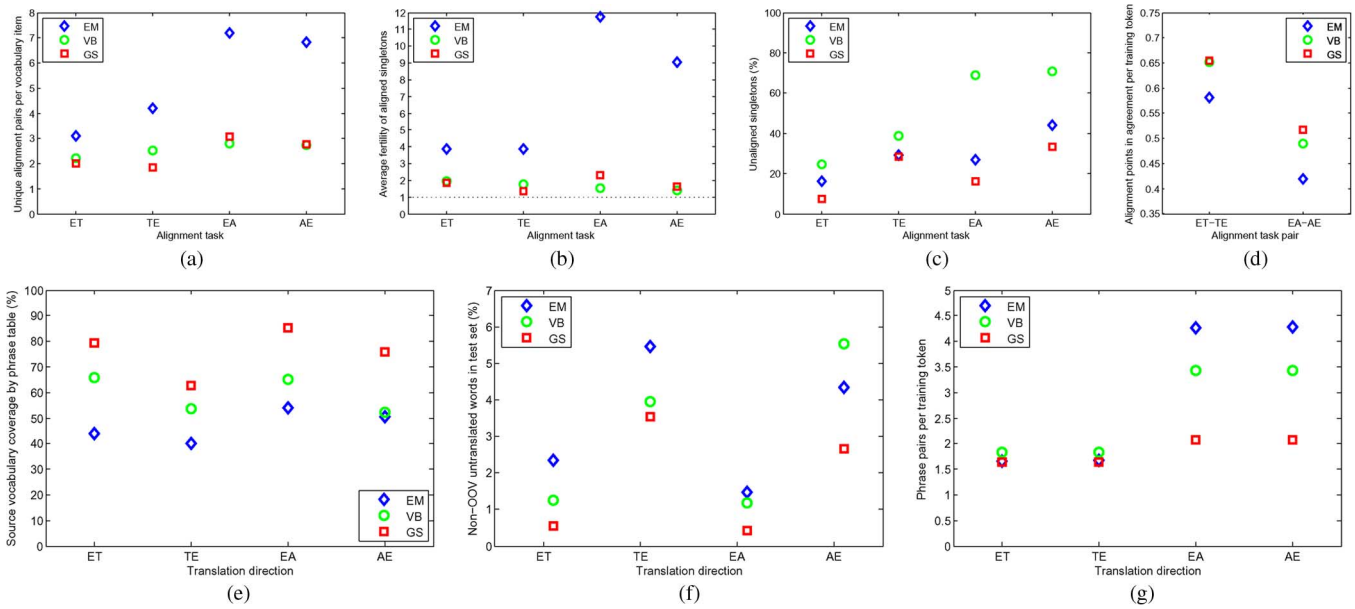

Fig. 9. Intrinsic and extrinsic evaluation of alignments in the small data experiments. (a) Alignment dictionary size normalized by the average of source and target vocabulary sizes. (b) Average alignment fertility of aligned singletons. (c) Percentage of unaligned singletons. (d) Number of symmetric alignments normalized by the average of source and target tokens. (e) Percentage of training set vocabulary covered by single-word phrases in the phrase table. (f) Decode-time rate of input words that are in the training vocabulary but without a translation in the phrase table. (g) Phrase table size normalized by the average of source and target tokens.

observation is that, while EM-estimated alignments suffer from rare words being assigned high fertilities (Fig. 9(b)), VB suffers from a high percentage of the rare words (e.g., about 70% of singletons in A $\leftrightarrow$ E) being left unaligned, resulting in lower translation performance (Section IV-C). Our analysis agrees with the findings of Guzman *et al.* [53] that unaligned words in an alignment results in lower-quality phrase tables.

4) Alignment Points in Agreement: Since the IBM alignment models are one(source)-to-many(target), switching the source and target languages usually result in a different set of alignment links (or points in an alignment matrix). The intersection of the two sets consists of high-precision alignment points where both alignment models agree [7]. Since the number of alignment points in each direction is constant (equal to the number of target words), increasing precision at the expense of recall by predicting fewer alignment points is not applicable in these models. Therefore higher agreement rate implies not only higher precision but higher recall as well. Fig. 9(d) shows that GS has the highest alignment agreement rate among the alignment methods for both language pairs.

5) Training Set Vocabulary Coverage by Phrase Table: We can also evaluate the inferred alignments extrinsically, e.g., by evaluating the SMT systems trained using those alignments. A desirable feature in a SMT system is to have as high vocabulary coverage as possible. This metric is highly sensitive to the performance of an alignment algorithm on infrequent words since they represent the majority of the vocabulary of a corpus (see Table II). Fig. 9(e) shows that alignment by GS leads to the best vocabulary coverage in all four alignment tasks. Note that word types that appear in the phrase table only as part of larger phrase(s) are excluded from this metric, since such words are practically out-of-vocabulary (OOV) except only in those specific contexts.

TABLE V
ALIGNMENT ERROR RATE (%) OF THE UNI-DIRECTIONAL AND SYMMETRIZED CZECH-ENGLISH ALIGNMENTS

Training set	1M sentences			15M sentences		
Direction	EC	CE	Sym.	EC	CE	Sym.
EM-5	45.1	41.4	30.9	40.6	38.4	27.7
GS-5	41.9	40.0	31.6	36.4	34.9	26.7
VB-5	37.8	36.5	28.9	31.9	32.1	24.1

Poor training set vocabulary coverage results in some non-OOV words being treated by the system as OOV, either dropping them from the output or leaving them untranslated. Such *pseudo-OOV* words further degrade the translation performance in addition to the OOV words. Fig. 9(f) shows that GS alignments lead to the lowest rate of pseudo-OOV words.

6) Phrase Table Size: In most machine translation applications, having a small model size is valuable, e.g., to reduce the memory requirement or the start-up/access time. Alignment methods can affect the induced phrase table sizes. Fig. 9(g) compares the number of phrase pairs in the SMT systems trained by different alignment methods. In the A $\leftrightarrow$ E task, where model size is of more concern compared to the smaller T $\leftrightarrow$ E task, GS results in significantly smaller phrase tables. This result is particularly remarkable since it means that a system using GS-inferred alignments achieves more vocabulary coverage (Section V-A-5) and higher BLEU scores (Section IV-B) with a smaller model size. Thanks to a larger intersection during alignment symmetrization (Fig. 9(d)), GS-based phrase tables contain a higher number of single-word phrase pairs (Fig. 9(e)). Moreover, fewer unaligned words after symmetrization lead to fewer poor-quality long phrase pairs [53].

7) Alignment Error Rate: Table V shows the alignment error rates (AERs) [10] obtained in the C $\leftrightarrow$ E alignment tasks

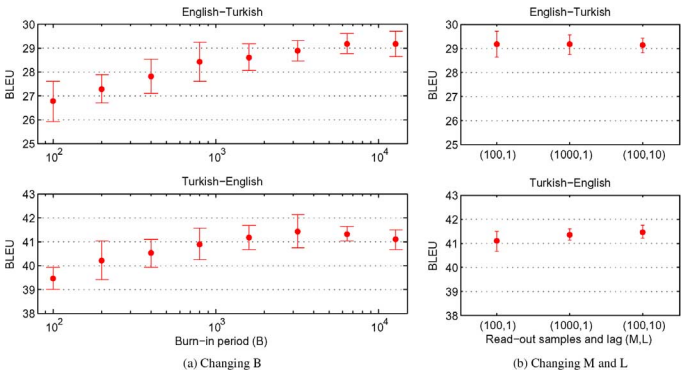

Fig. 10. BLEU scores obtained by different sampling settings (Section III-C). Averages and standard deviations are over 8 separate Gibbs chains. (a) $\mathrm{M} = 100$ and $\mathrm{L} = 1$. (b) $\mathrm{B} = 12800$. (a) Changing B; (b) changing M and L.

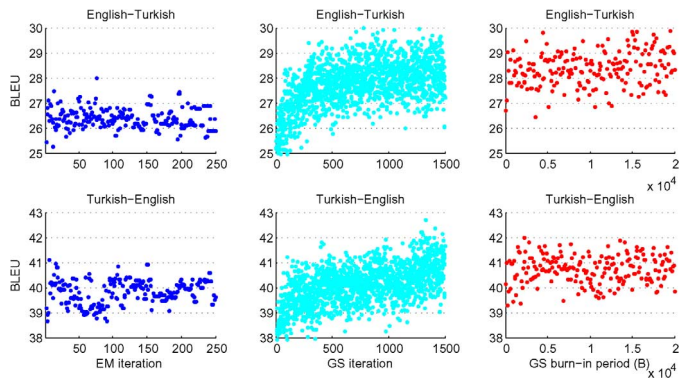

Fig. 11. BLEU scores of alignments estimated at different iterations. Left: EM, middle: samples from the Gibbs chain, right: GS viterbi estimates with $M = 100$, $L = 1$. A separate SMT system is trained at each shown data point on the plots. Note the difference in x-axis scales.

using a publicly available 515-sentence manually-aligned reference set [54]. The Bayesian methods achieve better AERs than EM in both alignment directions (denoted by "EC" and "CE"). Contrary to the ranking of the methods according to BLEU (Fig. 5), VB achieves the best AER, which also holds true after symmetrization (denoted by "Sym."). Furthermore, the symmetrized GS-5 alignment has the worst AER in the 1M-sentence experiment. These discrepancies support earlier findings by several others that AER is generally not a good predictor of BLEU performance [55].

As a final remark, in Table V EM-5 enjoys a larger amount of reduction in AER via symmetrization compared to GS-5, which suggests the possibility that the default alignment symmetrization heuristic in Moses ("grow-diag-final-and") has been fine-tuned for the default EM-based alignments, and thus other symmetrization/phrase extraction methods might work better for the GS- and VB-based alignments. For example, Bayesian alignment inference could be complemented with a probabilistic model of phrase extraction, e.g. [23], which is left as a future work.

B. Effect of Sampling Settings

We investigated the effect of changing the sampling settings B, M, and L (Section III-C) on $\mathrm{T} \leftrightarrow \mathrm{E}$ GS-N alignments. To account for the variability due to the randomness of the sampling process, we present in Fig. 10 the mean and the standard deviation of BLEU scores over eight separate chains with different random seeds. At each B value shown, eight separate SMT systems were trained. These eight runs each comprise a separate MERT run, thus error bars in Fig. 10 also include the variation due to MERT.

Fig. 10(a) shows the effect of changing B with $M = 100$ and $L = 1$. In this experiment, the sampler converges after roughly a few thousand iterations. Comparing the BLEU scores in Fig. 10(a) to those of the three EM-initialized samplers in Fig. 1, where $B = 400$, for the same language pair suggests that running more iterations of Gibbs sampling can compensate for poor initializations, or equivalently, initializing with EM alignments can provide a head start in the convergence of the Gibbs chain.

Fig. 10(b) compares the effect of different read-out schemes. The (M, L) settings of both (1000,1) and (100,10) collect samples over the same 1000-sample interval. We can deduce from their comparison in Fig. 10(b) that including or discarding the intermediate samples does not make a significant difference. On the other hand, comparing the settings (100,1) and (1000,1) confirms our intuition that increasing the number of samples (M) leads to more reliable (smaller variance) estimates of the Viterbi alignments.

C. Convergence and Variance Between Iterations

Fig. 11 compares the change in BLEU scores as iterations progress during both EM and GS. Each dot in the graphs correspond to a separate SMT system trained and optimized from the alignment estimated at that iteration. In the figure, there are two main sources of BLEU score variation between the iterations: updated alignments at each iteration and randomness due to MERT.

Comparing the BLEU scores of sample and Viterbi alignments obtained by GS, we observe smaller variance and higher average BLEU scores using Viterbi alignments. Compared to EM, GS achieves higher average BLEU scores, albeit with a larger amount of variation between iterations due to the random nature of sampling. To reduce the variation, a larger value of M (Section V-B) and/or a combination of alignments at different iterations can be used.

D. Computational Complexity

The computational complexity of the Gibbs sampling algorithm in Table I is linear in the number of sentences and roughly quadratic in the average number of words per sentence. Running Gibbs sampling (Model 1) on the largest of our datasets, the 15.4M-sentence Czech-English corpus, takes on average 33 seconds per iteration (steps 3–7 in Table I) using 24 threads on a 3.47GHz Intel Xeon X5690.[5] In the case of Model 2, the average time per Gibbs sampling iteration increases to 48 seconds. For comparison, a Model 1 EM iteration on the same hardware

[5]Our multi-threaded implementation is actually an approximation of Gibbs sampling, where the counts $N_{e,f}$ and C_r are not updated until the end of an iteration. Similar approximations have been done in scaling Gibbs sampling to large datasets using multiple parallel processors, e.g., in [56]. All large-data experiments reported in Sections IV-E and IV-F have been performed using this multi-threaded implementation.

and number of threads using MGIZA [57] takes 326 seconds on average (excluding pre-processing and initializations)[6].

VI. Conclusion

We developed a Gibbs sampling-based word alignment inference method for Bayesian IBM Models 1 and 2 and showed that it compares favorably to EM estimation in terms of translation BLEU scores. We observe the largest improvement when data is sparse, e.g., in the cases of smaller corpora and/or more morphological complexity. The proposed method successfully overcomes the well-known "garbage collection" problem of rare words in EM-estimated current models and learns a compact, sparse word translation distribution with more training vocabulary coverage. We also found Gibbs sampling to perform better than variational Bayes inference, which leaves a substantially high portion of source singletons unaligned. Additionally, we utilized alignment combination techniques to further improve the performance and robustness. Future research avenues include estimation of the hyperparameters from data/auxiliary sources and utilization of the proposed algorithm in either initialization or inference of more advanced alignment models.

Appendix
Derivation of the Gibbs Sampling Formula

In this section, we describe the derivation of the Gibbs sampler for IBM Model 2 given in (10). Since IBM Model 1 is a special case of Model 2 where $\mathbf{d}$ is fixed (Section III-D), the derivation of the sampler for Model 1 given in (6) would follow exactly the same steps, except that there would be no prior $P(\mathbf{d};\mathbf{\Phi})$ and the related terms.

A. The Dirichlet Priors

We choose a simple prior for the parameters $\mathbf{T}$ where each $\mathbf{t}_e$ has an independent[7] Dirichlet prior with hyperparameters $\mathbf{\Theta}_e$ (Section III-B):

$$P(\mathbf{t}_e;\mathbf{\Theta}_e)=\frac{1}{B(\mathbf{\Theta}_e)}\prod_{f=1}^{V_F}(t_{e,f})^{\theta_{e,f}-1}, \tag{11}$$

where $\theta_{e,f}>0\forall\{e,f\}$ and

$$B(\mathbf{\Theta}_e)\stackrel{def}{=}\frac{\prod_{f=1}^{V_F}\Gamma(\theta_{e,f})}{\Gamma\left(\sum_{f=1}^{V_F}\theta_{e,f}\right)} \tag{12}$$

Hence, the complete prior for $\mathbf{T}$ is given by:

$$P(\mathbf{T};\mathbf{\Theta})=\prod_{e=1}^{V_E}\frac{1}{B(\mathbf{\Theta}_e)}\prod_{f=1}^{V_F}(t_{e,f})^{\theta_{e,f}-1}. \tag{13}$$

[6]In the case of Model 2, for which multi-threading is not implemented in MGIZA, an EM iteration took 1960 seconds on average.

[7]While the prior knowledge about $\mathbf{T}$ could have been possibly expressed as a more refined, correlated distribution; we show that a simple, independent prior is also successful in biasing the parameters away from flat distributions.

Similarly, from Section III-D:

$$P(\mathbf{d};\mathbf{\Phi})=\frac{1}{B(\mathbf{\Phi})}\prod_{r=-\max_s I}^{\max_s I}(d_r)^{\phi_r-1}. \tag{14}$$

We further define the priors for the translation and distortion parameters to be independent so that $P(\mathbf{T},\mathbf{d})=P(\mathbf{T})P(\mathbf{d})$.

B. The Complete Distribution

Since we are only interested in inferring $\mathbf{A}$, we integrate out the unknowns $\mathbf{T}$ and $\mathbf{d}$ in (9) using (13) and (14):

$$P(\mathbf{F},\mathbf{A}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})=\iint_{\mathbf{T}\,\mathbf{d}}P(\mathbf{T};\mathbf{\Theta})P(\mathbf{d};\mathbf{\Phi})P(\mathbf{F},\mathbf{A}|\mathbf{E},\mathbf{T},\mathbf{d}) \tag{15}$$

$$=\int_{\mathbf{T}}\prod_{e=1}^{V_E}\frac{1}{B(\mathbf{\Theta}_e)}\prod_{f=1}^{V_F}(t_{e,f})^{N_{e,f}+\theta_{e,f}-1}\cdot\int_{\mathbf{d}}\frac{1}{B(\mathbf{\Phi})}\prod_{r=-\max_s I}^{\max_s I}(d_r)^{C_r+\phi_r-1} \tag{16}$$

$$=\prod_{e=1}^{V_E}\frac{1}{B(\mathbf{\Theta}_e)}\int_{\mathbf{t}_e}\prod_{f=1}^{V_F}(t_{e,f})^{N_{e,f}+\theta_{e,f}-1}\cdot\frac{1}{B(\mathbf{\Phi})}\int_{\mathbf{d}}\prod_{r=-\max_s I}^{\max_s I}(d_r)^{C_r+\phi_r-1}. \tag{17}$$

As a result of choosing conjugate priors, the integrands with respect to $\mathbf{t}_e$ and $\mathbf{d}$ in (17) can be recognized to be in the same form as the priors (i.e., Dirichlet distributions) with new sets of parameters $\mathbf{N}_e+\mathbf{\Theta}_e$ and $\mathbf{C}+\mathbf{\Phi}$, respectively, where we have defined $\mathbf{N}_e=N_{e,1}\cdots N_{e,V_F}$ and $\mathbf{C}=C_{-\max_s I}\cdots C_{\max_s I}$. Since the integral of a probability distribution is equal to 1, we obtain the closed-form expression:

$$P(\mathbf{F},\mathbf{A}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})=\prod_{e=1}^{V_E}\frac{B(\mathbf{N}_e+\mathbf{\Theta}_e)}{B(\mathbf{\Theta}_e)}\cdot\frac{B(\mathbf{C}+\mathbf{\Phi})}{B(\mathbf{\Phi})}. \tag{18}$$

C. Gibbs Sampler Derivation

Given the complete distribution in (18), the Gibbs sampling formula $P(z_j|\mathbf{z}^{\neg j})$ (Section III-C) can be derived as:

$$P(a_j|\mathbf{E},\mathbf{F},\mathbf{A}^{\neg j};\mathbf{\Theta},\mathbf{\Phi})$$

$$=\frac{P(\mathbf{F},\mathbf{A}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})}{P(\mathbf{F},\mathbf{A}^{\neg j}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})} \tag{19}$$

$$\propto\frac{P(\mathbf{F},\mathbf{A}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})}{P(\mathbf{F}^{\neg j},\mathbf{A}^{\neg j}|\mathbf{E};\mathbf{\Theta},\mathbf{\Phi})} \tag{20}$$

$$=\prod_{e=1}^{V_E}\frac{B(\mathbf{N}_e+\mathbf{\Theta}_e)}{B(\mathbf{N}_e^{\neg j}+\mathbf{\Theta}_e)}\cdot\frac{B(\mathbf{C}+\mathbf{\Phi})}{B(\mathbf{C}^{\neg j}+\mathbf{\Phi})} \tag{21}$$

$$=\frac{B(\mathbf{N}_{e_{a_j}}+\mathbf{\Theta}_{e_{a_j}})}{B(\mathbf{N}_{e_{a_j}}^{\neg j}+\mathbf{\Theta}_{e_{a_j}})}\cdot\frac{B(\mathbf{C}+\mathbf{\Phi})}{B(\mathbf{C}^{\neg j}+\mathbf{\Phi})} \tag{22}$$

$$
\begin{aligned}
&= \frac{\prod_{f=1}^{V_F} \Gamma\left(N_{e_{a_j},f} + \theta_{e_{a_j},f}\right)}{\prod_{f=1}^{V_F} \Gamma\left(N^{\neg j}_{e_{a_j},f} + \theta_{e_{a_j},f}\right)} \\
&\quad \cdot \frac{\Gamma\left(\sum_{f=1}^{V_F}\left(N^{\neg j}_{e_{a_j},f} + \theta_{e_{a_j},f}\right)\right)}{\Gamma\left(\sum_{f=1}^{V_F}\left(N_{e_{a_j},f} + \theta_{e_{a_j},f}\right)\right)} \\
&\quad \cdot \frac{\prod_{r=-\max_s I}^{\max_s I} \Gamma(C_r + \phi_r)}{\prod_{r=-\max_s I}^{\max_s I} \Gamma\left(C^{\neg j}_r + \phi_r\right)} \\
&\quad \cdot \frac{\Gamma\left(\sum_{r=-\max_s I}^{\max_s I}\left(C^{\neg j}_r + \phi_r\right)\right)}{\Gamma\left(\sum_{r=-\max_s I}^{\max_s I}(C_r + \phi_r)\right)} \qquad (23) \\
&= \frac{\Gamma\left(N_{e_{a_j},f_j} + \theta_{e_{a_j},f_j}\right)}{\Gamma\left(N^{\neg j}_{e_{a_j},f_j} + \theta_{e_{a_j},f_j}\right)} \\
&\quad \cdot \frac{\Gamma\left(\sum_{f=1}^{V_F}\left(N^{\neg j}_{e_{a_j},f} + \theta_{e_{a_j},f}\right)\right)}{\Gamma\left(\sum_{f=1}^{V_F}\left(N_{e_{a_j},f} + \theta_{e_{a_j},f}\right)\right)} \\
&\quad \cdot \frac{\Gamma\left(C_{a_j-\lfloor j\frac{I}{J}\rfloor} + \phi_{a_j-\lfloor j\frac{I}{J}\rfloor}\right)}{\Gamma\left(C^{\neg j}_{a_j-\lfloor j\frac{I}{J}\rfloor} + \phi_{a_j-\lfloor j\frac{I}{J}\rfloor}\right)} \\
&\quad \cdot \frac{\Gamma\left(\sum_{r=-\max_s I}^{\max_s I}\left(C^{\neg j}_r + \phi_r\right)\right)}{\Gamma\left(\sum_{r=-\max_s I}^{\max_s I}(C_r + \phi_r)\right)} \qquad (24) \\
&= \left(N^{\neg j}_{e_{a_j},f_j} + \theta_{e_{a_j},f_j}\right) \\
&\quad \cdot \frac{1}{\sum_{f=1}^{V_F}\left(N^{\neg j}_{e_{a_j},f} + \theta_{e_{a_j},f}\right)} \\
&\quad \cdot \left(C^{\neg j}_{a_j-\lfloor j\frac{I}{J}\rfloor} + \phi_{a_j-\lfloor j\frac{I}{J}\rfloor}\right) \\
&\quad \cdot \frac{1}{\sum_{r=-\max_s I}^{\max_s I}\left(C^{\neg j}_r + \phi_r\right)} \qquad (25) \\
&\propto \frac{N^{\neg j}_{e_{a_j},f_j} + \theta_{e_{a_j},f_j}}{\sum_{f=1}^{V_F} N^{\neg j}_{e_{a_j},f} + \sum_{f=1}^{V_F}\theta_{e_{a_j},f}} \\
&\quad \cdot \left(C^{\neg j}_{a_j-\lfloor j\frac{I}{J}\rfloor} + \phi_{a_j-\lfloor j\frac{I}{J}\rfloor}\right), \qquad (26)
\end{aligned}
$$

where (20) follows since $P(f_j|\mathbf{A}^{\neg j}, \mathbf{E}; \boldsymbol{\Theta})$ is independent of a_j, in (21) we used (18), in (23) we used (12) and grouped similar factors, in (25) each fraction is simplified using the property of the gamma function $\Gamma(x+1) = x\Gamma(x)$, and in (26) the proportionality comes from the omission of the last term in (25), which is constant for all values of a_j.

ACKNOWLEDGMENT

The authors would like to thank Kemal Oflazer for providing the Turkish morphological analyzer, and anonymous reviewers for their comments to improve the paper.

REFERENCES

[1] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[2] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. HLT-NAACL*, Edmonton, AB, Canada, May–Jun. 2003, pp. 48–54.
[3] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguist.*, vol. 33, no. 2, pp. 201–228, 2007.
[4] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable inference and training of context-rich syntactic translation models," in *Proc. ACL-COLING*, Sydney, Australia, Jul. 2006, pp. 961–968.
[5] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.
[6] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proc. COLING*, 1996, pp. 836–841.
[7] P. Liang, B. Taskar, and D. Klein, "Alignment by agreement," in *Proc. HLT-NAACL*, New York, NY, Canada, Jun. 2006, pp. 104–111.
[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[9] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer, and S. Mohanty, "But dictionaries are data too," in *Proc. HLT*, Plainsboro, NJ, USA, 1993, pp. 202–205.
[10] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003.
[11] R. C. Moore, "Improving IBM word alignment model 1," in *Proc. ACL*, Barcelona, Spain, Jul. 2004, pp. 518–525.
[12] C. Mermer and M. Saraclar, "Bayesian word alignment for statistical machine translation," in *Proc. ACL-HLT: Short Papers*, Portland, OR, USA, Jun. 2011, pp. 182–187.
[13] B. Zhao and E. P. Xing, "BiTAM: Bilingual topic admixture models for word alignment," in *Proc. COLING-ACL: Poster Sessions*, Sydney, Australia, Jul. 2006, pp. 969–976.
[14] A. Vaswani, L. Huang, and D. Chiang, "Smaller alignment models for better translations: Unsupervised word alignment with the l_0-norm," in *Proc. ACL*, 2012, pp. 311–319.
[15] S. Zhao and D. Gildea, "A fast fertility hidden Markov model for word alignment using MCMC," in *Proc. EMNLP*, Cambridge, MA, USA, Oct. 2010, pp. 596–605.
[16] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Proc. ACL*, Prague, Czech Republic, Jun. 2007, pp. 744–751.
[17] J. Gao and M. Johnson, "A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers," in *Proc. EMNLP*, Honolulu, HI, USA, Oct. 2008, pp. 344–352.
[18] S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," in *Proc. ACL-COLING*, Sydney, Australia, Jul. 2006, pp. 673–680.
[19] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proc. ACL-AJCNLP*, Suntec, Singapore, Aug. 2009, pp. 100–108.
[20] M. Johnson, T. L. Griffiths, and S. Goldwater, "Bayesian inference for PCFGs via Markov Chain Monte Carlo," in *Proc. NAACL-HLT*, Rochester, NY, USA, Apr. 2007, pp. 139–146.
[21] D. Chiang, J. Graehl, K. Knight, A. Pauls, and S. Ravi, "Bayesian inference for finite-state transducers," in *Proc. NAACL-HLT*, Los Angeles, CA, USA, Jun. 2010, pp. 447–455.
[22] P. Blunsom, T. Cohn, C. Dyer, and M. Osborne, "A Gibbs sampler for phrasal synchronous grammar induction," in *Proc. ACL-AJCNLP*, Suntec, Singapore, Aug. 2009, pp. 782–790.
[23] J. DeNero, A. Bouchard-Côté, and D. Klein, "Sampling alignment structure under a Bayesian translation model," in *Proc. EMNLP*, Honolulu, HI, USA, Oct. 2008, pp. 314–323.
[24] J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised Chinese word segmentation for statistical machine translation," in *Proc. COLING*, Manchester, U.K., Aug. 2008, pp. 1017–1024.
[25] T. Nguyen, S. Vogel, and N. A. Smith, "Nonparametric word segmentation for machine translation," in *Proc. COLING*, 2010, pp. 815–823.
[26] T. Chung and D. Gildea, "Unsupervised tokenization for machine translation," in *Proc. EMNLP*, Singapore, Aug. 2009, pp. 718–726.
[27] D. Riley and D. Gildea, "Improving the IBM alignment models using variational Bayes," in *Proc. ACL: Short Papers*, 2012, pp. 306–310.
[28] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov... 1984.
[29] F. J. Och and H. Ney, "A comparison of alignment models for statistical machine translation," in *Proc. COLING*, 2000, pp. 1086–1090.

[30] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1674–1682, Sep. 2006.
[31] M. Paul, M. Federico, and S. Stücker, "Overview of the IWSLT 2010 evaluation campaign," in *Proc. IWSLT*, Dec. 2010, pp. 3–27.
[32] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, pp. 359–394, 1999.
[33] [Online]. Available: http://aclweb.org/supplementals/P/P11/P11-2032.Software.txt
[34] 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 *Proc. ACL: Demo and Poster Sessions*, Prague, Czech Republic, Jun. 2007, pp. 177–180.
[35] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. ICSLP*, 2002, vol. 3.
[36] O. F. Zaidan, "Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems," *Prague Bull. Math. Linguist.*, vol. 91, no. 1, pp. 79–88, 2009.
[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.
[38] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proc. ACL:HLT*, Portland, OR, USA, 2011, pp. 176–181.
[39] W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2007 MT system," in *Proc. IWSLT*, Trento, Italy, 2007.
[40] C. Dyer, J. H. Clark, A. Lavie, and N. A. Smith, "Unsupervised word alignment with arbitrary features," in *Proc. ACL:HLT*, Portland, OR, USA, Jun. 2011, pp. 409–419.
[41] D. Riley and D. Gildea, Improving the performance of GIZA++ using variational Bayes Univ. of Rochester, Comput. Sci. Dept., Rochester, NY, USA, Tech. Rep. 963, Dec. 2010.
[42] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguist. Comput.*, vol. 9, no. 2, 1994.
[43] A. Eisele and Y. Chen, "MultiUN: A multilingual corpus from United Nation documents," in *Proc. LREC*, 2010, pp. 2868–2872.
[44] O. R. Nizar Habash and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proc. 2nd Int. Conf. Arabic Lang. Resources and Tools*, 2009.
[45] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 Workshop on Statistical Machine Translation," in *Proc. WMT*, 2012, pp. 10–51.
[46] O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna, "The Joy of Parallelism with CzEng 1.0," in *Proc. LREC*, 2012.
[47] S. Chatterjee and N. Cancedda, "Minimum error rate training by sampling the translation lattice," in *Proc. EMNLP*, 2010, pp. 606–615.
[48] D. Cer, D. Jurafsky, and C. D. Manning, "Regularization and search for minimum error rate training," in *Proc. WMT*, 2008, pp. 26–34.
[49] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proc. HLT-NAACL*, 2004, pp. 169–176.
[50] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. AMTA*, 2006, pp. 223–231.
[51] T. Bodrumlu, K. Knight, and S. Ravi, "A new objective function for word alignment," in *Proc. NAACL-HLT Workshop Integer Linear Program. for Natural Lang. Process.*, Boulder, CO, USA, Jun. 2009, pp. 28–35.
[52] T. Schoenemann, "Probabilistic word alignment under the l_0-norm," in *Proc. CoNLL*, 2011, pp. 172–180.
[53] F. Guzman, Q. Gao, and S. Vogel, "Reassessment of the role of phrase extraction in PBSMT," in *Proc. MT Summit*, 2009.
[54] O. Bojar and M. Prokopová, "Czech-English Word Alignment," in *Proc. LREC*, 2006, pp. 1236–1239.
[55] A. Fraser and D. Marcu, "Measuring word alignment quality for statistical machine translation," *Comput. Linguist.*, vol. 33, no. 3, pp. 293–303, 2007.
[56] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *J. Mach. Learn. Res.*, vol. 10, pp. 1801–1828, 2009.
[57] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Proc. ACL-HLT Workshop Software Eng., Testing, Quality Assurance for NLP*, 2008, pp. 49–57.

Coşkun Mermer received his B.S. degree in electrical and electronics engineering in 1998 from Bilkent University, Ankara, Turkey, and his M.S. degree in electrical engineering in 2001 from University of Washington, Seattle, WA. He is currently a Ph.D. candidate in the Electrical and Electronics Engineering Department at Boğaziçi University, Istanbul, Turkey. In 2006, he joined the Speech and Language Technologies Laboratory at TÜBİTAK-BİLGEM, where he has led the TÜBİTAK team in several machine translation evaluations.

He is currently serving on the Student Board of EACL.

Murat Saraçlar (M'00) received his B.S degree in 1994 from the Electrical and Electronics Engineering Department at Bilkent University, M.S.E. degree in 1997 and Ph.D. degree in 2001 from the Electrical and Computer Engineering Department at the Johns Hopkins University. He is currently an associate professor at the Electrical and Electronic Engineering Department of Boğaziçi University. From 2000 to 2005, he was with the multimedia services department at AT&T Labs Research. Dr. Saraçlar was a member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007–2009). He is currently serving as an associate editor for IEEE SIGNAL PROCESSING LETTERS and IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He is on the editorial boards of Computer Speech and Language, and Language Resources and Evaluation.

Ruhi Sarikaya (SM'08) received his B.S. degree from Bilkent University, Turkey in 1995, M.S. degree from Clemson University, SC in 1997 and Ph.D. degree from Duke University, NC in 2001 all in electrical and computer engineering. He is currently a principal scientist and the manager of language understanding and dialog systems science group at Microsoft. He was a research staff member and team lead in the Human Language Technologies Group at IBM T.J. Watson Research Center for ten years. He is currently serving as the general co-chair of IEEE SLT'12. He is also serving as associate editors of IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS. He also served as the lead guest editor of the special issue on "Processing Morphologically-Rich Languages" for IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.