Bayesian Word Alignment for Statistical Machine Translation

Coşkun Mermer[1,2]
[1]BILGEM
TUBITAK
Gebze 41470 Kocaeli, Turkey
coskun@uekae.tubitak.gov.tr

Murat Saraçlar[2]
[2]Electrical and Electronics Eng. Dept.
Bogazici University
Bebek 34342 Istanbul, Turkey
murat.saraclar@boun.edu.tr

Abstract

In this work, we compare the translation performance of word alignments obtained via Bayesian inference to those obtained via expectation-maximization (EM). We propose a Gibbs sampler for fully Bayesian inference in IBM Model 1, integrating over all possible parameter values in finding the alignment distribution. We show that Bayesian inference outperforms EM in all of the tested language pairs, domains and data set sizes, by up to 2.99 BLEU points. We also show that the proposed method effectively addresses the well-known rare word problem in EM-estimated models; and at the same time induces a much smaller dictionary of bilingual word-pairs.

1 Introduction

Word alignment is a crucial early step in the training of most statistical machine translation (SMT) systems, in which the estimated alignments are used for constraining the set of candidates in phrase/grammar extraction (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006). State-of-the-art word alignment models, such as IBM Models (Brown et al., 1993), HMM (Vogel et al., 1996), and the jointly-trained symmetric HMM (Liang et al., 2006), contain a large number of parameters (e.g., word translation probabilities) that need to be estimated in addition to the desired hidden alignment variables.

The most common method of inference in such models is expectation-maximization (EM) (Dempster et al., 1977) or an approximation to EM when exact EM is intractable. However, being a maximization (e.g., maximum likelihood (ML) or maximum *a posteriori* (MAP)) technique, EM is generally prone to local optima and overfitting. In essence, the alignment distribution obtained via EM takes into account only the most likely point in the parameter space, but does not consider contributions from other points.

Problems with the standard EM estimation of IBM Model 1 was pointed out by Moore (2004) and 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 was not reported. Zhao and Xing (2006) note that the parameter estimation (for which they use variational EM) suffers from data sparsity and use symmetric Dirichlet priors, but they find the MAP solution.

Bayesian inference, the approach in this paper, have recently been applied to several unsupervised learning problems in NLP (Goldwater and Griffiths, 2007; Johnson et al., 2007) as well as to other tasks in SMT such as synchronous grammar induction (Blunsom et al., 2009) and learning phrase alignments directly (DeNero et al., 2008).

Word alignment learning problem was addressed jointly with segmentation learning in Xu et al. (2008), Nguyen et al. (2010), and Chung and Gildea (2009). The former two works place *nonparametric* priors (also known as cache models) on the parameters and utilize Gibbs sampling. However, alignment inference in neither of these works is exactly Bayesian since the alignments are updated by running GIZA++ (Xu et al., 2008) or by local maximization (Nguyen et al., 2010). On the other hand,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 182–187,
Portland, Oregon, June 19-24, 2011.

Chung and Gildea (2009) apply a sparse Dirichlet prior on the multinomial parameters to prevent overfitting. They use variational Bayes for inference, but they do not investigate the effect of Bayesian inference to word alignment in isolation. Recently, Zhao and Gildea (2010) proposed fertility extensions to IBM Model 1 and HMM, but they do not place any prior on the parameters and their inference method is actually stochastic EM (also known as Monte Carlo EM), a ML technique in which sampling is used to approximate the expected counts in the E-step. Even though they report substantial reductions in alignment error rate, the translation BLEU scores do not improve.

Our approach in this paper is fully Bayesian in which the alignment probabilities are inferred by integrating over all possible parameter values assuming an intuitive, sparse prior. We develop a Gibbs sampler for alignments under IBM Model 1, which is relevant for the state-of-the-art SMT systems since: (1) Model 1 is used in bootstrapping the parameter settings for EM training of higher-order alignment models, and (2) many state-of-the-art SMT systems use Model 1 translation probabilities as features in their log-linear model. We evaluate the inferred alignments in terms of the end-to-end translation performance, where we show the results with a variety of input data to illustrate the general applicability of the proposed technique. To our knowledge, this is the first work to directly investigate the effects of Bayesian alignment inference on translation performance.

2 Bayesian Inference with IBM Model 1

Given a sentence-aligned parallel corpus $(\mathbf{E}, \mathbf{F})$, let e_i (f_j) denote the i-th (j-th) source (target)[1] word in $\mathbf{e}$ ($\mathbf{f}$), which in turn consists of I (J) words and denotes the s-th sentence in $\mathbf{E}$ ($\mathbf{F}$).[2] Each source sentence is also hypothesized to have an additional imaginary "null" word e_0. Also let V_E (V_F) denote the size of the observed source (target) vocabulary.

In Model 1 (Brown et al., 1993), each target word f_j is associated with a hidden alignment variable a_j whose value ranges over the word positions in the corresponding source sentence. 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 $t_{e,f} = P(f|e)$.

The joint distribution of the Model-1 variables is given by the following generative model[3]:

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

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

In the proposed Bayesian setting, we treat $\mathbf{T}$ as a random variable with a prior $P(\mathbf{T})$. To find a suitable prior for $\mathbf{T}$, we re-write (2) as:

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

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

where in (3) the count variable $n_{e,f}$ 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) $N_{e,f} = \sum_s n_{e,f}$. Since the distribution over $\{t_{e,f}\}$ in (4) is in the *exponential family*, specifically being a multinomial distribution, we choose the conjugate prior, in this case the Dirichlet distribution, for computational convenience.

For each source word type e, we assume the prior distribution for $\mathbf{t}_e = t_{e,1} \cdots t_{e,V_F}$, which is itself a distribution over the target vocabulary, to be a Dirichlet distribution (with its own set of hyperparameters $\mathbf{\Theta}_e = \theta_{e,1} \cdots \theta_{e,V_F}$) independent from the priors of other source word types:

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

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

We choose symmetric Dirichlet priors identically for all source words e with $\theta_{e,f} = \theta = 0.0001$ to obtain a sparse Dirichlet prior. A sparse prior favors

[1] We use the "source" and "target" labels following the generative process, in which $\mathbf{E}$ generates $\mathbf{F}$ (*cf.* Eq. 1).

[2] Dependence of the sentence-level variables $\mathbf{e}$, $\mathbf{f}$, I, J (and $\mathbf{a}$ and n, which are introduced later) on the sentence index s should be understood even though not explicitly indicated for notational simplicity.

[3] We omit $P(J|\mathbf{e})$ since both J and $\mathbf{e}$ are observed and so this term does not affect the inference of hidden variables.

distributions that peak at a single target word and penalizes flatter translation distributions, even for rare words. This choice addresses the well-known problem in the IBM Models, and more severely in Model 1, in which rare words act as "garbage collectors" (Och and Ney, 2003) and get assigned excessively large number of word alignments.

Then we obtain the joint distribution of all (observed + hidden) variables as:

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

where $\mathbf{\Theta} = \mathbf{\Theta}_1 \cdots \mathbf{\Theta}_{V_E}$.

To infer the posterior distribution of the alignments, we use Gibbs sampling (Geman and Geman, 1984). One possible method is to derive the Gibbs sampler from $P(\mathbf{E}, \mathbf{F}, \mathbf{A}, \mathbf{T}; \mathbf{\Theta})$ obtained in (5) and sample the unknowns $\mathbf{A}$ and $\mathbf{T}$ in turn, resulting in an *explicit* Gibbs sampler. In this work, we marginalize out $\mathbf{T}$ by:

$$P(\mathbf{E}, \mathbf{F}, \mathbf{A}; \mathbf{\Theta}) = \int_{\mathbf{T}} P(\mathbf{E}, \mathbf{F}, \mathbf{A}, \mathbf{T}; \mathbf{\Theta}) \quad (6)$$

and obtain a *collapsed* Gibbs sampler, which samples only the alignment variables.

Using $P(\mathbf{E}, \mathbf{F}, \mathbf{A}; \mathbf{\Theta})$ obtained in (6), the Gibbs sampling formula for the individual alignments is derived as:[4]

$$P(a_j = i|\mathbf{E}, \mathbf{F}, \mathbf{A}^{\neg j}; \mathbf{\Theta}) = \frac{N^{\neg j}_{e_i, f_j} + \theta_{e_i, f_j}}{\sum_{f=1}^{V_F} N^{\neg j}_{e_i, f} + \sum_{f=1}^{V_F} \theta_{e_i, f}} \quad (7)$$

where the superscript $\neg j$ denotes the exclusion of the current value of a_j.

The algorithm is given in Table 1. Initialization of $\mathbf{A}$ in Step 1 can be arbitrary, but for faster convergence special initializations have been used, e.g., using the output of EM (Chiang et al., 2010). Once the Gibbs sampler is deemed to have converged after B burn-in iterations, we collect M samples of $\mathbf{A}$ with L iterations in-between[5] to estimate $P(\mathbf{A}|\mathbf{E}, \mathbf{F})$. To obtain the Viterbi alignments, which are required for phrase extraction (Koehn et al., 2003), we select for each a_j the most frequent value in the M collected samples.

[4]The derivation is quite standard and similar to other Dirichlet-multinomial Gibbs sampler derivations, e.g. (Resnik and Hardisty, 2010).

[5]A lag is introduced to reduce correlation between samples.

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

Table 1: Gibbs sampling algorithm for IBM Model 1 (implemented in the accompanying software).

3 Experimental Setup

For Turkish↔English experiments, we used the 20K-sentence travel domain BTEC dataset (Kikui et al., 2006) from the yearly IWSLT evaluations[6] for training, the CSTAR 2003 test set for development, and the IWSLT 2004 test set for testing[7]. For Czech↔English, we used the 95K-sentence news commentary parallel corpus from the WMT shared task[8] for training, *news2008* set for development, *news2009* set for testing, and the 438M-word English and 81.7M-word Czech monolingual news corpora for additional language model (LM) training. For Arabic↔English, we used the 65K-sentence LDC2004T18 (news from 2001-2004) for training, the AFP portion of LDC2004T17 (news from 1998, single reference) for development and testing (about 875 sentences each), and the 298M-word English and 215M-word Arabic AFP and Xinhua subsets of the respective Gigaword corpora (LDC2007T07 and LDC2007T40) for additional LM training. All language models are 4-gram in the travel domain experiments and 5-gram in the news domain experiments.

For each language pair, we trained standard phrase-based SMT systems in both directions (including alignment symmetrization and log-linear model tuning) using Moses (Koehn et al., 2007), SRILM (Stolcke, 2002), and ZMERT (Zaidan, 2009) tools and evaluated using BLEU (Papineni et al., 2002). To obtain word alignments, we used the accompanying Perl code for Bayesian inference and

[6]International Workshop on Spoken Language Translation. http://iwslt2010.fbk.eu

[7]Using only the first English reference for symmetry.

[8]Workshop on Machine Translation. http://www.statmt.org/wmt10/translation-task.html

Method	TE	ET	CE	EC	AE	EA
EM-5	38.91	26.52	14.62	10.07	15.50	15.17
EM-80	39.19	26.47	14.95	10.69	15.66	15.02
GS-N	41.14	27.55	14.99	**10.85**	14.64	15.89
GS-5	40.63	27.24	**15.45**	10.57	**16.41**	15.82
GS-80	**41.78**	**29.51**	15.01	10.68	15.92	**16.02**
M4	39.94	27.47	15.47	11.15	16.46	15.43

Table 2: BLEU scores in translation experiments. E: English, T: Turkish, C: Czech, A: Arabic.

GIZA++ (Och and Ney, 2003) for EM.

For each translation task, we report two EM estimates, obtained after 5 and 80 iterations (EM-5 and EM-80), respectively; and three Gibbs sampling estimates, two of which were initialized with those two EM Viterbi alignments (GS-5 and GS-80) and a third was initialized *naively*[9] (GS-N). Sampling settings were B = 400 for T↔E, 4000 for C↔E and 8000 for A↔E; M = 100, and L = 10. For reference, we also report the results with IBM Model 4 alignments (M4) trained in the standard bootstrapping regimen of $1^5H^53^34^3$.

4 Results

Table 2 compares the BLEU scores of Bayesian inference and EM estimation. In all translation tasks, Bayesian inference outperforms EM. The improvement range is from 2.59 (in Turkish-to-English) up to 2.99 (in English-to-Turkish) BLEU points in travel domain and from 0.16 (in English-to-Czech) up to 0.85 (in English-to-Arabic) BLEU points in news domain. Compared to the state-of-the-art IBM Model 4, the Bayesian Model 1 is better in all travel domain tasks and is comparable or better in the news domain.

Fertility of a source word is defined as the number of target words aligned to it. Table 3 shows the distribution of fertilities in alignments obtained from different methods. Compared to EM estimation, including Model 4, the proposed Bayesian inference dramatically reduces "questionable" high-fertility (4 $\leq$ fertility $\leq$ 7) alignments and almost entirely eliminates "excessive" alignments (fertility $\geq$ 8)[10].

[9]Each target word was aligned to the source candidate that co-occured the most number of times with that target word in the entire parallel corpus.

Method	TE	ET	CE	EC	AE	EA
All	140K	183K	1.63M	1.78M	1.49M	1.82M
EM-80	5.07K	2.91K	52.9K	45.0K	69.1K	29.4K
M4	5.35K	3.10K	36.8K	36.6K	55.6K	36.5K
GS-80	755	419	14.0K	10.9K	47.6K	18.7K
EM-80	426	227	10.5K	18.6K	21.4K	24.2K
M4	81	163	2.57K	10.6K	9.85K	21.8K
GS-80	1	1	39	110	689	525
EM-80	24	24	28	30	44	46
M4	9	9	9	9	9	9
GS-80	8	8	13	18	20	19

Table 3: Distribution of inferred alignment fertilities. The four blocks of rows from top to bottom correspond to (in order) the total number of source tokens, source tokens with fertilities in the range 4–7, source tokens with fertilities higher than 7, and the maximum observed fertility. The first language listed is the *source* in alignment (Section 2).

Method	TE	ET	CE	EC	AE	EA
EM-80	52.5K	38.5K	440K	461K	383K	388K
M4	57.6K	40.5K	439K	441K	422K	405K
GS-80	**23.5K**	**25.4K**	**180K**	**209K**	**158K**	**176K**

Table 4: Sizes of bilingual dictionaries induced by different alignment methods.

The number of distinct word-pairs induced by an alignment has been recently proposed as an objective function for word alignment (Bodrumlu et al., 2009). Small dictionary sizes are preferred over large ones. Table 4 shows that the proposed inference method substantially reduces the alignment dictionary size, in most cases by more than 50%.

5 Conclusion

We developed a Gibbs sampling-based Bayesian inference method for IBM Model 1 word alignments and showed that it outperforms EM estimation in terms of translation BLEU scores across several language pairs, data sizes and domains. As a result of this increase, Bayesian Model 1 alignments perform close to or better than the state-of-the-art IBM

[10]The GIZA++ implementation of Model 4 artificially limits fertility parameter values to at most nine.

Model 4. The proposed method learns a compact, sparse translation distribution, overcoming the well-known "garbage collection" problem of rare words in EM-estimated current models.

Acknowledgments

Murat Saraçlar is supported by the TÜBA-GEBİP award.

References

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 782–790, Suntec, Singapore, August.

Tugba Bodrumlu, Kevin Knight, and Sujith Ravi. 2009. A new objective function for word alignment. In *Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing*, pages 28–35, Boulder, Colorado, June. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Chiang, Jonathan Graehl, Kevin Knight, Adam Pauls, and Sujith Ravi. 2010. Bayesian inference for finite-state transducers. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 447–455, Los Angeles, California, June.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 718–726, Singapore, August.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia, July.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 6(6):721–741, November.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic, June.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–146, Rochester, New York, April.

Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1674–1682.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003, Main Papers*, pages 48–54, Edmonton, May-June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June.

Robert C. Moore. 2004. Improving IBM word alignment Model 1. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 518–525, Barcelona, Spain, July.

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for ma-

chine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 815–823, Beijing, China, August.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. *University of Maryland Computer Science Department; CS-TR-4956*, June.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING*, pages 836–841.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–10124, Manchester, UK, August.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91(1):79–88.

Shaojun Zhao and Daniel Gildea. 2010. A fast fertility hidden Markov model for word alignment using MCMC. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 596–605, Cambridge, MA, October.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 969–976, Sydney, Australia, July. Association for Computational Linguistics.