The TÜBİTAK Statistical Machine Translation System for IWSLT 2012

Coşkun Mermer, Hamza Kaya, İlknur Durgar El-Kahlout, Mehmet Uğur Doğan

TÜBİTAK BİLGEM
Gebze 41470 Kocaeli, Turkey
{coskun.mermer,hamza.kaya,ilknur.durgar,mugur.dogan}@tubitak.gov.tr

Abstract

We describe the TÜBİTAK submission to the IWSLT 2012 Evaluation Campaign. Our system development focused on utilizing Bayesian alignment methods such as variational Bayes and Gibbs sampling in addition to the standard GIZA++ alignments. The submitted tracks are the Arabic-English and Turkish-English TED Talks translation tasks.

1. Introduction

In the 2012 IWSLT Evaluation Campaign [1], we participated in the TED task for the Arabic-English and Turkish-English language pairs. Our major focus this year was improving the word alignment.

Maximum-likelihood (ML) word alignments obtained using GIZA++ [2] can exhibit overfitting, e.g., rare words can have excessively high alignment fertilities [3], also known as "garbage collection" [2, 4]. Furthermore, ML estimation gives a point-estimate of the parameters, which assumes that the unknown parameters are fixed (as opposed to being a random variable). Finally, the expectation-maximization (EM) method used in obtaining the ML-estimates can get stuck in local optima.

As an alternative approach, in our submission we experimented with the Bayesian approach to word alignment. In the Bayesian framework, the parameters are treated as random variables with a prior distribution. By choosing a suitable prior, we can bias the inferred solution towards what we would expect from our prior knowledge and away from unlikely solutions such as garbage collection.

The remainder of this paper is organized as follows. Section 2 summarizes the word alignment methods and their parameter settings used in our systems. Sections 3 and 4 describe the data used and the common aspects of system development in both language tracks. The specifics of the Arabic-English and Turkish-English submissions and the experimental results are described in Sections 5 and 6, respectively, followed by the conclusions.

2. Word alignment methods

In most commonly-used word alignment methods, such as those used in GIZA++ [2], the model parameters are estimated via EM, which is a ML approach. For this evaluation, we experimented with two additional methods that use a Bayesian approach, where the parameters are treated as random variables with a prior and they are integrated over for alignment inference.

The main difference between the ML and Bayesian approaches to word alignment can be summarized as follows [5]. Given a parallel corpus $\{\mathbf{E}, \mathbf{F}\}$, let $\mathbf{A}$ denote the hidden word alignments. The IBM word alignment models [6] assign a probability to each possible alignment through $P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T})$, where $\mathbf{T}$ denotes the (unknown) translation parameters. The ML solution returns the posterior distribution of the alignments $P(\mathbf{A}|\mathbf{E}, \mathbf{F}, \mathbf{T}^*)$, such that:

$$T^* = \arg\max_T P(\mathbf{F}|\mathbf{E}, \mathbf{T}) \quad (1)$$

$$= \arg\max_T \sum_{\mathbf{A}} P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T}). \quad (2)$$

On the other hand, the Bayesian solution returns the posterior $P(\mathbf{A}|\mathbf{E}, \mathbf{F})$, which is obtained from:

$$P(\mathbf{F}, \mathbf{A}|\mathbf{E}) = \int_{\mathbf{T}} P(\mathbf{T}) P(\mathbf{F}, \mathbf{A}|\mathbf{E}, \mathbf{T}). \quad (3)$$

2.1. EM

We used the GIZA++ [2] software to obtain the EM-estimated IBM Model 4 alignments. The default bootstrapping regimen was used, i.e., 5 iterations each of IBM Model 1 and HMM, followed by 3 iterations each of Models 3 and 4, in that order.

2.2. Gibbs sampling

It was shown in [5] that, compared to EM, Bayesian word alignment using Gibbs sampling (GS) reduces overfitting (e.g., high-fertility rare words), induces smaller models, and improves the BLEU score. In our system, we obtained two GS-inferred alignments; one for IBM Model 1 [5] and one for IBM Model 2 [7]. The following settings were common to both samplers:

- *Initialization:* The samplers were initialized with the EM-estimated Model 4 alignments obtained in 2.1.
- *Hyperparameters:* A sparse prior $P(\mathbf{T})$ was imposed on the translation parameters, specifically, a symmetric Dirichlet distribution with $\theta = 0.0001$.

- *Sample collection:* A total of 200 iterations of the sampler was run, with only the last 100 iterations used for Viterbi estimation (i.e., the burn-in period was 100 iterations).

For Bayesian Model 2, we used a uniform prior on the distortion parameters, specifically, a symmetric Dirichlet distribution with $\theta = 1$. We used relative distortion [8] for Model 2 in order to reduce the number of parameters.

2.3. Variational Bayes

Variational Bayes (VB) is a Bayesian inference method sometimes preferred over GS 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. Word alignment using Dirichlet priors and VB inference was investigated in [9, 10]. In our experiments, we used the publicly available software[1]. VB training was used in all models of the bootstrapping regimen for training IBM Model 4. As done in [9, 10], we set the Dirichlet hyperparameter $\theta = 0$ (the default setting) and ran 5 iterations of VB for each of IBM Model 1, HMM, Model 3 and Model 4[2].

2.4. Alignment Combination

We used the four different alignment methods explained above (EM with Model 4, GS with Models 1 and 2, and VB with Model 4) and combined the phrases extracted from before extracting phrases and estimating the phrase table probabilities. Our alignment combination method is similar to those previously used by others, e.g., [11]. The only change to the standard Moses training procedure is that we 4-fold replicated the training corpus, ran a different alignment method on each replica, and concatenated the obtained individual alignments. Alignments in each direction were further combined (symmetrized) using the default heuristic in Moses (grow-diag-final-and).

3. Data

Tables 1 and 2 present the main characteristics of the parallel corpora used in our experiments for translation model training. For the Arabic-English task, we utilized only the TED parallel corpus [12], while for the Turkish-English task, we utilized both the TED and SE Times parallel corpora.

We trained three separate language models from the English sides of the following parallel corpora (Table 3): the TED corpus (ted), the News Commentary corpus (nc), and the Gigaword French-English corpus (gigafren). The combination weights of these language models were optimized during the tuning step, together with the other log-linear model features.

[1] http://cs.rochester.edu/∼gildea/mt/giza-vb.tgz

[2] This is achieved by specifying the following options in the Moses training: model1tvb=1,modelhmmtvb=1,model3tvb=1,model4tvb=1.

Table 1: *Statistics of the parallel training data used in the Arabic-English experiments.*

Translation Model	Arabic	English
Sentences	136,729	
Tokens (M)	2.5	2.6
Types (k)	68.5	51.3
Singletons (k)	28.7	21.5

Table 2: *Statistics of the parallel training data used in the Turkish-English experiments.*

	TED		SETimes	
	Turkish	English	Turkish	English
Sentences	124,193		161,408	
Tokens (M)	1.8	2.4	3.9	4.4
Types (k)	153.9	47.3	135.9	66.6
Singletons (k)	87.6	19.6	66.2	29.8

Among the available development corpora, we used dev2010 for tuning and tst2010 for internal testing. We also present the experimental results for the tst2011 dataset, which was made available to the participants after the submission period.

Table 3: *Statistics of the language model training data.*

	ted	nc	gigafren
Tokens (M)	2.8	5.1	672
Unigrams (k)	53	69	2000

4. Common system features

Our submissions for both language pairs feature phrase-based statistical machine translation systems trained using the Moses toolkit [13]. Truecasing models were trained on tokenized training data, and subsequently all models were trained on truecased data. All language models were standard 4-gram models trained with modified Kneser-Ney discounting and interpolation using the SRILM toolkit [14]. The minimum error rate training (MERT) algorithm [15] with lattice sampling [16] and search in random directions [17] was used with BLEU [18] as the metric to be optimized. Evaluation was also performed using BLEU.

5. Arabic-English

5.1. Preprocessing

Arabic data was morphologically decomposed using MADA+TOKAN [19] with BAMA 2.0 (LDC2004L02) [20] and the default tokenization scheme. For English, the default tokenizer in the Moses package was used together with some post-processing. The final tokenization convention can be summarized as follows:

- Map unicode punctuation marks to ASCII.
- Merge and standardize consecutive hyphens and dots.
- Separate hyphens only if both sides are numbers (default in MADA+TOKAN).
- Merge back separated apostrophes.

Moreover, in order to reduce data sparsity in word alignment, all numbers were reduced to their last digits during training. For example, the tokens "60,000" and "2,000" were both replaced with "0".

5.2. Experiments

Table 4 compares the translation performance of the various alignment methods discussed in Section 2. For IBM Models 1 and 2, both Bayesian approaches (VB and GS) outperform EM. However, for Model 4, EM turned out to be better than VB[3]. The alignment combination described in Section 2.4 (last row in Table 4) did not provide the expected improvement, yielding a BLEU score somewhere between the highest and the lowest of the combined individual BLEU scores. Nevertheless, we chose it as our official submission for the Arabic-English track.

Table 4: *Performance of alignment inference schemes and their combination in the Arabic-English experiments.*

	Alignment		BLEU		
	Method	Model	dev10	tst10	tst11
1	EM	1	24.11	22.68	22.34
2	VB	1	24.34	23.21	**22.95**
3	GS	1	**24.59**	**23.22**	22.68
4	EM	2	24.33	22.65	22.37
5	VB	2	25.01	23.64	23.19
6	GS	2	**25.34**	**23.80**	**23.50**
7	EM	4	**25.48**	**23.83**	**23.93**
8	VB	4	25.09	23.71	23.28
9	(3)+(6)+(7)+(8)		25.01	23.58	23.13

Reducing model size was previously proposed as an objective in unsupervised word alignment, e.g., in [21, 22]. To see whether the Bayesian methods indeed achieve smaller models, we analyzed the outputs of each alignment method in terms of the total number of unique word translations in the produced alignments. Table 5 shows that both Bayesian methods induce significantly smaller alignment dictionaries than EM.

A contributing factor for the high dictionary size in ML-estimated alignments is that the rare source words in the training corpus are aligned to excessively many target words, also known as "garbage collection" [3]. To measure the effect of this phenomenon, the average fertility of singletons ($\tilde{\phi}_{sing}$) was used in [23] and [22]. We present $\tilde{\phi}_{sing}$ values in both alignment directions for the different alignment methods in Tables 6 and 7. We see that both Bayesian methods

[3]A Model-4 implementation of GS is not yet available

Table 5: *Number of distinct word translations (unique alignment pairs) induced by the alignment methods in the Arabic-English experiments.*

	Alignment		Dictionary Size (k)		
	Method	Model	en-ar	ar-en	sym.
1	EM	1	508	528	412
2	VB	1	182	187	258
3	GS	1	282	318	321
4	EM	2	558	548	659
5	VB	2	195	199	281
6	GS	2	289	317	395
7	EM	4	496	487	546
8	VB	4	207	218	292
9	(3)+(6)+(7)+(8)		743	771	821

dramatically reduce the average alignment fertility of singletons.

However, $\tilde{\phi}_{sing}$ can sometimes be misleading because a smaller value is not necessarily better. For example, the lowest possible value 0 can be trivially achieved by leaving all singletons unaligned, which is clearly not desirable. Tables 6 and 7 also show the ratio of unaligned singletons ($|sing0|/|sing|$)[4], which reveals that VB for Model 1 leaves nearly half of the singletons unaligned. The rightmost column in the table presents $\tilde{\phi}_{sing+}$, which averages the fertilities only over aligned singletons and has the minimum attainable value of 1.

6. Turkish-English

6.1. Preprocessing

For both languages, the default tokenizer in the Moses package was used, without any morphological processing.

6.2. Experiments

Our first system used a single phrase-table trained on the combined TED+SETimes corpus and used only VB (2.3) as the alignment inference method. Our second system used four different alignment methods as in our Arabic-English submission (Section 5), separately for each of the TED and the SETimes corpora, and then used the resulting two phrase tables in decoding. However, due to a bug at the time of the submission, the internal BLEU scores of this second system were significantly lower than our first system. Therefore, we submitted the first system as our primary submission.

Table 8 compares the BLEU scores of different alignment methods on the Turkish-English TED corpus. As opposed to the Arabic-English case, we observe in Table 8 that alignment combination provides a significant gain over the individual alignments.

[4]We further denote the aligned singletons by "$sing+$" so that $|sing| = |sing0| + |sing+|$.

Table 6: *Singleton alignment performance (en-ar) of the alignment methods in the Arabic-English experiments.*

Method	Model	$\tilde{\phi}_{sing}$	$\lvert sing0\rvert/\lvert sing\rvert$	$\tilde{\phi}_{sing+}$
EM	1	5.0	0.20	6.2
VB	1	0.8	0.47	1.6
GS	1	1.2	0.26	1.6
EM	2	3.7	0.001	3.7
VB	2	0.9	0.27	1.3
GS	2	1.1	0.23	1.4
EM	4	4.1	0.001	4.1
VB	4	1.3	0.08	1.5

Table 7: *Singleton alignment performance (ar-en) of the alignment methods in the Arabic-English experiments.*

Method	Model	$\tilde{\phi}_{sing}$	$\lvert sing0\rvert/\lvert sing\rvert$	$\tilde{\phi}_{sing+}$
EM	1	6.0	0.20	7.4
VB	1	0.9	0.46	1.6
GS	1	1.6	0.17	1.9
EM	2	4.4	0.001	4.4
VB	2	1.1	0.23	1.4
GS	2	1.4	0.16	1.7
EM	4	4.8	0.002	4.8
VB	4	1.4	0.08	1.5

7. Conclusion and Future Work

We described our submission to IWSLT 2012. The main innovation tested was using Bayesian word alignment methods (both variational Bayes and Gibbs sampling) in combination with the standard EM. As future work, we plan to apply the same technique on the MultiUN corpus for the Arabic-English task, and other larger corpora for other language pairs.

8. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[2] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[3] R. C. Moore, "Improving IBM word alignment Model 1," in *Proc. ACL*, Barcelona, Spain, July 2004, pp. 518–525.

[4] 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, New Jersey, 1993, pp. 202–205.

Table 8: *Performance of alignment inference schemes and their combination in the TED Turkish-English experiments.*

	Alignment		BLEU	
	Method	Model	dev10	tst10
1	EM	1	10.68	11.43
2	VB	1	**10.80**	11.87
3	GS	1	10.61	**12.10**
4	EM	2	10.21	11.67
5	VB	2	**11.16**	**11.92**
6	GS	2	10.68	11.47
7	EM	4	10.28	11.28
8	VB	4	**10.38**	**11.33**
9	(3)+(6)+(7)+(8)		**11.78**	**12.90**

[5] C. Mermer and M. Saraclar, "Bayesian word alignment for statistical machine translation," in *Proc. ACL-HLT: Short Papers*, Portland, Oregon, June 2011, pp. 182–187.

[6] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[7] C. Mermer, M. Saraclar, and R. Sarikaya, "Improving statistical machine translation using Bayesian word alignment and Gibbs sampling," *IEEE Transactions on Audio, Speech and Language Processing (in review)*, 2012.

[8] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Proc. COLING*, 1996, pp. 836–841.

[9] D. Riley and D. Gildea, "Improving the performance of GIZA++ using variational Bayes," The University of Rochester, Computer Science Department, Tech. Rep. 963, December 2010.

[10] ——, "Improving the IBM alignment models using variational Bayes," in *Proc. ACL: Short Papers*, 2012, pp. 306–310.

[11] W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2007 MT system," in *Proc. IWSLT*, Trento, Italy, 2007.

[12] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[13] 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, June 2007, pp. 177–180.

[14] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, vol. 3, 2002.

[15] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL*, Sapporo, Japan, July 2003, pp. 160–167.

[16] S. Chatterjee and N. Cancedda, "Minimum error rate training by sampling the translation lattice," in *Proc. EMNLP*, 2010, pp. 606–615.

[17] D. Cer, D. Jurafsky, and C. D. Manning, "Regularization and search for minimum error rate training," in *Proc. WMT*, 2008, pp. 26–34.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, Pennsylvania, July 2002, pp. 311–318.

[19] O. R. Nizar Habash and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proc. Second International Conference on Arabic Language Resources and Tools*, 2009.

[20] T. Buckwalter, "Buckwalter Arabic morphological analyzer version 2.0," *Linguistic Data Consortium*, 2004.

[21] T. Bodrumlu, K. Knight, and S. Ravi, "A new objective function for word alignment," in *Proc. NAACL-HLT Wk. Integer Linear Programming for Natural Language Processing*, Boulder, Colorado, June 2009, pp. 28–35.

[22] A. Vaswani, L. Huang, and D. Chiang, "Smaller alignment models for better translations: Unsupervised word alignment with the l0-norm," in *Proc. ACL*, 2012, pp. 311–319.

[23] C. Dyer, J. H. Clark, A. Lavie, and N. A. Smith, "Unsupervised word alignment with arbitrary features," in *Proc. ACL:HLT*, Portland, Oregon, June 2011, pp. 409–419.