Leave-one-out Word Alignment without Garbage Collector Effects

Xiaolin Wang **Masao Utiyama** **Andrew Finch**
Taro Watanabe* **Eiichiro Sumita**
Advanced Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan
{xiaolin.wang,mutiyama,andrew.finch,eiichiro.sumita}@nict.go.jp
tarow@google.com

Abstract

Expectation-maximization algorithms, such as those implemented in GIZA++ pervade the field of unsupervised word alignment. However, these algorithms have a problem of over-fitting, leading to "garbage collector effects," where rare words tend to be erroneously aligned to untranslated words. This paper proposes a leave-one-out expectation-maximization algorithm for unsupervised word alignment to address this problem. The proposed method excludes information derived from the alignment of a sentence pair from the alignment models used to align it. This prevents erroneous alignments within a sentence pair from supporting themselves. Experimental results on Chinese-English and Japanese-English corpora show that the F_1, precision and recall of alignment were consistently increased by 5.0% – 17.2%, and BLEU scores of end-to-end translation were raised by 0.03 – 1.30. The proposed method also outperformed l_0-normalized GIZA++ and Kneser-Ney smoothed GIZA++.

1 Introduction

Unsupervised word alignment (WA) on bilingual sentence pairs serves as an essential foundation for building most statistical machine translation (SMT) systems. A lot of methods have been proposed to raise the accuracy of WA in an effort to improve end-to-end translation quality. This paper contributes to this effort through refining the widely used expectation-maximization (EM) algorithm for WA (Dempster et al., 1977; Brown et al., 1993b; Och and Ney, 2000).

The EM algorithm for WA has a great influence in SMT. Many well-known toolkits including GIZA++ (Och and Ney, 2003), the Berkeley Aligner (Liang et al., 2006; DeNero and Klein, 2007), Fast Align (Dyer et al., 2013) and SyMGIZA++ (Junczys-Dowmunt and Sza, 2012), all employ this algorithm. GIZA++ in particular is frequently used in systems participating in many shared tasks (Goto et al., 2011; Cettolo et al., 2013; Bojar et al., 2013).

However, the EM algorithm for WA is well-known for introducing "garbage collector effects." Rare words have a tendency to collect garbage, that is they have a tendency to be erroneously aligned to untranslated words (Brown et al., 1993a; Moore, 2004; Ganchev et al., 2008; V Graça et al., 2010). Figure 1(a) shows a real sentence pair, denoted $\mathbf{s}$, from the GALE Chinese-English Word Alignment and Tagging Training corpus (GALE WA corpus)[1] with it's human-annotated word alignment. The Chinese word "HE ZHANG," denoted w_r, which means river custodian, only occurs once in the whole corpus. We performed EM training using GIZA++ on this corpus concatenated with 442,967 training sentence pairs from the NIST Open Machine Translation (OpenMT) 2006 evaluation[2]. The resulting alignment is shown in Figure 1(b). It can be seen that w_r is erroneously aligned to multiple English words.

To find the cause of this, we checked the alignments in each iteration i of $\mathbf{s}$, denoted $\mathbf{a_s^i}$. We found that in $\mathbf{a_s^1}$, w_r together with the other source-side words were aligned with uniform probability to all the target-side words since the alignment models provided no prior information. However, in $\mathbf{a_s^2}$, w_r became erroneously aligned,

* The author now is affiliated with Google, Japan.

[1] Released by Linguistic Data Consortium, catalog number LDC2012T16, LDC2012T20, LDC2012T24 and LDC2013T05.

[2] http://www.itl.nist.gov/iad/mig/tests/mt/2006/

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1817–1827,
Lisbon, Portugal, 17-21 September 2015.

because the alignment distribution[3] of w_r was only learned from $\mathbf{a_s^1}$, thus consisted of non-zero values only for generating the target-side words in $\mathbf{s}$. Therefore, the alignment probabilities from the rare word w_r to the unaligned words in $\mathbf{s}$ were extraordinarily high, since almost all of the probability mass was distributed among them. In other words, the story behind these garbage collector effects is that erroneous alignments are able to provide support for themselves; the probability distribution learned only from $\mathbf{s}$ is re-applied to $\mathbf{s}$. In this way, these "garbage collector effects" are a form of over-fitting.

Motivated by this observation, we propose a leave-one-out EM algorithm for WA in this paper. Recently this technique has been applied to avoid over-fitting in kernel density estimation (Roux and Bach, 2011); instead of performing maximum likelihood estimation, maximum leave-one-out likelihood estimation is performed. Figure 1(c) shows the effect of using our technique on the example. The garbage collection has not occurred, and the alignment of the word "HE ZHANG" is identical to the human annotation.

2 Related Work

The most related work to this paper is training phrase translation models with leave-one-out forced alignment (Wuebker et al., 2010; Wuebker et al., 2012). The differences are that their work operates at the phrase level, and their aim is to improve translation models; while our work operates at the word level, and our aim is to provide better word alignment. As word alignment is a foundation of most MT systems, our method have a wider application.

Recently, better estimation methods during the maximization step of EM have been proposed to avoid the over-fitting in WA, such as using Kneser-Ney Smoothing to back-off the expected counts (Zhang and Chiang, 2014) or integrating the smoothed l_0 prior to the estimation of probability (Vaswani et al., 2012). Our work differs from theirs by addressing the over-fitting directly in the EM algorithm by adopting a leave-one-out approach.

Bayesian methods (Gilks et al., 1996; Andrieu et al., 2003; DeNero et al., 2008; Neubig et al., 2011), also attempt to address the issue of over-fitting, however EM algorithms related to the proposed method have been shown to be more efficient (Wang et al., 2014).

[3]The probability distribution of generating target language words from w_r. The description here is only based on IBM model1 for simplicity, and the other alignment models are similar.

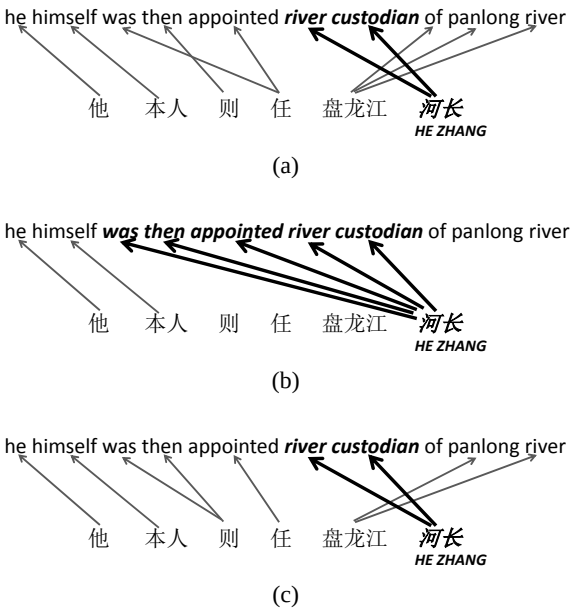


Figure 1: Examples of supervised word alignment. (a) gold alignment; (b) standard EM (GIZA++); (c) Leave-one-out alignment (proposed).

3 Methodology

This section first formulates the standard EM for WA, then presents the leave-one-out EM for WA, and finally briefly discusses handling singletons and effecient implementation. The main notation used in this section is shown in Table 1.

3.1 Standard EM for IBM Models 1, 2 and HMM Model

To perform WA through EM, the parallel corpus is taken as observed data, the alignments are taken as latent data. In order to maximize the likelihood of the alignment model θ given the data $\mathbf{S}$, the following two steps are conducted iteratively (Brown et al., 1993b; Och and Ney, 2000; Och and Ney, 2003),

Expectation Step (E step): calculating the conditional probability of alignments for each sentence pair,

$$P(\mathbf{a}|\mathbf{s},\theta) \quad = \prod_{j=1}^{J} \theta_{\text{ali}}(a_j|a_{j-1},I)\theta_{\text{lex}}(f_j|e_{a_j}) \quad (1)$$

where $\theta_{\text{ali}}(i|i',I)$ is the alignment probability and $\theta_{\text{lex}}(f|e)$ is the translation probability. Note that

$\mathbf{f}$	a foreign sentence $(f_1, \ldots, f_J)$
$\mathbf{e}$	an English sentence $(e_1, \ldots, e_I)$
$\mathbf{s}$	a sentence pair $(\mathbf{f}, \mathbf{e})$
$\mathbf{a}$	an alignment $(a_1, \ldots, a_J)$ where f_j is aligned to e_{a_j}
B_i	a list of the indexes of the foreign words which are aligned to e_i
$B_{i,k}$	the index of the k-th foreign word which is aligned to e_i
$\overline{B_i}$	is the average of all elements in B_i
ρ_i	the largest index of an English word $s.t.\ \rho_i < i$ and $\lvert B_{\rho_i}\rvert > 0$
ϕ_i	the fertility of e_i
E_i	the word class of e_i
$\theta_\cdot$	an probabilistic model
$\theta_\cdot^{\bar{\mathbf{s}}}$	a leave-one-out probabilistic model for $\mathbf{s}$
$n_x(\mathbf{s}, \mathbf{a})$	the number of times that an event x happens in $(\mathbf{s}, \mathbf{a})$
$N_x(\mathbf{s})$	the marginal number of times that an event x happens in $\mathbf{s}$

Table 1: Main Notation. Note that $N_x(\mathbf{s}) = \sum_{\mathbf{a}} n_x(\mathbf{s}, \mathbf{a}) P(\mathbf{a}|\mathbf{s})$. In practical calculation, for IBM models 1, 2 and HMM model, this summation is performed by dynamic programming; for IBM model 4, it is performed approximately using the best alignment and its neighbors.

(1) is a general form for IBM model 1, model 2 and the HMM model.

Maximization step (M step): re-estimating the probability models,

$$\theta_{\text{ali}}(i|i', I) \leftarrow \frac{\sum_{\mathbf{s}} N_{i|i',I}(\mathbf{s})}{\sum_{\mathbf{s}} N_{i',I}(\mathbf{s})} \quad (2)$$

$$\theta_{\text{lex}}(f|e) \leftarrow \frac{\sum_{\mathbf{s}} N_{f|e}(\mathbf{s})}{\sum_{\mathbf{s}} n_e(\mathbf{s})} \quad (3)$$

where $N_{i',I}(\mathbf{s})$ is the marginal number of times $e_{i'}$ is aligned to some foreign word if the length of $\mathbf{e}$ is I, or 0 otherwise; $N_{i|i',I}(\mathbf{s})$ is the marginal number of times the next alignment position after i' is i in $\mathbf{a}$ if the length of $\mathbf{e}$ is I, or 0 otherwise; $n_e(\mathbf{s})$ is the count of e in $\mathbf{e}$; $N_{f|e}(\mathbf{s}, \mathbf{a})$ is the marginal number of times e is aligned to f.

3.2 Leave-one-out EM for IBM Models 1, 2 and HMM Model

Leave-one-out EM for WA differs from standard EM in the way the alignment and translation probabilities are calculated. Each sentence pair will have its own alignment and translation probability models calculated by excluding the sentence pair itself. More formally, leave-one-out EM for WA are formulated as follows,

Leave-one-out E step: employing leave-one-out models for each $\mathbf{s}$ to calculate the conditional probability of alignments

$$P(\mathbf{a}|\mathbf{s}, \theta_{\bar{\mathbf{s}}}) = \prod_{j=1}^{J} \theta_{\text{ali}}^{\bar{\mathbf{s}}}(a_j|a_{j-1}, I) \theta_{\text{lex}}^{\bar{\mathbf{s}}}(f_j|e_{a_j}) \quad (4)$$

where $\theta_{\text{ali}}^{\bar{\mathbf{s}}}(i|i', I)$ and $\theta_{\text{lex}}^{\bar{\mathbf{s}}}(f_j|e_{a_j})$ are the leave-one-out alignment probability and translation probability, respectively.

Leave-one-out M step: re-estimating leave-one-out probability models,

$$\theta_{\text{ali}}^{\bar{\mathbf{s}}}(i|i', I) \leftarrow \frac{\sum_{\mathbf{s}' \neq \mathbf{s}} N_{i|i',I}(\mathbf{s}')}{\sum_{\mathbf{s}' \neq \mathbf{s}} N_{i',I}(\mathbf{s}')} \quad (5)$$

$$\theta_{\text{lex}}^{\bar{\mathbf{s}}}(f|e) \leftarrow \frac{\sum_{\mathbf{s}' \neq \mathbf{s}} N_{f|e}(\mathbf{s}')}{\sum_{\mathbf{s}' \neq \mathbf{s}} n_e(\mathbf{s}')}. \quad (6)$$

3.3 Standard EM for IBM Model 4

The framework of the standard EM for IBM Model 4 is similar with the one for IBM Models 1, 2 and HMM Model, but the calculation of alignment probability is more complicated.

E step: calculating the conditional probability through the reverted alignment (Och and Ney, 2003),

$$P(\mathbf{a}|\mathbf{s}, \theta) = P(B_0|B_1, \ldots, B_I) \cdot \prod_{i=1}^{I} P(B_i|B_{i-1}, e_i) \cdot \prod_{i=1}^{I} \prod_{j \in B_i} \theta_{\text{lex}}(f_j|e_i), \quad (7)$$

where B_0 means the set of foreign words aligned with the empty word; $P(B_0|B_1, \ldots, B_I)$ is assumed to be a binomial distribution for the size of B_0 (Brown et al., 1993b) or an modified distribution to relieve deficiency (Och and Ney, 2003).

The distribution $P(B_i|B_{i-1}, e_i)$ is decomposed as

$$P(B_i|B_{i-1}, e_i) = \theta_{\text{fer}}(\phi_i|e_i) \cdot \theta_{\text{hea}}(B_{i,1} - \overline{B_{\rho_i}}|E_{\rho_i}) \cdot \prod_{k=2}^{\phi_i} \theta_{\text{oth}}(B_{i,k} - B_{i,k-1}), \quad (8)$$

where θ_{fer} is a fertility model; θ_{hea} is a probability model for the head (first) aligned foreign word; θ_{oth} is a probability model for the other aligned foreign words. θ_{hea} is assumed to be conditioned

on the word class E_{ρ_i}, following the paper of (Och and Ney, 2003) and the implementation of GIZA++ and CICADA.

M step: re-estimating the probability models,

$$\theta_{\text{fer}}(\phi|e) \leftarrow \frac{\sum_{\mathbf{s}} N_{\phi|e}(\mathbf{s})}{\sum_{\mathbf{s}} \sum_{\phi'} N_{\phi'|e}(\mathbf{s})} \quad (9)$$

$$\theta_{\text{hea}}(\Delta i|E) \leftarrow \frac{\sum_{\mathbf{s}} N^{\text{hea}}_{\Delta i|E}(\mathbf{s})}{\sum_{\mathbf{s}} \sum_{\Delta i'} N^{\text{hea}}_{\Delta i'|E}(\mathbf{s})} \quad (10)$$

$$\theta_{\text{oth}}(\Delta i) \leftarrow \frac{\sum_{\mathbf{s}} N^{\text{oth}}_{\Delta i}(\mathbf{s})}{\sum_{\mathbf{s}} \sum_{\Delta i'} N^{\text{oth}}_{\Delta i'}(\mathbf{s})}, \quad (11)$$

where Δi is a difference of the indexes of two foreign words.

3.4 Leave-one-out EM for IBM Model 4

The leave-one-out treatment were applied to the three component probability models θ_{fer}, θ_{hea} and θ_{oth} of IBM model 4.

Leave-one-out E step: calculating the conditional probability through leave-one-out probability models

$$P(\mathbf{a}|\mathbf{s}, \theta^{\bar{\mathbf{s}}}) = P(B_0|B_1, \ldots, B_I) \cdot \prod_{i=1}^{I} P^{\bar{\mathbf{s}}}(B_i|B_{i-1}, e_i) \cdot \prod_{i=1}^{I} \prod_{j \in B_i} \theta^{\bar{\mathbf{s}}}_{\text{lex}}(f_j|e_i), \quad (12)$$

$$P^{\bar{\mathbf{s}}}(B_i|B_{i-1}, e_i) = \theta^{\bar{\mathbf{s}}}_{\text{fer}}(\phi_i|e_i) \cdot \theta^{\bar{\mathbf{s}}}_{\text{hea}}(B_{i,1} - \overline{B_{\rho_i}}|E_{\rho_i}) \cdot \prod_{k=2}^{\phi_i} \theta^{\bar{\mathbf{s}}}_{\text{oth}}(B_{i,k} - B_{i,k-1}). \quad (13)$$

Leave-one-out M step: re-estimating the leave-one-out probability models,

$$\theta^{\bar{\mathbf{s}}}_{\text{fer}}(\phi|e) \leftarrow \frac{\sum_{\mathbf{s}' \neq \mathbf{s}} N_{\phi|e}(\mathbf{s}')}{\sum_{\mathbf{s}' \neq \mathbf{s}} \sum_{\phi'} N_{\phi'|e}(\mathbf{s}')} \quad (14)$$

$$\theta^{\bar{\mathbf{s}}}_{\text{hea}}(\Delta i|E) \leftarrow \frac{\sum_{\mathbf{s}' \neq \mathbf{s}} N^{\text{hea}}_{\Delta i|E}(\mathbf{s}')}{\sum_{\mathbf{s}' \neq \mathbf{s}} \sum_{\Delta i'} N^{\text{hea}}_{\Delta i'|E}(\mathbf{s}')} \quad (15)$$

$$\theta^{\bar{\mathbf{s}}}_{\text{oth}}(\Delta i) \leftarrow \frac{\sum_{\mathbf{s}' \neq \mathbf{s}} N^{\text{oth}}_{\Delta i}(\mathbf{s}')}{\sum_{\mathbf{s}' \neq \mathbf{s}} \sum_{\Delta i'} N^{\text{oth}}_{\Delta i'}(\mathbf{s}')}. \quad (16)$$

3.5 Handling Singletons

Singletons are the words that occur only once in corpora. Singletons cause problems when applying leave-one-out to lexicalized models such as the translation model $\theta^{\bar{\mathbf{s}}}_{\text{lex}}$ and the fertility model $\theta^{\bar{\mathbf{s}}}_{\text{fer}}$. When calculating (6) and (14) for singletons, the denominators become zero, thus the probabilities are undefined.

For singletons, there is no prior information to guide their alignment, so we back off to uniform distributions. In that case, the alignments are primarily determined by the rest of the sentence.

In addition, singletons can be in the target side of the translation model $\theta^{\bar{\mathbf{s}}}_{\text{lex}}$. In that case, the probabilities become zero. This is handled by setting a minimum probability value of 1.0×10^{-12}, which was decided by pilot experiments.

3.6 Implementation Details

To alleviate memory requirements and increase speed, our implementation did not build or store the local alignment models explicitly for each sentence pair. The following formula was used to efficiently calculate (5), (6) and (14–16) to build temporary probability models,

$$\sum_{\mathbf{s}' \neq \mathbf{s}} N_x(\mathbf{s}') = (\sum_{\mathbf{s}'} N_x(\mathbf{s}')) - N_x(\mathbf{s}), \quad (17)$$

where x is a alignment event. Our implementation maintained global counts of all alignment events $\sum_{\mathbf{s}'} N_x(\mathbf{s}')$, and (considerably smaller) local counts $N_x(\mathbf{s})$ from each sentence pair $\mathbf{s}$.

Take the translation model $\theta^{\bar{\mathbf{s}}}_{\text{lex}}$ for example. For a sentence pair $\mathbf{s} = (f_1 \ldots f_J, e_1 \ldots e_I)$, it is caculated as,

$$\theta^{\bar{\mathbf{s}}}_{\text{lex}}(f_j|e_i) = \frac{(\sum_{\mathbf{s}'} N_{(f_j|e_i)}(\mathbf{s}')) - N_{(f_j|e_i)}(\mathbf{s})}{(\sum_{\mathbf{s}'} n_{e_i}(\mathbf{s}')) - n_{e_i}(\mathbf{s})}. \quad (18)$$

The global counts to be maintained are $\sum_{\mathbf{s}'} N_{(f_j|e_i)}(\mathbf{s}')$ and $n_{e_i}(\mathbf{s}')$, and the local counts are $\sum_{\mathbf{s}} N_{(f_j|e_i)}(\mathbf{s})$ and $n_{e_i}(\mathbf{s})$. Therefore the memory cost is,

$$|\mathcal{E}| \cdot (|\mathcal{F}| + 1) + \sum_{\mathbf{s}} I_{\mathbf{s}}(J_{\mathbf{s}} + 1), \quad (19)$$

where $|\mathcal{E}|$ is the size of English vocabulary, $|\mathcal{F}|$ is the size of foreign language vocabulary, $I_{\mathbf{s}}$ is the length of the English sentence of $\mathbf{s}$, and $J_{\mathbf{s}}$ is the length of the foreign sentence of $\mathbf{s}$.

The calculation of the leave-one-out translation model is performed for each English word and foreign word in $\mathbf{s}$. Therefore, the time cost is,

$$\sum_{\mathbf{s}} I_{\mathbf{s}}(J_{\mathbf{s}} + 1). \quad (20)$$

In addition, because the local counts $N_{(f_j|e_i)}(\mathbf{s})$ and $n_{e_i}(\mathbf{s})$ are read in order, storing them in a external memory such as a hard disk will not slow down the running speed much. This will reduce the memory cost to

$$|\mathcal{E}| \cdot (|\mathcal{F}| + 1). \quad (21)$$

This cost is independent to the number of sentence pairs[4].

The speed of the proposed method can be boosted through parallelism. These calculations on each sentence pair can be performed independently. We found empirically that when our implementation of the proposed method is run on a 16-core computer, it finishes the task earlier than GIZA++[5].

4 Experiments

The proposed WA method was tested on two language pairs: Chinese-English and Japanese-English (Table 2). Performance was measured both directly using the agreement with reference to manual WA annotations, and indirectly using the BLEU score in end-to-end machine translation tasks. GIZA++ and our own implementation of standard EM were used as baselines.

4.1 Experimental Settings

The Chinese-English experimental data consisted of the GALE WA corpus and the OpenMT corpus. They are from the same domain, both contain newswire texts and web blogs. The OpenMT evaluation 2005 was used as a development set for MERT tuning (Och, 2003), and the OpenMT evaluation 2006 was used as a test set. The Japanese-English experimental data was the Kyoto Free Translation Task (Neubig, 2011)[6]. The corpus contains a set of 1,235 sentence pairs that are manually word aligned.

The corpora were processed using a standard procedure for machine translation. The English texts were tokenized with the tokenization script released with Europarl corpus (Koehn, 2005) and converted to lowercase; the Chinese texts were segmented into words using the Stanford Word Segmenter (Xue et al., 2002)[7]; the Japanese texts were segmented into words using the Kyoto Text Analysis Toolkit (KyTea[8]). Sentences longer than 100 words or those with foreign/English word length ratios between larger than 9 were filtered out.

GIZA++ was run with the default Moses settings (Koehn et al., 2007). The IBM model 1, HMM model, IBM model 3 and IBM model 4 were run with 5, 5, 3 and 3 iterations. We implemented the proposed leave-one-out EM and standard EM in IBM model 1, HMM model and IBM model 4. In the original work (Och and Ney, 2003) this combination of models achieved comparable performance to the default Moses settings. They were run with 5, 5 and 6 iterations.

The standard EM was re-implemented as a baseline to provide a solid basis for comparison, because GIZA++ contains many undocumented details. Our implementation is based on the toolkit of CICADA (Watanabe and Sumita, 2011; Watanabe, 2012; Tamura et al., 2013)[9]. We named the implemented aligner AGRIPPA, to support our in-house decoders OCTAVIAN and AUGUSTUS.

In all experiments, WA was performed independently in two directions: from foreign languages to English, and from English to foreign languages. Then the grow-diag-final-and heuristic was used to combine the two alignments from both directions to yield the final alignments for evaluation (Och and Ney, 2000; Och and Ney, 2003).

4.2 Word Alignment Accuracy

Word alignment accuracy of the baseline and the proposed method is shown in Table 3 in terms of precision, recall and F_1 (Och and Ney, 2003). The proposed method gave rise to higher quality alignments in all our experiments. The improvement in F_1, precision and recall based on IBM Model 4 is in the range 8.3% to 9.1% compared with the GIZA++ baseline, and in the range 5.0% to 17.2% compared with our own baseline.

The most meaningful result comes from the comparison of the models trained using standard EM log-likelihood training, and the proposed EM leave-one-out log-likelihood training. These models are identical except for way in which the model likelihood is calculated. In all our experiments the proposed method gave rise to higher quality alignments. The standard EM implementation achieved

[4] We found the memory of our server is large enough, so we did not implement it

[5] We plan to make our code public available.

[6] http://www.phontron.com/kftt/

[7] http://nlp.stanford.edu/software/segmenter.shtml

[8] http://www.phontron.com/kytea/

[9] http://www2.nict.go.jp/univ-com/multi_trans/cicada/

Corpus	# Sent. pairs	# Foreign Words	# English Words
Chinese-English (GALE WA, OpenMT)			
WA	18,057	392,447	518,137
Train	442,967	12,265,072	13,444,927
Eval. 05	1,082†	29,688	138,952
Eval. 06	1,664†	37,827	189,059
Japanese-English (Kyoto Free Translation)			
WA	1,235	34,403	30,822
Train	329,882	6,085,131	5,911,486
Develop	1,166	26,856	24,309
Test	1,160	28,501	26,734

Table 2: Experimental Data. † Each consists of one foreign sentence and four English reference sentences.

Models	standard EM (GIZA++)			standard EM (ours)			Leave-one-out(prop.)		
	F_1	P	R	F_1	P	R	F_1	P	R
Chinese-English (GALE WA, OpenMT)									
Model 1	0.498	0.656	0.401	0.518	0.670	0.423	**0.553**	**0.689**	**0.461**
HMM	0.584	0.720	0.491	0.593	0.722	0.503	**0.665**	**0.774**	**0.583**
Model 4	0.624	0.698	0.565	0.593	0.688	0.522	**0.677**	**0.756**	**0.612**
Japanese-English (Kyoto Free Translation)									
Model 1	0.508	0.601	0.439	0.513	0.606	0.444	**0.535**	**0.618**	**0.471**
HMM	0.573	0.667	0.502	0.579	0.665	0.512	**0.626**	**0.687**	**0.575**
Model 4	0.577	0.594	0.561	0.570	0.617	0.530	**0.628**	**0.648**	**0.609**

Table 3: Word alignment accuracy measured by F_1, precision and recall.

alignment performance approximately comparable to GIZA++, whereas the proposed method exceeded the performance of both implementations.

4.3 End-to-end Translation Quality

BLEU scores achieved by the phrase-based and hierachical SMT systems[10] which were trained from different alignment results, are shown in Table 4. Each experiment was conducted three times to mitigate the variance in the results due to MERT. The results show that the proposed alignment method achieved the highest BLEU score in all experiments. The improvement over the baseline is in range 0.03 to 1.03 for phrase-based systems, and ranged from 0.43 to 1.30 for hierarchical systems.

Hierarchical systems benifit more from the proposed method than phrase-based systems. We think this is because that hierarchical systems are more sensitive to word alignment quality than phrase-based systems. Phrase-based systems only

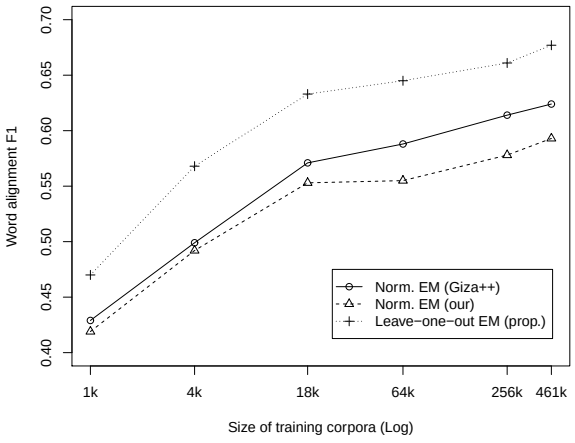


Figure 2: Curve of word alignment accuracy (F_1) under training corpora of different sizes.

[10] from the Moses toolkit

SMT Systems	standard EM (GIZA++)	standard EM (ours)	Leave-one-out (prop.)
Chinese-English (GALE WA, OpenMT)			
Phrase-based	31.85 ± 0.26	31.01 ± 0.18	$\mathbf{32.04 \pm 0.08}$
Hierarchical	32.27 ± 0.23	31.40 ± 0.26	$\mathbf{32.70 \pm 0.14}$
Japanese-English (Kyoto Free Translation)			
Phrase-based	18.35 ± 0.27	18.20 ± 0.20	$\mathbf{18.38 \pm 0.11}$
Hierarchical	19.48 ± 0.08	19.39 ± 0.02	$\mathbf{20.10 \pm 0.07}$

Table 4: End-to-end translation quality measured by BLEU

Corpus size	standard EM (GIZA++)			standard EM (ours)			Leave-one-out(prop.)		
	F_1	P	R	F_1	P	R	F_1	P	R
1K	0.429	0.466	0.397	0.419	0.463	0.382	**0.470**	**0.568**	**0.402**
4K	0.499	0.547	0.459	0.492	0.549	0.445	**0.568**	**0.668**	**0.494**
18K†	0.571	0.630	0.521	0.553	0.621	0.499	**0.633**	**0.721**	**0.565**
64K	0.588	0.659	0.531	0.555	0.638	0.492	**0.645**	**0.712**	**0.590**
256K	0.614	0.687	0.554	0.578	0.667	0.511	**0.661**	**0.718**	**0.612**
461K	0.624	0.698	0.565	0.593	0.688	0.522	**0.677**	**0.756**	**0.612**

Table 5: Effect of training corpus size on word alignment accuracy measured by F_1, precision and recall (Chinese-English). † the whole manually word aligned corpus

Corpus size	stan.(GIZA++)	stan.(ours)	LOO(prop.)	Gold
Phrase-based				
1k	7.86	7.66	9.38	**10.01**
4k	15.27	15.49	17.06	**17.57**
18K†	22.15	21.72	**24.41**	24.11
64K	28.10	27.91	**29.23**	NA
256K	31.05	30.82	**31.51**	NA
461K	31.85	31.01	**32.04**	NA
Hierarchical				
1k	7.53	7.54	9.19	**10.62**
4k	14.89	15.51	17.91	**18.31**
18K†	22.85	22.56	**24.66**	24.52
64K	28.82	28.22	**29.78**	NA
256K	31.47	30.21	**31.72**	NA
461K	32.27	31.04	**32.70**	NA

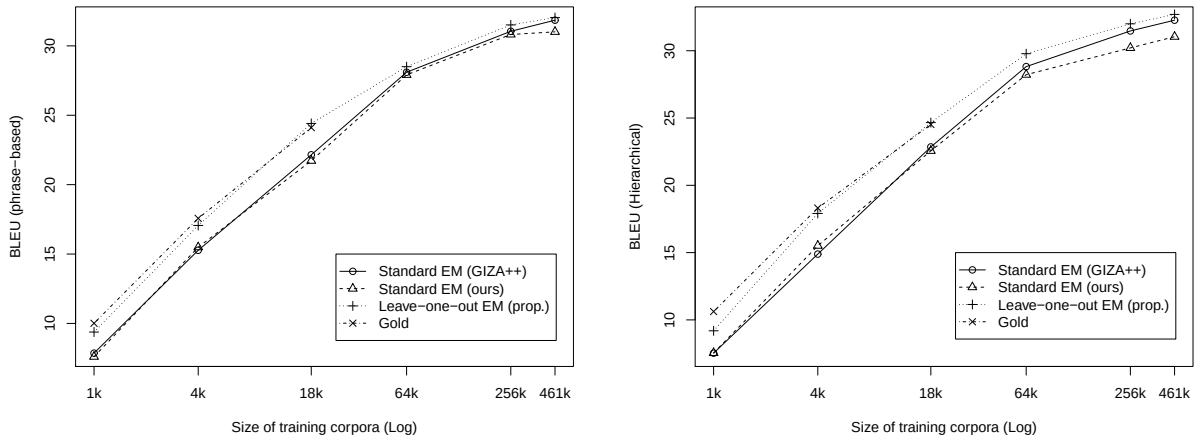
Table 6: Effect of training corpus size on end-to-end translation quality measured by BLEU (Chinese-English). † the whole manually word aligned corpus

take contiguous parallel phrase pairs as translation rules, while hierarchical systems also use patterns made by subtracting (inner) short parallel phrases from (outer) longer parallel phrases. Both the outer and inner phrases typically need to be noise-free in order to produce high quality rules. This puts a high demand on the alignment quality.

4.4 Effect of Training Corpus Size

Training corpora of different sizes were employed to perform unsupervised WA experiments and MT experiments (see Tables 5 and 6).

The training corpora were randomly sampled from the Chinese-English manual WA corpora and the parallel training corpus. The manual WA corpus has a priority for being sampled so that the gold WA annotation is available for MT experi-



(a) (b)

Figure 3: Curves of translation quality (BLEU) under training corpora of different sizes. (a) Phrase-based MT; (b) Hierarchical MT.

ments.

The settings of the unsupervised WA experiments and the MT experiments are the same with the previous experiments. In the WA experiments, GIZA++, our implemented standard EM and the proposed leave-one-out EM are applied to training corpora with the same parameter settings as the previous. In the MT experiments, the WA results of different methods and the gold WA (if available) are employed to extract translation rules; the rest settings including language models, development and test corpus, and parameters are the same as the previous.

On word alignment accuracy, the proposed method achieved improvements of F_1 from 0.041 to 0.090 under the different training corpora (Table 5. The maximum improvement compared with GIZA++ is 0.069 when the training corpus has 4,000 sentence pairs. The maximum improvement compared with our own implement is 0.090 when the training corpus has 64,000 sentence pairs.

Figure 2 shows that the extent of improvements slightly changes under different training corpora, but they are all quite stable and obvious.

On translation quality, the proposed method achieved improvements of BLEU under the different training corpora. The improvements ranged from 0.19 to 1.72 for phrase-based MT and ranged from 0.25 to 3.02 (see Table 5). The improvements are larger under smaller training corpora (see Figure 3).

In addition, the BLEUs achieved by the proposed method is close to the ones achieved by gold WA annotations. The proposed method slightly outperforms the gold WA annotations when using the full manual WA corpus of 18,057 sentence pairs.

4.5 Comparison to l_0-Normalization and Kneser-Ney Smoothing Methods

The proposed leave-one-word word alignment method was empirically compared to l_0-normalized GIZA++ (Vaswani et al., 2012)[11] and Kneser-Ney smoothed GIZA++ (Zhang and Chiang, 2014)[12]. l_0-normalization and Kneser-Ney smoothing methods are established methods to overcome the sparse problem. This enables the probability distributions on rare words to be estimated more effectively. In this way, these two GIZA++ variants are related to the proposed method.

l_0-normalized GIZA++ and Kneser-Ney smoothed GIZA++ were run with the same settings as GIZA++, which came from the default settings of MOSES. For the settings of l_0-normalized GIZA++ that are not in common with GIZA++ were the default settings. As for Kneser-Ney smoothed GIZA++, the smooth switches of IBM models 1 – 4 and HMM model

[11] http://www.isi.edu/~avaswani/giza-pp-l0.html

[12] https://github.com/hznlp/giza-kn

	GIZA++			l_0-Normalization			Kneser-Ney Smooth.			Leave-one-out(prop.)		
Word Alignment Quality												
	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R
All Words	0.624	0.698	0.565	0.629	0.700	0.571	0.656	0.726	0.599	**0.678**	**0.755**	**0.615**
S.W.F=1	0.458	0.435	0.483	0.448	0.471	0.427	**0.515**	0.532	**0.499**	0.398	**0.693**	0.279
S.W.F$\leq$2	0.466	0.451	0.481	0.461	0.485	0.440	**0.522**	0.545	**0.501**	0.450	**0.707**	0.330
S.W.F$\leq$5	0.476	0.480	0.473	0.478	0.509	0.451	**0.534**	0.572	**0.501**	0.502	**0.722**	0.385
S.W.F$\leq$10	0.485	0.505	0.466	0.491	0.531	0.456	**0.541**	0.593	**0.498**	0.529	**0.733**	0.414
Translation Quality (BLEU)												
Phrase-based	31.85 ± 0.26			31.52 ± 0.06			31.94 ± 0.19			**32.04 ± 0.08**		
Hierarchical	32.27 ± 0.23			32.20 ± 0.04			32.47 ± 0.33			**32.70 ± 0.14**		

Table 7: Empirical Comparision with l_0-Normalized and Kneser-Ney Smoothed GIZA++'s

were turned on.

The experimental results are presented in Table 7. The experiments were run on the Chinese-English language pair. The word alignment quality was evaluated separately for all words and for various levels of rare words. The leave-one-out method outperformed related methods in terms of precision, recall and F_1 when evaluated on all words.

Rare words were categorized based on the number of occurences in the source-language text of the training data. The evaluations were carried out on the subset of alignment links that had a rare word on the source side. Table 7 presents the results for thresholds 1, 2, 5 and 10. The proposed method achieved much higher precision on rare words than the other methods, but performed poorly on recall. The Kneser-Ney Smoothed GIZA++ had higher recall. The explanation might be that the leave-one-out method punishes rare words more than the Kneser-Ney smoothing method, by totally removing the derived expected counts of current sentence pair from the alignment models. This leads to rare words being passively aligned. In other words, the leave-one-out method would align rare words unless the confidence is high. Therefore, we plan to seek a method to integrate Kneser-Ney smoothing into the proposed leave-one-out method in the future work.

The BLEU scores achieved by phrase-based SMT and hierarchical SMT for different alignment methods are presented in Table 7. The proposed method outperforms the other methods. The Kneser-Ney Smoothed GIZA++ performed the second best. We tried to further analyze the relation between word alignment and BLEU, but found the analysis was obscured by the many processing stages. These stages include parallel phrase extraction (or translation rule extraction from hierarchical SMT), log-linear model, MERT tuning and practical decoding where a lot of pruning happened.

5 Conclusion

This paper proposes a leave-one-out EM algorithm for WA to overcome the over-fitting problem that occurs when using standard EM for WA. The experimental results on Chinese-English and Japanese-English corpora show that both the WA accuracy and the end-to-end translation are improved.

In addition, we have a interesting finding about the effect of manual WA annotations on training MT systems. In a Chinese-English parallel training corpus of 18,057 sentence pairs, the manual WA annotation outperformed the unsupervised WA results produced by standard EM algorithms. However, the unsupervised WA results produced by proposed leave-one-out EM algorithm outperformed the manual WA annotation.

Our future work will focus on increasing the gains in end-to-end translation quality through the proposed leave-one-out aligner. It is a interesting question why GIZA++ achieved competitive BLEU scores though its alignment accuracy measured by F_1 was substantially lower. The answer to this question which may reveal essence of good word alignment for MT and eventually help to improve MT. In addition, we plan to improve the proposed method by integrating Kneser-Ney smoothing.

Acknowledgments

We appreciated the valuable comments from the reviewers.

References

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993a. But dictionaries are data too. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 29–38.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics*, pages 17–24.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a bayesian translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 314–323, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Kuzman Ganchev, Joao V. Graça, and Ben Taskar. 2008. Better alignments = better translations? *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics*, page 42.

Walter R. Gilks, Sylvia Richardson, and David J. Spiegelhalter. 1996. *Markov chain Monte Carlo in practice*, volume 2. CRC press.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578.

Marcin Junczys-Dowmunt and Arkadiusz Sza. 2012. Symgiza++: Symmetrized word alignment models for machine translation. In Pascal Bouvry, Mieczyslaw A. Klopotek, Franck Leprvost, Malgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybinski, editors, *Security and Intelligent Information Systems (SIIS)*, volume 7053 of *Lecture Notes in Computer Science*, pages 379–390, Warsaw, Poland. Springer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies*, pages 104–111. Association for Computational Linguistics.

Robert C. Moore. 2004. Improving IBM word-alignment model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 518. Association for Computational Linguistics.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *ACL*, pages 632–641.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1086–1090. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Nicolas Le Roux and Francis Bach. 2011. Local component analysis. Technical report.

Akihiro Tamura, Taro Watanabe, Eiichiro Sumita, Hiroya Takamura, and Manabu Okumura. 2013. Part-of-speech induction in dependency trees for statistical machine translation. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pages 841–851.

João V Graça, Kuzman Ganchev, and Ben Taskar. 2010. Learning tractable word alignment models with complex constraints. *Computational Linguistics*, 36(3):481–504.

Ashish Vaswani, Liang Huang, and David Chiang. 2012. Smaller alignment models for better translations: unsupervised word alignment with the l_0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 311–319. Association for Computational Linguistics.

Xiaolin Wang, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014. Empirical study of unsupervised chinese word segmentation methods for smt on large-scale corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–758, Baltimore, Maryland, June. Association for Computational Linguistics.

Taro Watanabe and Eiichiro Sumita. 2011. Machine translation system combination by confusion forest. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1249–1257. Association for Computational Linguistics.

Taro Watanabe. 2012. Optimized online rank learning for machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 253–262. Association for Computational Linguistics.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 475–484. Association for Computational Linguistics.

Joern Wuebker, Mei-Yuh Hwang, and Chris Quirk. 2012. Leave-one-out phrase model training for large-scale deployment. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 460–467. Association for Computational Linguistics.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Hui Zhang and David Chiang. 2014. Kneser-ney smoothing on expected counts. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 765–774, Baltimore, Maryland, June. Association for Computational Linguistics.