Semi-Supervised Training for Statistical Word Alignment

Alexander Fraser
ISI / University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
fraser@isi.edu

Daniel Marcu
ISI / University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

Abstract

We introduce a semi-supervised approach to training for statistical machine translation that alternates the traditional Expectation Maximization step that is applied on a large training corpus with a discriminative step aimed at increasing word-alignment quality on a small, manually word-aligned sub-corpus. We show that our algorithm leads not only to improved alignments but also to machine translation outputs of higher quality.

1 Introduction

The most widely applied training procedure for statistical machine translation — IBM model 4 (Brown et al., 1993) unsupervised training followed by post-processing with symmetrization heuristics (Och and Ney, 2003) — yields low quality word alignments. When compared with gold standard parallel data which was manually aligned using a high-recall/precision methodology (Melamed, 1998), the word-level alignments produced automatically have an F-measure accuracy of 64.6 and 76.4% (see Section 2 for details).

In this paper, we improve word alignment and, subsequently, MT accuracy by developing a range of increasingly sophisticated methods:

1. We first recast the problem of estimating the IBM models (Brown et al., 1993) in a discriminative framework, which leads to an initial increase in word-alignment accuracy.
2. We extend the IBM models with new (sub)models, which leads to additional increases in word-alignment accuracy. In the process, we also show that these improvements are explained not only by the power of the new models, but also by a novel search procedure for the alignment of highest probability.
3. Finally, we propose a training procedure that interleaves discriminative training with maximum likelihood training.

These steps lead to word alignments of higher accuracy which, in our case, correlate with higher MT accuracy.

The rest of the paper is organized as follows. In Section 2, we review the data sets we use to validate experimentally our algorithms and the associated baselines. In Section 3, we present iteratively our contributions that eventually lead to absolute increases in alignment quality of 4.8% for French/English and 4.8% for Arabic/English, as measured using F-measure for large word alignment tasks. These contributions pertain to the casting of the training procedure in the discriminative framework (Section 3.1); the IBM model extensions and modified search procedure for the Viterbi alignments (Section 3.2); and the interleaved, minimum error/maximum likelihood, training algorithm (Section 4). In Section 5, we assess the impact that our improved alignments have on MT quality. We conclude with a comparison of our work with previous research on discriminative training for word alignment and a short discussion of semi-supervised learning.

2 Data Sets and Baseline

We conduct experiments on alignment and translation tasks using Arabic/English and French/English data sets (see Table 1 for details). Both sets have training data and two gold standard word alignments for small samples of the training data, which we use as the alignment

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 769–776,
Sydney, July 2006.

		Arabic/English		French/English	
		A	E	F	E
Training	Sents	3,713,753		2,842,184	
	Words	102,473,086	119,994,972	75,794,254	67,366,819
	Vocab	489,534	231,255	149,568	114,907
	Singletons	199,749	104,155	60,651	47,765
Align Discr.	Sents	100		110	
	Words	1,712	2,010	1,888	1,726
	Links	2,129		2,292	
Align Test	Sents	55		110	
	Words	1,004	1,210	1,899	1,716
	Links	1,368		2,176	
Max BLEU	Sents	728 (4 references)		833 (1 reference)	
	Words	17664	22.0K to 24.5K	20,562	17,454
Trans. Test	Sents	663 (4 references)		2,380 (1 reference)	
	Words	16,075	19.0K to 21.6K	58,990	49,182

Table 1: Datasets

System	F-measure F to E	F-measure E to F	F-measure Best Symm.
A/E Model 4: Iteration 4	65.6 / 60.5	53.6 / 50.2	69.1 / 64.6 (union)
F/E Model 4: Iteration 4	73.8 / 75.1	74.2 / 73.5	76.5 / 76.4 (refined)

Table 2: Baseline Results. F-measures are presented on both the alignment discriminative training set and the alignment test set sub-corpora, separated by /.

discriminative training set and alignment test set. Translation quality is evaluated by translating a held-out translation test set. An additional translation set called the Maximum BLEU set is employed by the SMT system to train the weights associated with the components of its log-linear model (Och, 2003).

The training corpora are publicly available: both the Arabic/English data and the French/English Hansards were released by LDC. We created the manual word alignments ourselves, following the Blinker guidelines (Melamed, 1998).

To train our baseline systems we follow a standard procedure. The models were trained two times, first using French or Arabic as the source language and then using English as the source language. For each training direction, we run GIZA++ (Och and Ney, 2003), specifying 5 iterations of Model 1, 4 iterations of the HMM model (Vogel et al., 1996), and 4 iterations of Model 4. We quantify the quality of the resulting hypothesized alignments with F-measure using the manually aligned sets.

We present the results for three different conditions in Table 2. For the "F to E" direction the models assign non-zero probability to alignments consisting of links from one Foreign word to zero or more English words, while for "E to F" the models assign non-zero probability to alignments consisting of links from one English word to zero or more Foreign words. It is standard practice to improve the final alignments by combining the "F to E" and "E to F" directions using symmetrization heuristics. We use the "union", "refined" and "intersection" heuristics defined in (Och and Ney, 2003) which are used in conjunction with IBM Model 4 as the baseline in virtually all recent work on word alignment. In Table 2, we report the best symmetrized results.

The low F-measure scores of the baselines motivate our work.

3 Improving Word Alignments

3.1 Discriminative Reranking of the IBM Models

We reinterpret the five groups of parameters of Model 4 listed in the first five lines of Table 3 as sub-models of a log-linear model (see Equation 1). Each sub-model h_m has an associated weight λ_m. Given a vector of these weights λ, the alignment search problem, i.e. the search to return the best alignment $\hat{a}$ of the sentences e and f according to the model, is specified by Equation 2.

$$p_\lambda(f, a|e) = \frac{\exp(\sum_i \lambda_i h_i(a, e, f))}{\sum_{a', f'} \exp(\sum_i \lambda_i h_i(a', e, f'))} \quad (1)$$

$$\hat{a} = \operatorname*{argmax}_a \sum_i \lambda_i h_i(f, a, e) \quad (2)$$

m	Model 4	Description	m	Description
1	$t(f\|e)$	translation probs, f and e are words	9	translation table using approx. stems
2	$n(\phi\|e)$	fertility probs, ϕ is number of words generated by e	10	backoff fertility (fertility estimated over all e)
3	$null$	parameters used in generating Foreign words which are unaligned	11	backoff fertility for words with count $<= 5$
4	$d_1(\triangle j)$	movement probs of leftmost Foreign word translated from a particular e	12	translation table from HMM iteration 4
5	$d_{>1}(\triangle j)$	movement probs of other Foreign words translated from a particular e	13	zero fertility English word penalty
6		translation table from refined combination of both alignments	14	non-zero fertility English word penalty
7		translation table from union of both alignments	15	NULL Foreign word penalty
8		translation table from intersection of both alignments	16	non-NULL Foreign word penalty

Table 3: Sub-Models. Note that sub-models 1 to 5 are IBM Model 4, sub-models 6 to 16 are new.

Log-linear models are often trained to maximize entropy, but we will train our model directly on the final performance criterion. We use $1-$F-measure as our error function, comparing hypothesized word alignments for the discriminative training set with the gold standard.

Och (2003) has described an efficient exact one-dimensional error minimization technique for a similar search problem in machine translation. The technique involves calculating a piecewise constant function $f_m(x)$ which evaluates the error of the hypotheses which would be picked by equation 2 from a set of hypotheses if we hold all weights constant, except for the weight λ_m (which is set to x).

The discriminative reranking algorithm is initialized with the parameters of the sub-models θ, an initial choice of the λ vector, gold standard word alignments (labels) for the alignment discriminative training set, the constant N specifying the N-best list size used[1], and an empty master set of hypothesized alignments. The algorithm is a three step loop:

1. Enrich the master set of hypothesized alignments by producing an N-best list using λ. If all of the hypotheses in the N-best list are already in the master set, the algorithm has converged, so terminate the loop.
2. Consider the current λ vector and 999 additional randomly generated vectors, setting λ to the vector with lowest error on the master set.
3. Repeatedly run Och's one-dimensional error minimization step until there is no further error reduction (this results in a new vector λ).

[1] $N = 128$ for our experiments

3.2 Improvements to the Model and Search

3.2.1 New Sources of Knowledge

We define new sub-models to model factors not captured by Model 4. These are lines 6 to 16 of Table 3, where we use the "E to F" alignment direction as an example. We use word-level translation tables informed by both the "E to F" and the "F to E" translation directions derived using the three symmetrization heuristics, the "E to F" translation table from the final iteration of the HMM model and an "E to F" translation table derived using approximative stemming. The approximative stemming sub-model (sub-model 9) uses the first 4 letters of each vocabulary item as the stem for English and French while for Arabic we use the full word as the stem. We also use sub-models for backed off fertility, and direct penalization of unaligned English words ("zero fertility") and aligned English words, and unaligned Foreign words ("NULL-generated" words) and aligned Foreign words. This is a small sampling of the kinds of knowledge sources we can use in this framework; many others have been proposed in the literature.

Table 4 shows an evaluation of discriminative reranking. We observe:

1. The first line is the starting point, which is the Viterbi alignment of the 4th iteration of HMM training.
2. The 1-to-many alignments generated by discriminatively reranking Model 4 are better than the 1-to-many alignments of four iterations of Model 4.
3. The 1-to-many alignments of the discriminatively reranked extended model are much better than four iterations of Model 4.

System	F-measure F to E	F-measure E to F	F-measure Best Symm.
A/E last iteration HMM	58.6 / 54.4	47.7 / 39.9	62.1 / 57.0 (union)
A/E Model 4 reranking	65.3 / 59.5	55.7 / 51.4	69.7 / 64.6 (union)
A/E extended model reranking	68.4 / 62.2	61.6 / 57.7	72.0 / 66.4 (union)
A/E Model 4: Iteration 4	65.6 / 60.5	53.6 / 50.2	69.1 / 64.6 (union)
F/E last iteration HMM	72.4 / 73.9	71.5 / 71.8	76.4 / 77.3 (refined)
F/E Model 4 reranking	77.9 / 77.9	78.4 / 77.7	79.2 / 79.4 (refined)
F/E extended model reranking	78.7 / 80.2	79.3 / 79.6	79.6 / 80.4 (refined)
F/E Model 4: Iteration 4	73.8 / 75.1	74.2 / 73.5	76.5 / 76.4 (refined)

Table 4: Discriminative Reranking with Improved Search. F-measures are presented on both the alignment discriminative training set and the alignment test set sub-corpora, separated by /.

4. The discriminatively reranked extended model outperforms four iterations of Model 4 in both cases with the best heuristic symmetrization, but some of the gain is lost as we are optimizing the F-measure of the 1-to-many alignments rather than the F-measure of the many-to-many alignments directly.

Overall, the results show our approach is better than or competitive with running four iterations of unsupervised Model 4 training.

3.2.2 New Alignment Search Algorithm

Brown et al. (1993) introduced operations defining a hillclimbing search appropriate for Model 4. Their search starts with a complete hypothesis and exhaustively applies two operations to it, selecting the best improved hypothesis it can find (or terminating if no improved hypothesis is found). This search makes many search errors[2]. We developed a new alignment algorithm to reduce search errors:

- We perform an initial hillclimbing search (as in the baseline algorithm) but construct a priority queue of possible other candidate alignments to consider.
- Alignments which are expanded are marked so that they will not be returned to at a future point in the search.
- The alignment search operates by considering complete hypotheses so it is an "anytime" algorithm (meaning that it always has a current best guess). Timers can therefore be used to terminate the processing of the priority queue of candidate alignments.

The first two improvements are related to the well-known Tabu local search algorithm (Glover, 1986). The third improvement is important for restricting total time used when producing alignments for large training corpora.

We performed two experiments. The first evaluates the number of search errors. For each corpus we sampled 1000 sentence pairs randomly, with no sentence length restriction. Model 4 parameters are estimated from the final HMM Viterbi alignment of these sentence pairs. We then search to try to find the Model 4 Viterbi alignment with both the new and old algorithms, allowing them both to process for the same amount of time. The percentage of known search errors is the percentage of sentences from our sample in which we were able to find a more probable candidate by applying our new algorithm using 24 hours of computation for just the 1000 sample sentences. Table 5 presents the results, showing that our new algorithm reduced search errors in all cases, but further reduction could be obtained. The second experiment shows the impact of the new search on discriminative reranking of Model 4 (see Table 6). Reduced search errors lead to a better fit of the discriminative training corpus.

4 Semi-Supervised Training for Word Alignments

Intuitively, in approximate EM training for Model 4 (Brown et al., 1993), the E-step corresponds to calculating the probability of all alignments according to the current model estimate, while the M-step is the creation of a new model estimate given a probability distribution over alignments (calculated in the E-step).

In the E-step ideally all possible alignments should be enumerated and labeled with $p(a|e, f)$, but this is intractable. For the M-step, we would like to count over all possible alignments for each sentence pair, weighted by their probability according to the model estimated at the previous

[2] A search error in a word aligner is a failure to find the best alignment according to the model, i.e. in our case a failure to maximize Equation 2.

SYSTEM	F TO E ERRORS %	E TO F ERRORS %
A/E OLD	19.4	22.3
A/E NEW	8.5	15.3
F/E OLD	32.5	25.9
F/E NEW	13.7	10.4

Table 5: Comparison of New Search Algorithm with Old Search Algorithm

SYSTEM	F-MEASURE F TO E	F-MEASURE E TO F	F-MEASURE BEST SYMM.
A/E MODEL 4 RERANKING OLD	64.1 / 58.1	54.0 / 48.8	67.9 / 63.0 (UNION)
A/E MODEL 4 RERANKING NEW	65.3 / 59.5	55.7 / 51.4	69.7 / 64.6 (UNION)
F/E MODEL 4 RERANKING OLD	77.3 / 77.8	78.3 / 77.2	79.2 / 79.1 (REFINED)
F/E MODEL 4 RERANKING NEW	77.9 / 77.9	78.4 / 77.7	79.2 / 79.4 (REFINED)

Table 6: Impact of Improved Search on Discriminative Reranking of Model 4

step. Because this is not tractable, we make the assumption that the single assumed Viterbi alignment can be used to update our estimate in the M-step. This approximation is called Viterbi training. Neal and Hinton (1998) analyze approximate EM training and motivate this type of variant.

We extend approximate EM training to perform a new type of training which we call Minimum Error / Maximum Likelihood Training. The intuition behind this approach to semi-supervised training is that we wish to obtain the advantages of both discriminative training (error minimization) and approximate EM (which allows us to estimate a large numbers of parameters even though we have very few gold standard word alignments). We introduce the EMD algorithm, in which discriminative training is used to control the contributions of sub-models (thereby minimizing error), while a procedure similar to one step of approximate EM is used to estimate the large number of sub-model parameters.

A brief sketch of the EMD algorithm applied to our extended model is presented in Figure 1. Parameters have a superscript t representing their value at iteration t. We initialize the algorithm with the gold standard word alignments (labels) of the word alignment discriminative training set, an initial λ, N, and the starting alignments (the iteration 4 HMM Viterbi alignment). In line 2, we make iteration 0 estimates of the 5 sub-models of Model 4 and the 6 heuristic sub-models which are iteration dependent. In line 3, we run discriminative training using the algorithm from Section 3.1. In line 4, we measure the error of the resulting λ vector. In the main loop in line 7 we align the full training set (similar to the E-step of EM), while in line 8 we estimate the iteration-dependent sub-models (similar to the M-step of EM). Then

```
1: Algorithm EMD(labels, λ', N, starting alignments)
2: estimate θ^0_m for m = 1 to 11
3: λ^0 = Discrim(θ^0, λ', labels, N)
4: e^0 = E(λ^0, labels)
5: t = 1
6: loop
7:    align full training set using λ^{t-1} and θ^{t-1}_m
8:    estimate θ^t_m for m = 1 to 11
9:    λ^t = Discrim(θ^t, λ'', labels, N)
10:   e^t = E(λ^t, labels)
11:   if e^t >= e^{t-1} then
12:      terminate loop
13:   end if
14:   t = t + 1
15: end loop
16: return hypothesized alignments of full training set
```

Figure 1: Sketch of the EMD algorithm

we perform discriminative reranking in line 9 and check for convergence in lines 10 and 11 (convergence means that error was not decreased from the previous iteration). The output of the algorithm is new hypothesized alignments of the training corpus.

Table 7 evaluates the EMD semi-supervised training algorithm. We observe:

1. In both cases there is improved F-measure on the second iteration of semi-supervised training, indicating that the EMD algorithm performs better than one step discriminative reranking.
2. The French/English data set has converged[3] after the second iteration.
3. The Arabic/English data set converged after improvement for the first, second and third iterations.

We also performed an additional experiment for French/English aimed at understanding the potential contribution of the word aligned data without

[3]Convergence is achieved because error on the word alignment discriminative training set does not improve.

SYSTEM	F-MEASURE F TO E	F-MEASURE E TO F	BEST SYMM.
A/E STARTING POINT	58.6 / 54.4	47.7 / 39.9	62.1 / 57.0 (UNION)
A/E EMD: ITERATION 1	68.4 / 62.2	61.6 / 57.7	72.0 / 66.4 (UNION)
A/E EMD: ITERATION 2	69.8 / 63.1	64.1 / 59.5	74.1 / 68.1 (UNION)
A/E EMD: ITERATION 3	70.6 / 65.4	64.3 / 59.2	74.7 / 69.4 (UNION)
F/E STARTING POINT	72.4 / 73.9	71.5 / 71.8	76.4 / 77.3 (REFINED)
F/E EMD: ITERATION 1	78.7 / 80.2	79.3 / 79.6	79.6 / 80.4 (REFINED)
F/E EMD: ITERATION 2	79.4 / 80.5	79.8 / 80.5	79.9 / 81.2 (REFINED)

Table 7: Semi-Supervised Training Task F-measure

the new algorithm[4]. Like Ittycheriah and Roukos (2005), we converted the alignment discriminative training corpus links into a special corpus consisting of parallel sentences where each sentence consists only of a single word involved in the link. We found that the information in the links was "washed out" by the rest of the data and resulted in no change in the alignment test set's F-Measure. Callison-Burch et al. (2004) showed in their work on combining alignments of lower and higher quality that the alignments of higher quality should be given a much higher weight than the lower quality alignments. Using this insight, we found that adding 10,000 copies of the special corpus to our training data resulted in the highest alignment test set gain, which was a small gain of 0.6 F-Measure. This result suggests that while the link information is useful for improving F-Measure, our improved methods for training are producing much larger improvements.

5 Improvement of MT Quality

The symmetrized alignments from the last iteration of EMD were used to build phrasal SMT systems, as were the symmetrized Model 4 alignments (the baseline). Aside from the final alignment, all other resources were held constant between the baseline and contrastive SMT systems, including those based on lower level alignments models such as IBM Model 1. For all of our experiments, we use two language models, one built using the English portion of the training data and the other built using additional English news data. We run Maximum BLEU (Och, 2003) for 25 iterations individually for each system.

Table 8 shows our results. We report BLEU (Papineni et al., 2001) multiplied by 100. We also show the F-measure after heuristic symmetrization of the alignment test sets. The table shows that our algorithm produces heuristically symmetrized final alignments of improved F-measure. Using these alignments in our phrasal SMT system, we produced a statistically significant BLEU improvement (at a 95% confidence interval a gain of 0.78 is necessary) on the French/English task and a statistically significant BLEU improvement on the Arabic/English task (at a 95% confidence interval a gain of 1.2 is necessary).

5.1 Error Criterion

The error criterion we used for all experiments is $1 -$ F-measure. The formula for F-measure is shown in Equation 3. (Fraser and Marcu, 2006) established that tuning the trade-off between Precision and Recall in the F-Measure formula will lead to the best BLEU results. We tuned α by building a collection of alignments using our baseline system, measuring Precision and Recall against the alignment discriminative training set, building SMT systems and measuring resulting BLEU scores, and then searching for an appropriate α setting. We searched $\alpha = 0.1, 0.2, ..., 0.9$ and set α so that the resulting F-measure tracks BLEU to the best extent possible. The best settings were $\alpha = 0.2$ for Arabic/English and $\alpha = 0.7$ for French/English, and these settings of α were used for every result reported in this paper. See (Fraser and Marcu, 2006) for further details.

$$F(A, S, \alpha) = \frac{1}{\frac{\alpha}{\text{Precision}(A,S)} + \frac{(1-\alpha)}{\text{Recall}(A,S)}} \quad (3)$$

6 Previous Research

Previous work on discriminative training for word-alignment differed most strongly from our approach in that it generally views word-alignment as a supervised task. Examples of this perspective include (Liu et al., 2005; Ittycheriah and Roukos, 2005; Moore, 2005; Taskar et al., 2005). All of these also used knowledge from one of the IBM Models in order to obtain competitive results

[4]We would like to thank an anonymous reviewer for suggesting that this experiment would be useful even when using a small discriminative training corpus.

SYSTEM	BLEU	F-MEASURE
A/E UNSUP. MODEL 4 UNION	49.16	64.6
A/E EMD 3 UNION	50.84	69.4
F/E UNSUP. MODEL 4 REFINED	30.63	76.4
F/E EMD 2 REFINED	31.56	81.2

Table 8: Evaluation of Translation Quality

with the baseline (with the exception of (Moore, 2005)). We interleave discriminative training with EM and are therefore performing semi-supervised training. We show that semi-supervised training leads to better word alignments than running unsupervised training followed by discriminative training.

Another important difference with previous work is that we are concerned with generating many-to-many word alignments. Cherry and Lin (2003) and Taskar et al. (2005) compared their results with Model 4 using "intersection" by looking at AER (with the "Sure" versus "Possible" link distinction), and restricted themselves to considering 1-to-1 alignments. However, "union" and "refined" alignments, which are many-to-many, are what are used to build competitive phrasal SMT systems, because "intersection" performs poorly, despite having been shown to have the best AER scores for the French/English corpus we are using (Och and Ney, 2003). (Fraser and Marcu, 2006) recently found serious problems with AER both empirically and analytically, which explains why optimizing AER frequently results in poor machine translation performance.

Finally, we show better MT results by using F-measure with a tuned α value. The only previous discriminative approach which has been shown to produce translations of similar or better quality to those produced by the symmetrized baseline was (Ittycheriah and Roukos, 2005). They had access to 5000 gold standard word alignments, considerably more than the 100 or 110 gold standard word alignments used here. They also invested significant effort in sub-model engineering (producing both sub-models specific to Arabic/English alignment and sub-models which would be useful for other language pairs), while we use sub-models which are simple extensions of Model 4 and language independent.

The problem of semi-supervised learning is often defined as "using unlabeled data to help supervised learning" (Seeger, 2000). Most work on semi-supervised learning uses underlying distributions with a relatively small number of parameters. An initial model is estimated in a supervised fashion using the labeled data, and this supervised model is used to attach labels (or a probability distribution over labels) to the unlabeled data, then a new supervised model is estimated, and this is iterated. If these techniques are applied when there are a small number of labels in relation to the number of parameters used, they will suffer from the "overconfident pseudo-labeling problem" (Seeger, 2000), where the initial labels of poor quality assigned to the unlabeled data will dominate the model estimated in the M-step. However, there are tasks with large numbers of parameters where there are sufficient labels. Nigam et al. (2000) addressed a text classification task. They estimate a Naive Bayes classifier over the labeled data and use it to provide initial MAP estimates for unlabeled documents, followed by EM to further refine the model. Callison-Burch et al. (2004) examined the issue of semi-supervised training for word alignment, but under a scenario where they simulated sufficient gold standard word alignments to follow an approach similar to Nigam et al. (2000). We do not have enough labels for this approach.

We are aware of two approaches to semi-supervised learning which are more similar in spirit to ours. Ivanov et al. (2001) used discriminative training in a reinforcement learning context in a similar way to our adding of a discriminative training step to an unsupervised context. A large body of work uses semi-supervised learning for clustering by imposing constraints on clusters. For instance, in (Basu et al., 2004), the clustering system was supplied with pairs of instances labeled as belonging to the same or different clusters.

7 Conclusion

We presented a semi-supervised algorithm based on IBM Model 4, with modeling and search extensions, which produces alignments of improved F-measure over unsupervised Model 4 training.

We used these alignments to produce translations of higher quality.

The semi-supervised learning literature generally addresses augmenting supervised learning tasks with unlabeled data (Seeger, 2000). In contrast, we augmented an unsupervised learning task with labeled data. We hope that Minimum Error / Maximum Likelihood training using the EMD algorithm can be used for a wide diversity of tasks where there is not enough labeled data to allow supervised estimation of an initial model of reasonable quality.

8 Acknowledgments

This work was partially supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. We would like to thank the USC Center for High Performance Computing and Communications.

References

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. 2004. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proc. of the ACM SIGKDD international conference on knowledge discovery and data mining*, pages 59–68, New York. ACM Press.

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

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July.

Alexander Fraser and Daniel Marcu. 2006. Measuring word alignment quality for statistical machine translation. In *Technical Report ISI-TR-616. Available at http://www.isi.edu/ fraser/research.html*, ISI/University of Southern California, May.

Fred Glover. 1986. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13(5):533–549.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proc. of Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing*, Vancouver, BC.

Yuri A. Ivanov, Bruce Blumberg, and Alex Pentland. 2001. Expectation maximization for weakly labeled data. In *ICML '01: Proc. of the Eighteenth International Conf. on Machine Learning*, pages 218–225.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 459–466, Ann Arbor, MI.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-07, Institute for Research in Cognitive Science, Philadelphia, PA.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proc. of Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing*, Vancouver, BC, October.

Radford M. Neal and Geoffrey E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer.

Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

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 *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September.

Matthias Seeger. 2000. Learning with labeled and unlabeled data. In *Technical report, 2000. Available at http://www.dai.ed.ac.uk/ seeger/papers.html*.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proc. of Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing*, Vancouver, BC, October.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.