Squibs and Discussions

Measuring Word Alignment Quality for Statistical Machine Translation

Alexander Fraser*
University of Southern California

Daniel Marcu*
University of Southern California

Automatic word alignment plays a critical role in statistical machine translation. Unfortunately, the relationship between alignment quality and statistical machine translation performance has not been well understood. In the recent literature, the alignment task has frequently been decoupled from the translation task and assumptions have been made about measuring alignment quality for machine translation which, it turns out, are not justified. In particular, none of the tens of papers published over the last five years has shown that significant decreases in alignment error rate (AER) result in significant increases in translation performance. This paper explains this state of affairs and presents steps towards measuring alignment quality in a way which is predictive of statistical machine translation performance.

1. Introduction

Automatic word alignment (Brown et al. 1993) is a vital component of all statistical machine translation (SMT) approaches. There were a number of research papers presented from 2000 to 2005 at ACL, NAACL, HLT, COLING, WPT03, WPT05, and so forth, outlining techniques for attempting to increase word alignment quality. Despite this high level of interest, none of these techniques has been shown to result in a large gain in translation performance as measured by BLEU (Papineni et al. 2001) or any other metric. We find this lack of correlation between previous word alignment quality metrics and BLEU counterintuitive, because we and other researchers have measured this correlation in the context of building SMT systems that have benefited from using the BLEU metric in improving performance in open evaluations such as the NIST evaluations.[1]

We confirm experimentally that previous metrics do not predict BLEU well and develop a methodology for measuring alignment quality that is predictive of BLEU. We

* USC/ISI - Natural Language Group, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292-6601. E-mail: fraser@isi.edu, marcu@isi.edu.

1 Because in our experiments we use BLEU to compare the performance of systems built using a common framework where the only difference is the word alignment, we make no claims about the utility of BLEU for measuring translation quality in absolute terms, nor its utility for comparing two completely different MT systems.

also show that alignment error rate (AER) is not correctly derived from F-Measure and is therefore unlikely to be useful as a metric.

2. Experimental Methodology

2.1 Data

To build an SMT system we require a bitext and a word alignment of that bitext, as well as language models built from target language data. In all of our experiments, we will hold the bitext and target language resources constant, and only vary how we construct the word alignment.

The gold standard word alignment sets we use have been manually annotated using links between words showing translational correspondence. Links which must be present in a hypothesized alignment are called Sure links. Some of the alignment sets also have links which are not Sure links but are Possible links (Och and Ney 2003). Possible links which are not Sure[2] may be present but need not be present.

We evaluate the translation performance of SMT systems by translating a held-out translation test set and measuring the BLEU score of our hypothesized translations against one or more reference translations. We also have an additional held-out translation set, the development set, which is employed by the MT system to train the weights of its log-linear model to maximize BLEU (Och 2003). We work with data sets for three different language pairs, examining French to English, Arabic to English, and Romanian to English translation tasks.

The training data for the French/English data set is taken from the LDC Canadian Hansard data set, from which the word aligned data (presented in Och and Ney 2003) was also taken. The English side of the bitext is 67.4 million words. We used a separate Canadian Hansard data set (released by ISI) as the source of the translation test set and development set. We evaluate two different tasks using this data, a medium task where 1/8 of the data (8.4 million English words) is used as the fixed bitext, and a large task where all of the data is used as the fixed bitext. The 484 sentences in the gold standard word alignments have 4,376 Sure links and 19,222 Possible links.

The Arabic/English training corpus is the data used for the NIST 2004 machine translation evaluation.[3] The English side of the bitext is 99.3 million words. The translation development set is the "NIST 2002 Dry Run," and the test set is the "NIST 2003 evaluation set." We have annotated gold standard alignments for 100 parallel sentences using Sure links, following the Blinker guidelines (Melamed 1998), which call for Sure links only (there were 2,154 Sure links). Here we also examine a medium task using 1/8 of the data (12.4 million English words) and a large task using all of the data.

The Romanian/English training data was used for the tasks on Romanian/English alignment at WPT03 (Mihalcea and Pederson 2003) and WPT05 (Martin, Mihalcea, and Pedersen 2005). We carefully removed two sections of news bitext to use as the translation development and test sets. The English side of the training corpus is 964,000 words. The alignment set is the first 148 annotated sentences used for the 2003 task (there were 3,181 Sure links).

2 Sure links are by definition also Possible.

3 http://www.nist.gov/speech/tests/summaries/2004/mt04.htm.

2.2 Measuring Translation Performance Changes Caused By Alignment

In phrased-based SMT (Koehn, Och, and Marcu 2003) the knowledge sources which vary with the word alignment are the phrase translation lexicon (which maps source phrases to target phrases using counts from the word alignment) and some of the word level translation parameters (sometimes called lexical smoothing). However, many knowledge sources do not vary with the final word alignment, such as rescoring with IBM Model 1, *n*-gram language models, and the length penalty. In our experiments, we use a state-of-the-art phrase-based system, similar to Koehn, Och, and Marcu. The weights of the different knowledge sources in the log-linear model used by our system are trained using Maximum BLEU (Och 2003), which we run for 25 iterations individually for each system. Two language models are used, one built using the target language training data and the other built using additional news data.

2.3 Generating Alignments of Varying Quality

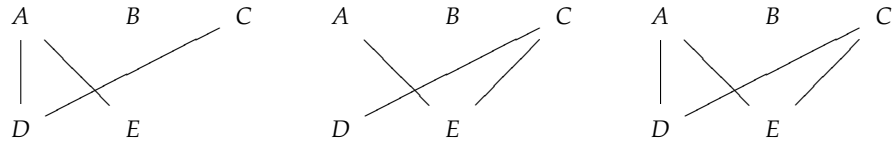
We have observed in the past that generative models used for statistical word alignment create alignments of increasing quality as they are exposed to more data. The intuition behind this is simple; as more co-occurrences of source and target words are observed, the word alignments are better. If we wish to increase the quality of a word alignment, we allow the alignment process access to extra data, which is used only during the alignment process and then removed. If we wish to decrease the quality of a word alignment, we divide the bitext into pieces and align the pieces independently of one another, finally concatenating the results together.

To generate word alignments we use GIZA++ (Och and Ney 2003), which implements both the IBM Models of Brown et al. (1993) and the HMM model (Vogel, Ney, and Tillmann 1996). We use Model 1, HMM, and Model 4, in that order. The output of these models is an alignment of the bitext which projects one language to another. GIZA++ is run end-to-end twice. In one case we project the source language to the target language, and in the other we project the target language to the source language. The output of GIZA++ is then post-processed using the three "symmetrization heuristics" described in Och and Ney (2003). We evaluate our approaches using these heuristics because we would like to account for alignments generated in different fashions. These three heuristics were used as the baselines in virtually all recent work on automatic word alignment, and most of the best SMT systems use these techniques as well.

When applying the union symmetrization heuristic, we take the transitive closure of the bipartite graph created, which results in fully connected components indicating translational correspondence.[4] Each of the presented alignments are equivalent from a translational correspondence perspective and the first two will be mapped to the third

4 We have no need to do this for the "refined" and "intersection" heuristics, because they only produce alignments in which the components are fully connected.

in order to ensure consistency between the number of links an alignment has and the translational equivalences licensed by that alignment.



3. Word Alignment Quality Metrics

3.1 Alignment Error Rate is Not a Useful Measure

We begin our study of metrics for word alignment quality by testing AER (Och and Ney 2003). AER requires a gold standard manually annotated set of Sure links and Possible links (referred to as S and P). Given a hypothesized alignment consisting of the link set A, three measures are defined:

$$\text{Precision}(A, P) = \frac{|P \cap A|}{|A|} \tag{1}$$

$$\text{Recall}(A, S) = \frac{|S \cap A|}{|S|} \tag{2}$$

$$\text{AER}(A, P, S) = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|} \tag{3}$$

In our graphs, we will present $1 - \text{AER}$ so that we have an accuracy measure.

We created alignments of varying quality for the medium French/English training set. We broke the data into separate pieces corresponding to 1/16, 1/8, 1/4, and 1/2 of the original data to generate degraded alignments, and we used 2, 4, and 8 times the original data to generate enhanced alignments. For the "fractional" alignments we report the average AER of the pieces.[5]

The graph in Figure 1 shows the correlation of $1 - \text{AER}$ with BLEU. High correlation would look like a line from the bottom left corner to the top right corner. As can be seen by looking at the graph, there is low correlation between $1 - \text{AER}$ and the BLEU score. A concise mathematical description of correlation is the coefficient of determination (r^2),

5 For example, for 1/16, we perform 16 pairs of alignments, each of which includes the full gold standard text, and another 16 pairs of alignments without the gold standard text. We then apply the symmetrization heuristics to these pairs. We use the symmetrized alignments including the text from the gold standard set to measure AER and we concatenate those not including the gold standard text to build SMT systems for measuring BLEU.

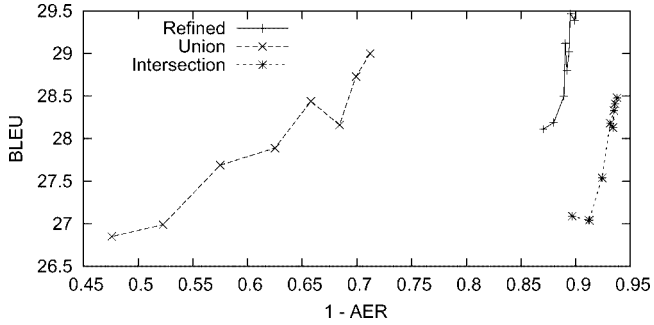


Figure 1
French 1 − AER versus BLEU, $r^2 = 0.16$.

which is the square of the Pearson product-moment correlation coefficient (r). Here, $r^2 = 0.16$, which is low.

The correlation is low because of a significant shortcoming in the mathematical formulation of AER, which to our knowledge has not been previously reported. Och and Ney (2003) state that AER is derived from F-Measure. But AER does not share a very important property of F-Measure, which is that unbalanced precision and recall are penalized, where $S \subset P$ (i.e., when we make the Sure versus Possible distinction).[6] We will show this using an example.

We first define the measure "F-Measure with Sure and Possible" using Och and Ney's Precision and Recall formulae together with the standard F-Measure formula (van Rijsbergen 1979). In the F-Measure formula (4) there is a parameter α which sets the trade-off between Precision and Recall. When an equal trade-off is desired, α is set to 0.5.

$$\text{F-Measure with Sure and Possible}(A, P, S, \alpha) = \frac{1}{\frac{\alpha}{\text{Precision}_{(A,P)}} + \frac{(1-\alpha)}{\text{Recall}_{(A,S)}}} \tag{4}$$

We compare two hypothesized alignments where $|A|$, the number of hypothesized alignment links, is the same, for instance, $|A| = 100$. Let $|S| = 100$. In the first case, let $|P \cap A| = 50$ and $|S \cap A| = 50$. Precision is 0.50 and Recall is 0.50. In the second case, let $|P \cap A| = 75$ and $|S \cap A| = 25$. Precision is 0.75 and Recall is 0.25.

The basic property of F-Measure, if we set α equal to 0.5, is that unbalanced precision and recall should be penalized. The first hypothesized alignment has an F-Measure with Sure and Possible score of 0.50, whereas the second has a worse score, 0.375.

However, if we substitute the relevant values into the formula for AER (Equation (3)), we see that 1 − AER for both of the hypothesized alignments is 0.5. Therefore AER does not share the property of F-Measure (with $\alpha = 0.5$) that unbalanced precision and recall are penalized. Because of this, it is possible to maximize AER by favoring precision over recall, which can be done by simply guessing very few alignment links. Unfortunately, when $S \subset P$, this leads to strong biases, which makes AER not useful as a metric.

6 Note that if $S = P$ then AER reduces to balanced F-Measure.

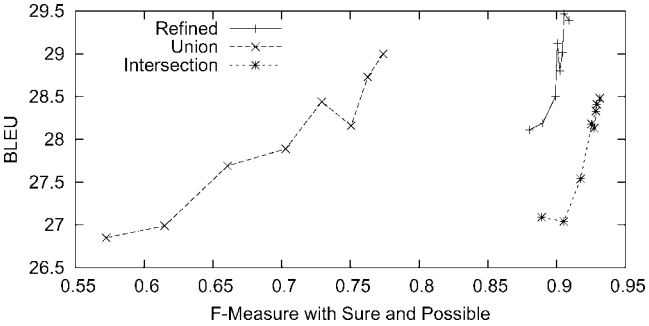


Figure 2
French F-Measure with Sure and Possible $\alpha = 0.5$ versus BLEU, $r^2 = 0.20$.

Goutte, Yamada, and Gaussier (2004) previously observed that AER could be unfairly optimized by using a bias toward precision which was unlikely to improve the usefulness of the alignments. Possible problems with AER were discussed at WPT 2003 and WPT 2005.

Examining the graph in Figure 2, we see that F-Measure with Sure and Possible has some predictive power for the data points generated using a single heuristic, but the overall correlation is still low, $r^2 = 0.20$. We need a measure that predicts BLEU without having a dependency on the way the alignments are generated.

3.2 Balanced F-Measure is Better, but Still Inadequate

We wondered whether the low correlation was caused by the Sure and Possible distinction. We reannotated the first 110 sentences of the French test set using the Blinker guidelines (there were 2,292 Sure links). We define F-Measure without the Sure versus Possible distinction (i.e., all links are Sure) in Equation (5), and set $\alpha = 0.5$. This measure has been extensively used with other word alignment test sets. Figure 3 shows the results. Correlation is higher: $r^2 = 0.67$.

$$\text{F-Measure}(A, S, \alpha) = \frac{1}{\frac{\alpha}{\text{Precision}(A,S)} + \frac{(1-\alpha)}{\text{Recall}(A,S)}} \tag{5}$$

3.3 Varying the Trade-Off Between Precision and Recall Works Well

We then hypothesized that the trade-off between precision and recall is important. This is controlled in both formulae by the constant α. We search $\alpha = 0.1, 0.2, \ldots, 0.9$. The best results are: $\alpha = 0.1$ for the original annotation annotated with Sure and Possible (see Figure 4), and $\alpha = 0.4$ for the first 110 sentences as annotated by us (see Figure 5).[7] The relevant r^2 scores were 0.80 and 0.85, respectively. With a good α setting, we are able

7 We also checked the first 110 sentences using the original annotation to ensure that the differences observed were not an effect of restricting our annotation to these sentences.

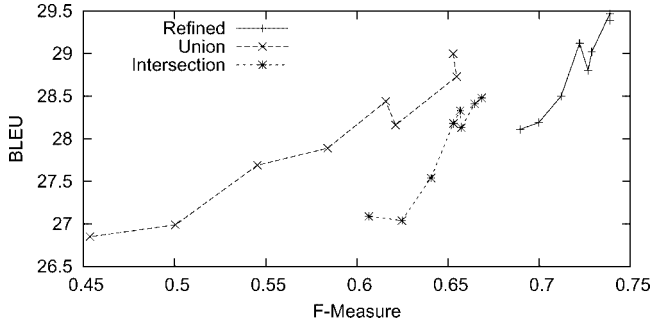


Figure 3
French F-Measure $\alpha = 0.5$ versus BLEU, $r^2 = 0.67$.

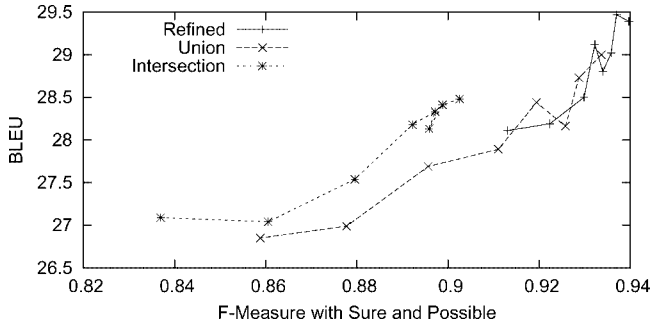


Figure 4
French F-Measure with Sure and Possible $\alpha = 0.1$ versus BLEU, $r^2 = 0.80$.

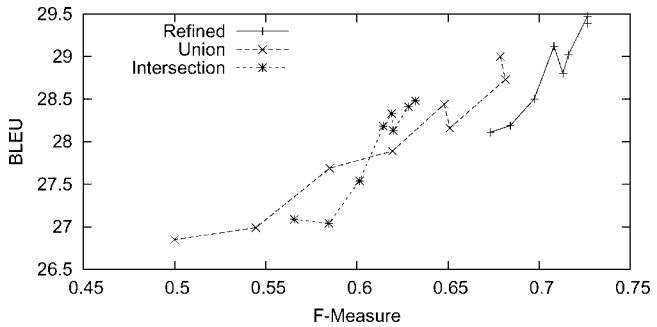


Figure 5
French F-Measure $\alpha = 0.4$ versus BLEU, $r^2 = 0.85$.

to predict the machine translation results reasonably well. For the original annotation, recall is very highly weighted, whereas for our annotation, recall is still more important than precision.[8] Our results also suggest that better correlation will be achieved when using Sure-only annotation than with Sure and Possible annotation.

8 α less than 0.5 weights recall higher, whereas α greater than 0.5 weights precision higher; see the F-Measure formulae.

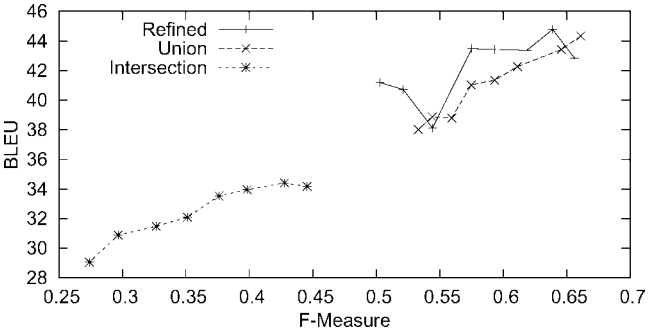


Figure 6
Arabic F-Measure $\alpha = 0.1$ versus BLEU, $r^2 = 0.93$.

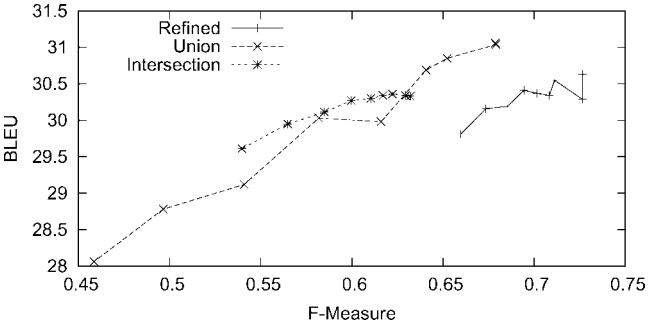


Figure 7
Large French F-Measure $\alpha = 0.4$ (110 sentences) versus BLEU, $r^2 = 0.64$.

We then tried the medium Arabic training set. Results are shown in Figure 6, the best setting of $\alpha = 0.1$, and $r^2 = 0.93$. F-Measure is effective in predicting machine translation performance for this set.

We also tried the larger tasks, where we can only decrease alignment quality, as we have no additional data. For the large French/English corpus the best results are with $\alpha = 0.2$ for the original annotation of 484 sentences and $\alpha = 0.4$ for the new

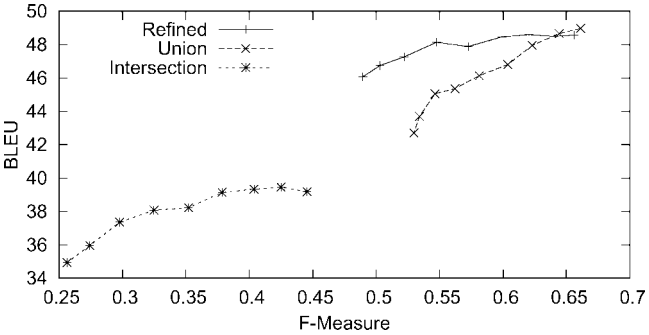


Figure 8
Large Arabic F-Measure $\alpha = 0.1$ (100 sentences) versus BLEU, $r^2 = 0.90$.

annotation of 110 sentences with only Sure links (see Figure 7). Relevant r^2 scores were 0.62 and 0.64, respectively. Disappointingly, our measures are not able to fully explain MT performance for the large French/English task.

For the large Arabic/English corpus, the results were better: the best correlation was at $\alpha = 0.1$, for which $r^2 = 0.90$ (see Figure 8). We can predict MT performance for this set. It is worth noting that the Arabic/English translation task and data set has been tested in conjunction with our translation system over a long period, but the French/English translation task and data has not. As a result, there may be hidden factors that affect the performance of our MT system, which only appear in conjunction with the large French/English task.

One well-studied task on a smaller data set is the Romanian/English shared word alignment task from the Workshop on Parallel Text at ACL 2005 (Martin, Mihalcea, and Pedersen 2005). We only decreased alignment quality and used 5 data points for each symmetrization heuristic due to the small bitext. The best setting of α was $\alpha = 0.2$, for which $r^2 = 0.94$, showing that F-Measure is again effective in predicting BLEU.

4. Conclusion

We have presented an empirical study of the use of simple evaluation metrics based on gold standard alignment of a small number of sentences to predict machine translation performance. Based on our experiments we can now draw the following conclusions:

1. When $S \subset P$, AER does not share the important property of F-Measure that unequal precision and recall are penalized, making it easy to obtain good AER scores by simply guessing fewer alignment links. As a result AER is a misleading metric that should no longer be used.
2. Good correlation was obtained for the medium French and Arabic data sets, the large Arabic data set, and the small Romanian data set. We have explained most of the effect of alignment quality on these sets, and if we are given the F-Measure of a hypothesized word alignment for the bitext, we can make a reasonable prediction as to what the resulting BLEU score will be.
3. We have only partially explained the effect of alignment quality on BLEU for the large French data set, and further investigation is warranted.
4. We recommend using the Blinker guidelines as a starting point for new alignment annotation efforts, and that Sure-only annotation be used. If a larger gold standard is available and was already annotated using the Sure versus Possible distinction, this is likely to have only slightly worse results.

Although we have addressed measuring alignment quality for phrasal SMT, similar work is now required to see how to measure alignment quality for other settings of machine translation and for other tasks. For an evaluation campaign the organizers should pick a specific task, such as improving phrasal SMT, and calculate an appropriate α to be used. Individual researchers working on the same phrasal SMT tasks as those reported here (or on very similar tasks) could use the values of α we calculated.

Our work invalidates some of the conclusions of recent alignment work which presented only evaluations based on metrics like AER or balanced F-Measure, and explains the lack of correlation in the few works which presented both such a metric

and final MT results. A good example of the former are our own results (Fraser and Marcu 2005). The work presented there had the highest balanced F-Measure scores for the Romanian/English WPT05 shared task, but based on the findings here it is possible that a different algorithm tuned for the correct criterion would have had better MT performance. Other work includes many papers working on alignment models where words are allowed to participate in a maximum of one link. These models generally have higher precision and lower recall than IBM Model 4 symmetrized using the "Refined" or "Union" heuristics. Recall that in Section 3.1 we showed that AER is broken in a way that favors precision. It is therefore likely that the results reported in these papers are affected by the AER bias and that the corresponding improvements in AER score do not correlate with increases in phrasal SMT performance.

We suggest comparing alignment algorithms by measuring performance in an identified final task such as machine translation. F-Measure with an appropriate setting of α will be useful during the development process of new alignment models, or as a maximization criterion for discriminative training of alignment models (Cherry and Lin 2003; Ayan, Dorr, and Monz 2005; Ittycheriah and Roukos 2005; Liu, Liu, and Lin 2005; Fraser and Marcu 2006; Lacoste-Julien et al. 2006; Moore, Yih, and Bode 2006).

Acknowledgments

This work was supported by DARPA-ITO grant NN66001-00-1-9814, NSF grant IIS-0326276, and the DARPA GALE Program. We thank USC High Performance Computing & Communications.

References

Ayan, Necip Fazil, Bonnie J. Dorr, and Christof Monz. 2005. Neuralign: Combining word alignments using neural networks. In *Proceedings of HLT-EMNLP*, pages 65–72, Vancouver, Canada.

Brown, Peter F., 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.

Cherry, Colin and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of ACL*, pages 88–95, Sapporo, Japan.

Fraser, Alexander and Daniel Marcu. 2005. Isi's participation in the Romanian-English alignment task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 91–94, Ann Arbor, MI.

Fraser, Alexander and Daniel Marcu. 2006. Semi-supervised training for word alignment. In *Proceedings of COLING-ACL*, pages 769–776, Sydney, Australia.

Goutte, Cyril, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *Proceedings of ACL*, pages 502–509, Barcelona, Spain.

Ittycheriah, Abraham and Salim Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of HLT-EMNLP*, pages 89–96, Vancouver, Canada.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 127–133, Edmonton, Canada.

Lacoste-Julien, Simon, Dan Klein, Ben Taskar, and Michael Jordan. 2006. Word alignment via quadratic assignment. In *Proceedings of HLT-NAACL*, pages 112–119, New York, NY.

Liu, Yang, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of ACL*, pages 459–466, Ann Arbor, MI.

Martin, Joel, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, MI.

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

Mihalcea, Rada and Ted Pederson. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts*, pages 1–10, Edmonton, Canada.

Moore, Robert C., Wen-Tau Yih, and Andreas Bode. 2006. Improved discriminative bilingual word alignment. In *Proceedings of COLING-ACL*, pages 513–520, Sydney, Australia.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan.

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

Papineni, Kishore A., 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.

van Rijsbergen, Keith. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA.

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