Word-Based Alignment, Phrase-Based Translation: What's the Link?

Adam Lopez
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742
alopez@cs.umd.edu

Philip Resnik
Department of Linguistics
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
resnik@umd.edu

Abstract

State-of-the-art statistical machine translation is based on alignments between *phrases* – sequences of words in the source and target sentences. The learning step in these systems often relies on alignments between *words*. It is often assumed that the quality of this word alignment is critical for translation. However, recent results suggest that the relationship between alignment quality and translation quality is weaker than previously thought. We investigate this question directly, comparing the impact of high-quality alignments with a carefully constructed set of degraded alignments. In order to tease apart various interactions, we report experiments investigating the impact of alignments on different aspects of the system. Our results confirm a weak correlation, but they also illustrate that more data and better feature engineering may be more beneficial than better alignment.

1 Introduction

Machine translation and alignment are closely related problems. In the translation problem, we are given a source sentence, and the task is to output a target sentence which conveys the same meaning. In the process, the system produces a mapping between words of the source sentence and target sentence. In the alignment problem, we are given both sentences, and the task is simply to find this mapping. Alignment can therefore be thought of as a more constrained version of the translation problem, in which we need find only one of the outputs instead of both.

As in other machine learning problems, statistical machine translation systems learn to make decisions by looking at previous examples of those decisions. Because the decisions made by translation systems produce alignments between source and target sentences, the training data must contain examples of these alignment decisions. However, the data available for training translation systems contains only pairs of sentences.

There are two solutions to this problem. The first is to treat the unseen alignments as a hidden input and apply unsupervised learning methods. The second is to first solve the easier problem of alignment, and then use supervised methods for learning. Current state-of-the-art models in machine translation are based on alignments between phrases – sequences of words within each sentence (Och et al., 1999; Koehn et al., 2003; Chiang, 2005; Simard et al., 2005). Unfortunately, unsupervised learning of phrase-based models is intractable. It requires numerous approximations and tradeoffs, and often produces poor results (Marcu and Wong, 2002; DeNero et al., 2006; Birch et al., 2006). Therefore, supervised methods are usually employed (Och et al., 1999; Koehn et al., 2003; Chiang, 2005). This requires a method for phrase alignment. However, there are only a few examples of research on the phrase alignment problem (Zhao and Waibel, 2005). An alternative is to first generate *word* alignments as if we were training a word-based system. Phrase alignments are then inferred heuristically from these word alignments. This approach was first described by Och et al. (1999) and later explored in some de-

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 90-99,
Cambridge, August 2006.

Citation	MT Test Corpus	AER			BLEU		
		worst	best	diff	worst	best	diff
Koehn et al. (2003) [1]	Europarl German	*	*	*	24.5	25.2	**0.7**
Callison-Burch et al. (2004)	Verbmobil German	12.17	7.52	**4.6**	27.0	28.2	**1.2**
Callison-Burch et al. (2004)	Hansards French	16.59	13.55	**3.0**	12.6	12.8	**0.2**
Ittycheriah and Roukos (2005) [2]	MT-Eval 2003 Arabic	23.7	12.2	**11.5**	45.9	47.9	**2.0**
Ittycheriah and Roukos (2005) [2]	MT-Eval 2004 Arabic	23.7	12.2	**11.5**	41.9	43.3	**1.4**
Ittycheriah and Roukos (2005) [2]	MT-Eval 2005 Arabic	23.7	12.2	**11.5**	45.6	46.5	**0.9**

Table 1: The impact of alignment performance on machine translation performance as reported in several recent studies. Alignment performance is measured using the alignment error rate (AER) (Och and Ney, 2000).[3] Translation performance is measured using BLEU (Papineni et al., 2002).[4]

tail by Koehn et al. (2003). It has since been widely adopted.

Because the heuristic approach depends on a word alignment, it is often assumed that the quality of the word alignment is critical to its success. A number of recent word alignment methods achieve impressive results on extrinsic metrics (Ayan et al., 2005; Ittycheriah and Roukos, 2005; Moore, 2005; Taskar et al., 2005). Often, it is implied that these improvements will propagate to a downstream translation system. However, several recent papers have reported that large gains in alignment accuracy often lead to, at best, minor gains in translation performance. Some examples are listed in Table 1. These results raise serious questions about the presumed utility of word alignment as an input to phrase-based statistical machine translation.

[1]Although the specific alignment error rate of the different methods used in this paper is unknown, we show the case in which the reported input alignments were obtained using IBM Model 1 and IBM Model 4. The difference in performance of these two methods is known to be large; under similar conditions in a German-English evaluation the difference in AER was reported to be 9.3 absolute (Och and Ney, 2003).

[2]The results in Ittycheriah and Roukos (2005) are reported in terms of alignment F-score. However, they point out that because their evaluation data for alignment contained only sure links (Section 2), we can obtain alignment error rate simply by subtracting the F-score from 1. We have done this here.

[3]It should be noted that the AER numbers reported in these experiments are not necessarily comparable. AER is sensitive to annotation differences, and in particular to the presence or absence of probable links (Section 2). For a thorough explanation refer to Fraser and Marcu (2006).

[4]Och and Ney (2000) report a similar relationship between AER and the word error rate metric for translation.

2 Word-Based Alignment

Word alignment originated in the training step of word-based translation models (Brown et al., 1993). In these models, the units of correspondence between sentences are individual words, and so word alignment corresponds exactly to the translation model. Over the past decade, a number of additional uses have been found for it, including the automatic acquisition of bilingual dictionaries (e.g. (Melamed, 1996; Resnik et al., 2001)) and cross-lingual syntactic learning (Yarowsky et al., 2001; Lopez et al., 2002; Smith and Smith, 2004; Hwa et al., 2005). For this reason, it is a topic of significant study in its own right.

For translation, the only truly important metric is the translation metric. However, since word alignment has uses outside of learning translation models, many word alignment studies report results using intrinsic metrics, which we briefly review here.

Formally, we say that the objective of the word alignment task is to discover the word-to-word correspondences in a sentence pair $(F = f_1 ... f_I, E = e_1 ... e_J)$ in which the source and target sentences contain I and J words, respectively. The alignment A of this pair is simply a set of these correspondences. We say that $A \subset \{1,2,...,I\} \times \{1,2,...,J\}$. If $(i,j) \in A$, then the ith source word is aligned to the jth target word.

Intrinsic evaluation is performed by comparison of the alignment set A with alignments created by human annotators. Annotations may contain two sets of links: the sure set S, containing only links about which all annotators are certain, and the probable set P, which includes all links in S as well as

links that were uncertain (Och and Ney, 2000).[5]

Given the set of hypothesized alignment links A, we compute the standard metrics precision (P), recall (R), and alignment error rate (AER) as follows:

$$Precision = \frac{|A \cap P|}{|A|}$$
$$Recall = \frac{|A \cap S|}{|S|}$$
$$AER = 1 - \frac{|S \cap A| + |P \cap A|}{|S| + |A|}$$

3 Phrase-Based Statistical Machine Translation

Phrase-based models represent the current state-of-the-art in statistical machine translation. In phrase-based models, the unit of translation is any contiguous sequence of words, which we call a phrase. Each phrase $\tilde{e}$ in E is nonempty and translates to exactly one nonempty phrase $\tilde{f}$ in F. This is done using a simple mechanism.

1. The source sentence is segmented into phrases.
2. Each phrase is translated.
3. The translated phrases are permuted into a final order.

The set of rules which governs this process is contained in a phrase table, which is simply a list of all source phrases and all of their translations. The phrase table is learned from the training data. The other rules (segmentation and permutation) are applied as described and do not need to be learned (in many systems, weights are not even learned for these rules). A derivation D consists of the set of rules used during decoding.

The decisions are guided by a log-linear model which scores each candidate translation.

$$E = \underset{\hat{E}}{\operatorname{argmax}} \langle w, \Phi(\hat{E}, F) \rangle$$

This model finds the candidate translation E that maximizes the dot product between a weight vector w and a vector Φ mapping each $\hat{E}, F$ pair onto a feature space.

In a departure from the large, sparse feature spaces consisting mainly of indicator functions, found elsewhere in natural language processing (Ratnaparkhi, 1996; Taskar et al., 2004), translation models typically use a small feature space in which all features are active, and have non-integer values. These features are estimated using maximum likelihood methods. While the former approach has attractive properties and can be optimized directly for translation accuracy metrics, as has been shown for other NLP tasks, it is very slow (Taskar et al., 2004). Maximum likelihood probabilistic estimation is much faster, and for this reason it is very attractive for machine translation, where very large corpora are used, and efficiency is at a premium. To train the small number of log-linear feature weights, we use minimum error rate training (Och, 2003).

The baseline translation model we consider in the following sections has eight features, following the example of the phrase-based Pharaoh system (Koehn, 2004).

1. A conditional phrase-to-phrase model that incorporates the probability of each phrase pair used in the derivation D (Equation 1).
2. The inverse conditional phrase-to-phrase probability model (Equation 2).
3. A lexical weighting feature (Equation 3). This feature operates over word alignments within phrase pairs.
4. The inverse lexical weighting (Equation 4).
5. A trigram language model feature.
6. A distortion count feature (Marcu and Wong, 2002; Koehn et al., 2003).
7. A feature counting the number of phrase pairs used in the translation.
8. A feature counting the number of target words.

$$\prod_{(\tilde{e},\tilde{f}) \in D} p(\tilde{e}|\tilde{f}) \quad (1)$$
$$\prod_{(\tilde{e},\tilde{f}) \in D} p(\tilde{f}|\tilde{e}) \quad (2)$$
$$\prod_{j=1}^{J} \left[\sum_{i:(i,j) \in A} 1\right]^{-1} \sum_{i:(i,j) \in A} p(e_j|f_i) \quad (3)$$
$$\prod_{i=1}^{I} \left[\sum_{j:(i,j) \in A} 1\right]^{-1} \sum_{j:(i,j) \in A} p(f_i|e_j) \quad (4)$$

The first five of these are probabilistic. They are converted to features by taking the negative logarithm. The first four are dependent in some way on the input alignments, as we will show. This configuration is representative of many phrase-based systems.

[5] Since it can be difficult for annotators to rate certainty of links, S sometimes contains links specified by *all* annotators, with P containing links specified by *any* annotator.

3.1 Word-Based Alignment and Phrase-Based Translation

The interaction between word alignments and phrase-based translation occurs in the learning step. Learning in statistical machine translation consists of two tasks: rule extraction and parameter estimation. In phrase-based systems, rule extraction consists of producing a phrase table by finding all corresponding phrases in the training data. This can be done using word alignments by extracting all phrases that are consistent with the word alignment. Consistency is defined as follows: if the source word f_i is aligned with the target word e_j, then a phrase pair containing f_i must also contain e_j; likewise, a phrase pair containing e_j must also contain f_i. Phrase pairs containing neither f_i nor e_j are not constrained in any way by the alignment point (i, j) (Och et al., 1999).[6]

Parameter estimation consists of maximum likelihood estimation for probabilistic submodels, and minimum error rate training. Alignment is used in training the lexical weighting features in the system we have described.

Word alignments can affect learning in three ways.

1. Alignments affect the phrase pairs that are extracted from the training corpus. The inventory of phrase pairs determines the space of translations that the model is capable of producing. Word alignment affects the recall of phrase extraction. In particular, an alignment error may cause the extraction to miss some phrase pair that is critical for translation of a particular source phrase. If this happens, the system will be unable to translate that phrase correctly.

2. Alignments affect the phrase probability feature. This is to some extent the inverse of the previous effect. An alignment error may result in the extraction of a phrase that is not a good translation. As such erroneous phrases pollute the phrase table, the phrase translation feature degrades.

3. Alignments affect the quality of the lexical weighting feature. Here the correlation is direct: poor alignment will cause this feature to favor translation featuring word pairs which are not translations of each other, while ignoring word pairs which are translations of each other.

With these points in mind, there are a few possible causes for the seeming disconnect between alignment improvement and translation improvement, each of which is worth considering.

1. As discussed above, it could be that alignments are useful to some aspect of the translation process, and harmful elsewhere. One way to investigate this is to tease apart the various effects of alignment on the translation process. This is the approach we take.

2. It could be that the relationship is obscured by the use of poor metrics for one or both tasks. The controversy over good evaluation metrics for MT is longstanding and beyond the scope of this paper. The question of alignment metrics is actually closer to the problem at hand. Assuming that we have decided upon a satisfactory translation metric, one possible approach would be to optimize our alignment for different alignment metrics, in order to see which one best correlates with the final MT metric. A slightly different approach would be to create a parameterized alignment metric, and tune its parameters for MT output performance using logistic regression or similar techniques. Some of these issues are explored by Ayan and Dorr (2006) and Fraser and Marcu (2006). In this paper, we do not address the issues of specific metrics. Our experiments address the issue of alignment quality directly by using alignments whose qualitative rankings are consistent across all metrics.

3. The answer could be the obvious one: phrase-based translation is simply insensitive to the quality of the underlying alignment. It may simply be that the quality of word alignment links does not significantly impact the quality of the extracted phrase tables.

4 Experiments

All of our experiments were performed on Chinese-English translation in the news domain. The data we used in our experiments were divided into four parts. For phrase extraction and training of submodels, we used a large training set consisting of over 1 million sentences from various newswire corpora. This corpus is roughly the same as the one used for large-scale experiments by Chiang et al. (2005). We

[6] An anonymous reviewer nicely summed up the relationship between word alignment and phrase extraction: "a couple currently if uneasily holding hands on the road to high-quality machine translation."

Training Data (various news)	
Sentences	1041792
English tokens	30175414
Chinese tokens	27379211
Alignment Test Set (MT Eval 2002)	
Sentences	441
English tokens	12123
Chinese tokens	10878
MERT Development Set (MT Eval 2003)	
Sentences	919
English tokens*	28445
Chinese tokens	27045
Translation Test Set (MT Eval 2005)	
Sentences	1082
English tokens*	34563
Chinese tokens	33216

Table 2: Characteristics of the experimental data. *For the MERT development and translation test sets, we show the average number of English tokens over four reference sets.

included in the training data a small set of sentences from MT Eval 2002 for which manual alignments were available. The alignments contained both sure and probable annotations. These were used to measure the accuracy of the word alignment methods used on the corpus. We used MT Eval 2003 as our development set for minimum error rate training. We used MT Eval 2005 as our translation test set. The details of the corpora are described in Table 2.

To generate alignments, we used GIZA++ (Och and Ney, 2003). We symmetrized bidirectional alignments using the grow-diag-final heuristic (Koehn et al., 2003). Although the accuracy of this method has been surpassed by numerous supervised methods in the last few years, particularly for small corpora, it still produces very good alignments for large corpora. The AER of the alignments we obtained (.226) was only slightly worse than the AER obtained (.197) by a supervised system on the same set (Ayan et al., 2005), although precision and recall profiles are different. We caution that these results are not directly comparable due to differing tokenization and the use of cross-validation for the supervised method. However, we believe that the GIZA++ alignments on this corpus are reasonably close to state-of-the-art. We refer to this alignment as Best.

We wished to avoid confounding our study by considering alignments with vastly different profiles, such as recall-oriented alignments versus precision-oriented alignments. For our purposes, we were interested in the impact of the quality of alignments, without regard to any particular quality metric. To this end, we designed a set of experiments in which we were able to compare alignments whose quality was consistent across metrics. We did this by creating a set of degraded alignments as follows: for each number of chunks $n \in \{1, 10, 100, 1000\}$: (a) divide the corpus into n equal-sized chunks, (b) align each chunk individually, and then (c) concatenate the results. Regardless of n, the resulting aligned bitext is the same size once the results have all been concatenated back together. As the number of chunks increases, the size of each chunk decreases, and the quality of the alignment degrades. We refer to the non-degraded alignment (full use of the available data) as Best, and degraded alignments with chunk sizes of 10, 100, and 1000 are referred to as Slightly Degraded, Moderately Degraded, and Highly Degraded, respectively.

Each of these four alignments contained similar numbers of alignment links. Furthermore, as desired, the accuracy of the alignments degraded consistently across all of the alignment metrics that we considered, so we are sure that our results are not affected by uncertainties regarding the efficacy of any one particular metric. The performance of our input alignments under AER, precision, and recall is reported in Table 3.

We ran minimum error rate training for each run reported in the sections that follow. Specifically, each row of Tables 4, 5, 6, and 8 represents a separate run of minimum error rate training. In our decoder we used a distortion limit of 4, translation table limit of 20, and a probability threshold of .0001 for pruning. The settings are similar to the default settings used by Pharaoh (Koehn, 2004).[7] To measure translation accuracy, we used the BLEU score

[7] Our decoder is a clone of Pharaoh, written by David Chiang. In preliminary experiments, we found the performance to be very similar to Pharaoh's.

Alignment	Prec.	Rec.	AER
Best	**.7089**	**.8625**	**.2258**
Slightly Degraded	.6510	.8251	.2765
Moderately Degraded	.5680	.7579	.3553
Highly Degraded	.5409	.7209	.3860

Table 3: Quality of the alignments used in the experiments, as measured by different alignment metrics. The alignment qualities correspond to division of the training corpus into 1, 10, 100, and 1000 equal-size chunks, respectively.

(Papineni et al., 2002).

4.1 Phrase Tables and the Translation Search Space

As we noted previously, an alignment error may prevent extraction of phrases that are critical to translation. Therefore, we wanted to measure the impact of the extracted phrases on the space of translations that can generated by our decoder.

We computed the source word coverage of the development and test corpora using the extracted phrase tables. This tells us how many source language words the translation model is able to translate, either correctly or incorrectly. We found that for all alignment accuracies on both data sets, the coverage was over 99%. We conclude from this that source language coverage is not affected by alignment accuracy in large data conditions.

Each phrase table generates a different translation search space, so we measured the coverage of the search space in an oracle experiment. We generated 1000-best outputs for each test sentence the best performing run for each alignment (Section 4.4). We then selected from these the set of output sentences that maximized the BLEU score against the reference set. [8] The results are given in Table 4. Although

Alignment Quality	Oracle BLEU score	
	Dev	Test
Best	.429	.406
Slightly Degraded	.414	.391
Moderately Degraded	.399	.380
Highly Degraded	.396	.374

Table 4: Oracle translation results for different alignment accuracies, using 1000-best lists.

we see a significant drop of .033 in the BLEU score between the best and worst quality alignments, the worst oracle scores are still substantially better than the 1-best decoder output, with scores of .396 and .374 versus .304 and .279 for development and test, respectively. This means that even with phrase tables produced with very poor alignment accuracy, it is *possible* to find good translations. The task of actually finding them in this space must then fall to our log-linear model and feature set.

It is likely that the high oracle scores are at least partially attributable to the size of the training corpus. Even a poor alignment algorithm is likely to produce some good phrases. With a sufficient amount of training data, the cumulative effect over a sufficiently large dataset is that good translations are present in the search space regardless of alignment quality.

4.2 Phrase Translation Features

Since the search space is large enough to find good translations regardless of input alignment quality, we studied the impact of alignment quality on the features which are used to find translations within this space.

Two features depend on the input alignments: the phrase translation probability, and the lexical weighting probability. The phrase translation probability is affected indirectly: when poor translations are introduced into the phrase table due to alignment error, we expect noise to degrade the utility of this feature.

In order to tease apart the effects of the two

[8] Because the BLEU score is computed using aggregate statistics over the output, the locally best output for any given input sentence is not necessarily the one that results in the best overall BLEU score (indeed, due to BLEU's use of a geometric average, most single sentences turn out to have a BLEU score of 0, which is not very useful even for determining the locally optimal sentence). Computing the choice of sentences which results in the best global BLEU is an intractable search problem, so we resorted to a greedy hill-climbing search (Och et al., 2004; Venugopal and Vogel, 2005). This works as follows: we first choose for each input sentence an output that maximizes a local non-zero approximation to BLEU. We then iterate over our input sentences and at each step choose a new output from the 1000-best list that optimizes the global BLEU score while holding all the other outputs constant. This is repeated until no further gains in BLEU score can be found.

Alignment Quality	BLEU score	
	Dev	Test
Best	.293	.272
Slightly Degraded	.289	.268
Moderately Degraded	.278	.262
Highly Degraded	.274	.252

Table 5: Translation results obtained using without alignment-based lexical weighting features.

Alignment Quality	BLEU score	
	Dev	Test
Best	.301	.280
Slightly Degraded	.297	.277
Moderately Degraded	.286	.265
Highly Degraded	.281	.259

Table 6: Translation results obtained using lexical weighting features.

alignment-based features, we ran the decoder without the lexical weighting feature. This allows us to study the impact of the phrase translation feature in isolation. The results are given in Table 5. We find that the translation quality degrades significantly between the best and worst alignment accuracies, by a BLEU score of .02. However, we note that the corresponding alignment difference required to create this gap is much larger, about 14% absolute under each alignment metric. The difference produced by the two best alignment accuracies, separated by about 5% absolute under alignment metrics, is .004 BLEU. The differences in accuracy between current state-of-the-art alignment methods are often much less than this.

4.3 Lexical Weighting Features

To study the effect of the lexical weighting, we added it to the model. For each alignment, we included the lexical weighting feature and reran the experiment. The results are shown in Table 6. With one exception, the lexical weighting resulted in a consistent improvement of between .007 and .008 BLEU for each run. Notably, the lexical weighting feature was about equally helpful for all of the alignment accuracies.

From this experiment, we drew two conclusions. First, it seems that the lexical weighting provides a modest improvement over the system with no lexical weighting. This is consistent with the results of Koehn et al. (2003). Second, because the increases are largely consistent, we conclude that alignment quality has very little affect on the utility of this feature. Here again, we suspect that the amount of training data may be at play, providing enough signal enough through a substantial amount of noise in the poor alignments.

4.4 Feature Selection and Feature Engineering

In the previous sections we noted that translation quality consistently tracked alignment quality in the phrase probability feature, but not the lexical weighting feature. In order to gain more insight into this, we ran a simple feature selection experiment using the minimum error rate criterion. We used the forward selection algorithm, which is the feature selection method presented for the maximum entropy criterion by Berger et al. (1996). In this algorithm, we begin with an empty feature set, and iteratively add features. At each step, we add the feature that leads to the the largest decrease in error, as measured over 1000-best lists using the iterative line optimization algorithm of Och (2003). Because we had a very small number of features, we did not use a stopping criterion. Instead, we used it as a mechanism to evaluate the contribution of each feature. We ran the algorithm through eight iterations, using the system derived from alignments created using Best. At the end of each iteration, we computed the BLEU score for development set using only the current set of weights and features. The results are given in Table 7.

The results support the results of our previous experiments showing that the phrase probability feature is more important to the overall translation quality than the lexical weighting feature. Features for phrase probability, language model, and word count (which receives a negative weight, counterbalancing the other two) combined account for 78% of the absolute translation performance. Interestingly, the model disprefers probabilistic features in the form $p(e|f)$.

Significantly, note that many of the features here do not contribute much to the overall translation performance. Considering the seemingly low impact of

Rank	Feature	BLEU
1	phrasal $p(\tilde{f}\|\tilde{e})$	.057
2	language model	.169
3	word count	.237
4	lexical $p(f\|e)$	.236
5	reordering	.256
6	lexical $p(e\|f)$	.295
7	phrase count	.299
8	phrasal $p(\tilde{e}\|\tilde{f})$	.301

Table 7: Feature selection results. The BLEU score represents results using only features of the same rank or lower.

Alignment Quality	BLEU score	
	Dev	Test
Best	.304	.279
Slightly Degraded	.301	.279
Moderately Degraded	.288	.271
Highly Degraded	.286	.267

Table 8: Translation results obtained using the Model 1 feature.

these features, we believe it is possible that gains could be made through better feature engineering. In order to explore this idea further, we added an alignment-free lexical feature, the IBM Model 1 feature. This model is computed as follows.

$$\prod_{j=1}^{J}\sum_{i=1}^{I} p(e_j|f_i)$$

This differs from the lexical weighting feature in that it contains all words in both source and target, leading to what Och et al. (2004) refer to as a *triggering effect*. This can be thought of as a coarse form of word sense disambiguation. We wanted to see if such a feature might begin to overcome the deficiencies of the poorer phrasal translation feature. The Model 1 feature was shown to be beneficial by Och et al. (2004). It is estimated directly from the training corpus using expectation-maximization. No alignment is involved.

Our results using the Model 1 feature are given in Table 8. We see no change in the discrepancy between the development set BLEU scores, but something interesting happens in the test set. Surprisingly, the gap between the best and worst BLEU scores closes from .021 to .012. This mainly due to the fact that the best alignments do not improve, while the poor alignments do. It appears that the Model 1 feature helps to overcome the deficiencies of the features based on poor alignment. This is evidence that better feature engineering may be able to overcome differences in the quality of the input alignments.

5 Conclusions and Future Work

We have presented an extensive analysis of the relationship between word alignment quality and resulting phrase translation quality. We presented results on a large corpus, comparing alignment quality that is near state-of-the-art with a set of consistently degraded alignments. Our results confirm that, while there is a definite correlation between alignment and translation quality, it takes large gains in alignment performance under any metric to achieve relatively small gains in translation performance. Furthermore, our results show that this primarily stems from noise reduction in the phrase probability feature. This means that it may be more useful to directly investigate ways to reduce noise in phrase extraction, rather than approaching the problem indirectly via alignment improvement. We found that alignment quality has little impact on the lexical weighting feature, which itself provides only a modest improvement in translation quality. Furthermore, the translation search spaces resulting from all input alignment qualities contain much better translations than the ones we are currently able to find. This suggests that there is an opportunity for substantial gain in translation quality by designing features and learning algorithms that make better decisions in this search space. Furthermore, the range for improvement is much greater than it is for improving alignment quality. Our results also suggest that large data conditions may help to overcome poor alignment performance generally, although this point requires further investigation.

We illustrated that improvements in feature engineering may be sufficient to overcome deficiencies of poor alignments. We further illustrated a simple feature selection method that raised interesting questions about the features commonly employed in

phrase-based systems, which we intend to explore further. We noticed that the probabilistic features did not contribute equally to the overall model – in particular, we noticed that features in the form $p(f|e)$ outperformed features in the form $p(e|f)$. Although a good deal of this effect is probably a result of feature overlap, it may be beneficial to explore features that incorporate marginals over both e and f in the same statistic, such as log-likelihood ratios or Dice values. We plan to explore this in future work, along with more novel features.

Acknowledgements

This research has been supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001, ONR MURI Contract FCPO.810548265, and Department of Defense contract RD-02-5700. The authors would like to thank David Chiang for his implementations of the phrase-based decoder and minimum error rate training, Necip Fazil Ayan for illuminating discussions, and the anonymous reviewers for helpful comments.

References

Necip Fazil Ayan and Bonnie Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL'2006)*, Jul.

Necip Fazil Ayan, Bonnie Dorr, and Christof Monz. 2005. Neuralign: Combining word alignments using neural networks. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 65–72, Oct.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, Mar.

Alexandra Birch, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proceedings of the NAACL 2006 Workshop on Statistical Machine Translation*, pages 154–157, Jun.

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

Chris Callison-Burch, David Talbot, and Miles Osbourne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 176–183, Jul.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 779–786, Oct.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL)*, June.

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of the NAACL 2006 Workshop on Statistical Machine Translation*, pages 31–38, Jun.

Alexander Fraser and Daniel Marcu. 2006. Measuring word alignment quality for statistical machine translation. Technical Report ISI-TR-616, ISI-University of Southern California.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, Sep.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 89–96, Oct.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 127–133, May.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of The 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, Sep.

Adam Lopez, Michael Nossal, Rebecca Hwa, and Philip Resnik. 2002. Word-level alignment for multilingual resource acquisition. In *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and Representation – Bootstrapping Annotated Language Data*, Jun.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 133–139, Jul.

I. Dan Melamed. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the Second Meeting of the Association for Machine Translation in the Americas (AMTA)*.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 81–88, Oct.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 1086–1090, Jul.

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

Franz Josef Och, Christoph Tillman, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, pages 20–28, Jun.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, pages 161–168, May.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul.

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

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 133–142, May.

Philip Resnik, Douglas Oard, and Gina Levow. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the Proceedings of the First International Conference on Human Language Technology Research (HLT)*, San Diego, CA, March.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 755–762, Oct.

David A. Smith and Noah Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 49–56, Jul.

Ben Taskar, Dan Klein, Michael Collins, Daphne Koller, and Christopher Manning. 2004. Max-margin parsing. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–8, Jul.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 73–80, Oct.

Ashish Venugopal and Stephan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, pages 271–279, May.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing mulilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the Proceedings of the First International Conference on Human Language Technology Research (HLT)*, pages 109–116.

Bing Zhao and Alex Waibel. 2005. Learning a log-linear model with bilingual phrase-pair features for statistical machine translation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 79–86, Oct.