Word Alignment Revisited

Francisco Guzman
Centro de Sistemas Inteligentes
ITESM
Monterrey,Mexico
guzmanhe@gmail.com

Jan Niehues
Institut für Theoretische Informatik
Universität Karlsruhe (TH)
Karlsruhe, Germany
jniehues@ira.uka.de

Qin Gao
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
qing@cs.cmu.edu

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
stephan.vogel@cs.cmu.edu

1 Introduction

Word alignments have been considered the backbones of Statistical Machine Translation. Even when Statistical Machine Translation has shifted from a word-based to a phrase-based paradigm, the word alignment has remained the base for most phrase-based [Koehn et al., 2003] and syntactic augmented phrase based SMT systems [Zollmann and Venugopal, 2006, Chiang, 2007, Marcu et al., 2006].

Most SMT systems use the freely available GIZA++ Toolkit [Al-Onaizan et al., 1999] to generate the word alignment. This toolkit implements the IBM Models and the HMM model introduced in [Brown et al., 1990, Vogel et al., 1996].

Generative models have the advantage that they are well suited for a noisy-channel approach. Unsupervised training can be used to align large amount of unlabeled parallel corpora. Nonetheless they have a major disadvantage: because these models are completely unsupervised, they can hardly make use of the increasingly available manual alignments. Also, given their complexity, to incorporate other sources of informations such as POS tags, word frequencies etc., is a non-trivial task. Moreover, because the IBM models are not symmetric, the alignments for different directions are quite different, which makes the search for a symmetrized combination of the word alignments a challenging procedure.

Given the fact that the word alignments serve as a starting point of the SMT pipeline, improving their quality has been a major focus of research in the SMT community. However, due to the amount of processing that a word alignment undergoes before being used in translation (for example, phrase extraction), the quality of word alignment is not directly related to the quality of translation. In fact, only weak correlation between alignment error rate (AER) and BLEU scores has been reported [Fraser and Marcu, 2006a]. The mismatch between the quality of word alignment models and that of phrase-based or syntactic based SMT may lead to the phenomenon of improved translation quality resulting from "degraded" alignment quality [Vilar et al., 2006]. This calls for more careful analysis of word alignment errors. There has been little effort doing a thorough error analysis of the alignment process. As a result, the role of the quality of word alignments in machine translation remains rather unclear.

Recently, different efforts have focused on the symmetrization of the word alignment models [Matusov et al., 2004, Liang et al., 2006], the inclusion of annotated data in the training of generative models [Fraser and Marcu, 2006b], and the use of discriminative models [Blunsom and Cohn, 2006, Taskar et al., 2005, Niehues and Vogel, 2008]. One of the advantages of the latter models is that the word alignment quality can be tuned towards a given word alignment quality measurement[1]. Moreover, their conditional probability model allows the inclusion of different features, enabling that any available knowledge source can be used to find the best alignment.

In this work, we present the results of an extensive error analysis of the alignments created by the generative models using GIZA++. By characterizing the errors, we hope to shed light on the behavior of the aligners, as well as to identify some opportunities for improvement. We also present our work on a discriminative word alignment framework, as presented in [Niehues and Vogel, 2008], which is easy to enhance with new features. We believe that with a proper analysis of the alignment behaviors, coupled with the use of discriminative word aligner, can help to overcome many of the weaknesses of the generative models.

The paper is organized as follows. In Section 2, the analysis of the alignment errors is presented. In Section 3 the discriminative aligner is introduced, along with proposed new features. The alignment experiments and analysis are presented in Section 4.

[1] For some measurements, smoothing is required.

2 Error Analysis

When analyzing the errors made by the automatically generated word alignments we compare the Viterbi alignments generated by GIZA++ against a gold standard of hand aligned data. In some gold standards, there is a distinction between Sure and Possible links. Sure links represent the hand alignments made by the annotator for which he is sure of the alignment. Possible links are those which represent a degree of uncertainty, e.g. were different annotators differ in the manual alignment. In their study, [Fraser and Marcu, 2006a] discourage the use of Possible links. They argue that they induce flaws. Therefore, in this study, we include only Sure links.

Based on the differences between Viterbi and hand alignments, there are three basic quantities that we can measure: the number of links in which these two alignments agree, i.e. true positives (tp); the number of links that are present in the output of the aligner but not in the gold standard, i.e. false positives (fp); and the number of links that are present in the gold standard, but not in the output of the aligner, i.e. false negatives (fn). There are several metrics that are used for measuring the quality of a word alignment, but most of them are based on these three quantities.

- Alignment Error Rate: AER, as defined in [Och and Ney, 2003], takes Sure and Possible links into account.

$$AER = 1 - \frac{|A \cup S| + |A \cup P|}{A + S}$$

  However when we only have Sure Alignments, the metric is related to the F measure :

$$AER = 1 - \frac{2tp}{2tp + fp + fn} = 1 - F \quad (1)$$

- Precision: This measure gives us a notion of how accurate is the output of the aligner. It is the ratio of the correct to all generated links.

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

- Recall: This measures how well we cover the desired links, i.e. those in the hand alignments, with the automatically generated ones. That is, of all the links in the gold standard, what is the amount of links that are also present in the aligner output.

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

For some time, AER has been the predominant metric in the area. However, recent studies have suggested that it does not correlate as well to the quality of Machine Translation as some variations of the F-measure [Fraser and Marcu, 2006a]. Despite that, [Vilar et al., 2006] encourage the use of AER as an alignment quality metric. They argue that the discrepancy between BLEU and AER is a result of the mismatch between the alignment and translation models. In this study, we use AER, given that it is a widely metric and that the balance between precision and recall is already fixed.

2.1 Data Analysis

In this section, we present a detailed error analysis of the Viterbi alignments, resulting from performing training through the standard sequence of word alignment models IBM1, HMM, IBM3 and finally IBM4, in both directions, i.e. source to target (S2T) and target to source (T2S). We use the modified GIZA toolkit [Gao and Vogel, 2008]. In addition, we generated the combined alignment, using the grow-diag-final heuristics implemented and used in the MOSES package[Koehn et al., 2007]. Our analysis was performed on alignments for Arabic-English and Chinese-English. The data sets using for training are displayed in Table 2. The data sets that we used are for evaluation are summarized in Table1.

Table 1: Data Statistics for the Evaluation set used in our Analysis

	#Sentences	#Words	Avg. SenLen (stdev)
Arabic - English			
Arabic	14K	287K	20.547 (10.21)
English	14K	358K	25.709 (12.88)
Chinese - English			
Chinese	19K	375K	19.384 (12.20)
English	19K	463K	23.942 (15.86)

Table 2: Statistics of corpora used in GIZA++ training

	#Sentences	#Words
Arabic - English		
Arabic	7.7 M	218M
English	7.7 M	216M
Chinese - English		
English	11.0 M	309M
Chinese	11.0 M	273M

From the statistics for the analyzed set we can see that the English sentences are on average longer that the Arabic and Chinese sentences. Therefore, it is to be expected that source words often need to be aligned to several English words (i.e English words have higher fertility), or that many English words are not aligned to Chinese or Arabic source words. As we know, this poses a problem for the generative models we are analyzing, because

they can align at most one source word to a target word. Therefore a greater asymmetry is to be expected.

In Table 3, we summarize the results for the evaluation of Arabic-English and Chinese-English. As expected, IBM4 models from source-to-target perform worse than the target-to-source ones. This difference is more striking for the Chinese-English case, where the AER difference is almost 20 points. What can be observed from the data, for source to target cases, is that the aligner has too many missing links, which makes the resulting alignments more precision oriented. This can be interpreted as Chinese and Arabic words being aligned to more than one English word in the gold-standard, which cannot be directly achieved in the IBM Models.

2.2 Unaligned Words

The generative word alignment models are not symmetric and therefore we may expect unbalance between the number of source and target words that are left unaligned. Looking at the statistics of unaligned words for the output of GIZA++, as well as the behavior of symmetrization heuristics, we can get a better sense of the impact of the underlying structure of these models. Table 4 gives the percentage of NULL-Alignments (source words left unaligned) and the percentage of target words not aligned. We also show these numbers for a symmetrized alignment and for the discriminative alignment described later in section 3.

Table 4: Statistics of unaligned words and null-alignment

Alignment	NULL Alignments	Unaligned Words
Arabic - English		
Manual Alignment	8.58%	11.84%
IBM4 S2T	3.49%	30.02%
IBM4 T2S	5.33%	15.72%
Combined	5.53%	7.79%
DWA	17.64%	18.68%
Chinese - English		
Manual Alignment	7.80%	11.90%
IBM4 S2T	5.46%	23.84%
IBM4 T2S	6.41%	34.53%
Combined	9.80%	14.64%
DWA	16.70%	23.67%

Compared to the hand aligned data, GIZA++ tends to have less NULL alignments, but a larger number of unaligned words. Depending on the language considered as source and target, we see that up to one third of the target words are not aligned. This highlights a weakness of the generative alignment models. On the other hand, the heuristic symmetrization (grow-diag-and) generates alignments which are more balanced between the number of NULL alignments and unaligned words, and reasonably close to the the hand alignment. In contrast, the discriminative alignment is sparser, leaving more source and target words unaligned. This is specially true for Chinese-English where we observe a large number of unaligned English words, which will impact the phrase pairs extracted from this kind of alignment.

2.3 Alignment Errors: Sentence Level Analysis

In the following subsection we analyze the distribution of alignment errors over all sentence pairs.

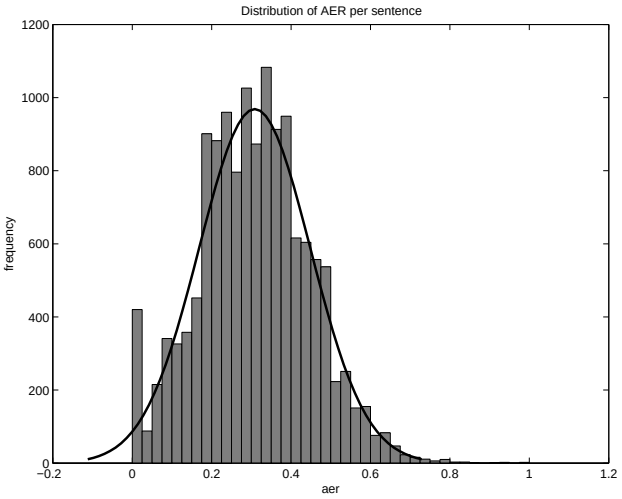

Figure 1: AER per Sentence distribution of the Arabic-English alignment generated by GIZA++ S2T. In the vertical axis, we observe the counts of sentences versus their corresponding AER evaluation in the horizontal axis

The distribution of the per-sentence AER shown in Figure 1 is almost Gaussian. However, we see a peak at AER=0. This is partially an artifact of the bucketing: to achieve an AER of 0.1 the sentence has to be already in the range of 10 words. What is interesting, however, is the observation that even longer sentences are sometimes perfectly aligned, as is shown in Table 5.

Table 5: Number and percentage of perfect alignments by source sentence length of the Arabic-English alignment generated by GIZA++ S2T.

SenLen	#Perfect	Percent
1	21	1.00
2	14	0.93
3	14	0.50
5	235	0.179
7	396	0.164
10	604	0.0563
15	377	0.0053

Table 6 shows some Arabic-English sentence pairs, which were perfectly aligned. We observe that all the

Table 3: Statistics of the Different Alignments (GS = Gold Standard Alignment)

Alignment	#Links GS	#Links	Correct	Misaligned	Missing	Precision	Recall	AER
Arabic - English								
IBM4 S2T	336,995	275,386	202,898	72,488	134,097	73.68	60.21	33.73
IBM4 T2S	336,995	339,281	232,840	106,441	104,155	68.63	69.09	31.14
Combined	336,995	334,469	244,817	89,652	92,178	73.20	72.65	27.08
Chinese - English								
IBM4 S2T	527803	359485	186,620	172,865	341,183	51.91	35.36	57.93
IBM4 T2S	527803	451222	299,744	151,478	228,059	66.43	56.79	38.77
Combined	527803	437241	296,312	140,929	231,491	67.77	56.14	38.59

perfectly aligned sentence pairs have no long distance reordering.

Table 6: Examples of perfectly aligned sentences of the Arabic-English alignment generated by GIZA++ S2T.

الملك حسين يبدا المرحلة الرابعة من العلاج الكيميائي

King Hussein Begins Fourth Phase of Chemotherapy
1-1 2-2 3-3 5-4 4-5 6-6 7-7 8-7

روبن كوك يستبعد اي تدخل للحكومة البريطانية في قضية بينوشيه

Robin Cook Dismisses Any British Government Intervention in Pinochet Case
1-1 2-2 3-3 4-4 7-5 6-6 5-7 8-8 10-9 9-10

وتستخدم اليونيسكوم في العراق نحو 120 شخصا بينهم 40 مفتشا .

UNSCOM employs about 120 persons in Iraq , including 40 inspectors .
2-1 1-2 5-3 6-4 7-5 3-6 4-7 8-9 9-10 10-11 11-12

Finally, it is interesting to see how well the alignment model scores correlate to the alignment quality. A high correlation would allow us to select sentence pairs, for which the alignment is more likely correct. In Figure 2 a scatter plot shows the correlation between the alignment error rate and the normalized alignment log probability (i.e. divided by the number of words in the sentence). We see a weak correlation, which might not be sufficient for a reliable data selection. We also observe a series of bands (at approx 0.3, 0.5, 0.6, etc.) which correspond to the discrete nature of AER, i.e. a sentence pair of length 2/2 can only have AERs of $\{0, 0.25, 0.5, 0.75, 1\}$.

2.4 Alignment Errors: Word Level Analysis

The distribution of words in the training corpus typically follows a Zipf curve. This leads to the question if generating correct alignments is more problematic for very low frequency words - as there is not much evidence from the data - or for high frequency words - as they are seen co-occurring with almost all of the other words. In Figure 3 the AER with respect to word frequency is displayed, showing that indeed both high frequency words and low frequency words have higher alignment error rates than

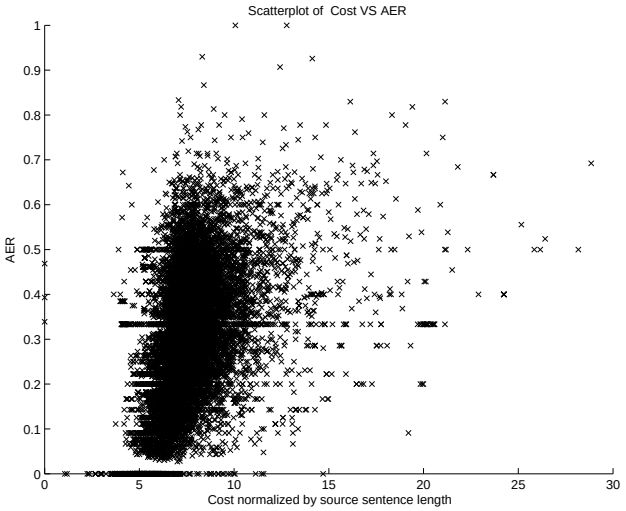

Figure 2: Scatter plot of AER per Sentence versus Model Cost of the Arabic-English alignment generated by GIZA++ S2T. In the vertical axis, we observe the per sentence AER versus the normalized model cost from GIZA++ in the horizontal axis

the mid frequency words. The word frequencies for this analysis were estimated with the counts over the corpus in tables 1,2.

The errors in high frequency words are, of course more problematic, as they add more to the overall alignment error rate. We therefore analyze the errors for these words in more detail. As we saw in table 3, there are some disparities between the number of errors for the two alignment directions, in particular the false negatives differ a lot. We therefore analyze which words contribute most to these errors. In Figure 4, we see the distribution of the misalignments (fp) and missing links (fn) by the most frequent English words in the GIZA source to target alignments for Chinese.

We can see that for Chinese the most frequently misaligned word is the stop sign. On the other side, the word for which the required link is missing most often is the word “the” on the English side. In Table 7 we present some examples that illustrate those cases. As we can

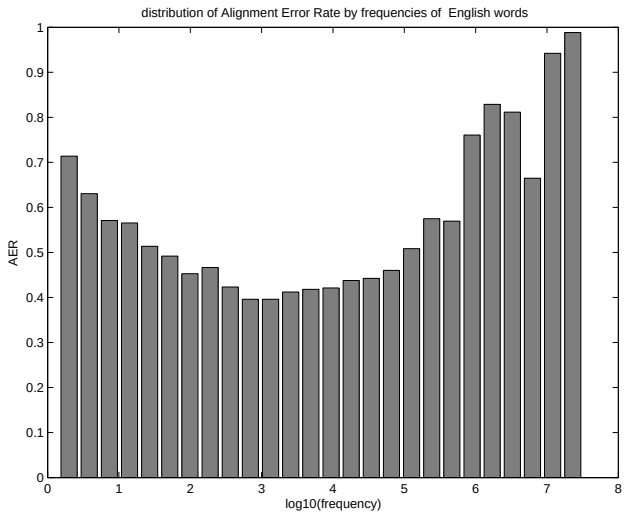

Figure 3: Distribution of AER per frequency of English words of the Chinese-English alignment generated by GIZA++ S2T. In the vertical axis, we observe the per word AER versus the log frequency of the English words in the horizontal axis

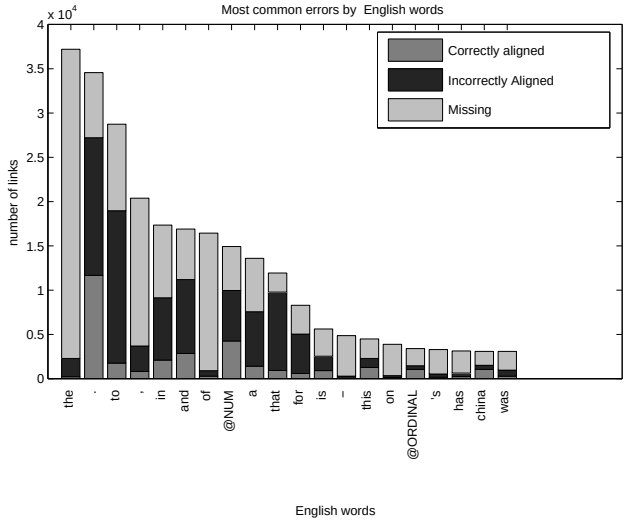

Figure 4: Alignment errors for high frequency English words from the Chinese-English alignment generated by GIZA++ S2T

see from the distribution, the dot is misaligned more than half of the time. Other punctuation marks have also high alignment error rates.

One conclusion that we can obtain from looking at these tables is that for Chinese, very frequent words such as the stop sign get frequently aligned to content words. On the other hand, there are many cases in which the English determiner "the" is missing from an alignment. These kind of behavior could be easily avoided if we could integrate some linguistic information into the generative models. Fortunately, there are other alignment models (such as the Discriminative Model) which enable us to do so.

Table 7: Summarization of alignments of high frequency English words "the" and "." extracted from Chinese-English alignment generated by GIZA++ S2T

Stop sign "."				**Determiner "the"**			
Chinese Word	TP	FP	FN	Chinese Word	Agrmt	FP	FN
,	1176	6540	4114	@NUM	3	19	573
。	10325	2231	2071	美国	0	2	507
”	3	389	9	这	96	200	393
了	0	248	0	的	1	3	367
嗯	1	229	0	政府	0	3	363
的	0	177	1	一	10	54	297
)	0	116	3	国	2	12	293
对	0	110	0	公司	0	0	213
报道	0	98	0	问题	0	4	211
@NUM	1	97	15	世界	0	0	208

3 Discriminative Aligner

In recent years several authors [Moore, 2005, Taskar et al., 2005, Blunsom and Cohn, 2006] proposed discriminative word alignment frameworks and showed that this leads to improved alignment quality. One main advantage of the discriminative framework is the ability to use all available knowledge sources by introducing additional features. In this study, we use the discriminative model presented in [Niehues and Vogel, 2008], which uses a conditional random field (CRF) to model the alignment matrix. The alignment matrix is described by a random variable y_{ji} for every source and target word pair (f_j, e_i). These variables can have two values, 0 and 1, indicating whether these words are translations of each other or not. By modeling the matrix, no restrictions to the alignment are required and even many to many alignments can be generated. As a result, the model is symmetric, and therefore will produce the same alignment regardless of the direction.

The structure of the CRF used in this discriminative framework is described by a factored graph, with two different types of nodes: hidden nodes, which correspond to the random variables y_{ji}, and the factored nodes c. The factored nodes define a potential Φ_c on the random variables V_c they are connected to. This potential is used to describe the probability of an alignment based on the information encoded in the features. This potential is a log-linear combination of some features $F_c(V_c) = (f_1(V_c), \ldots, f_n(V_c))$ and it can be written as:

$$\begin{aligned} \Phi_c(V_c) &= exp(\Theta * F_c(V_c)) \\ &= exp(\textstyle\sum_k \theta_k * f_k(V_c)) \end{aligned} \quad (4)$$

with the weights Θ. Then the probability of an assignment of the random variables, which corresponds to a word alignment, can be expressed as:

$$p_\Theta(y|e,f) = \frac{1}{Z(e,f)} \prod_{c \in V_{FN}} \Phi_c(V_c) \qquad (5)$$

where V_{FN} is the set of all factored nodes in the graph, and $Z(e,f)$ is a normalization factor

The structure of the described CRF is quite complex and there are many loops in the graph structure, so the inference cannot be done exactly. Consequently an approximation algorithm has to be used. For this implementation, the belief propagation algorithm introduced in [Pearl, 1988] is used. This algorithm is not exact in loopy graphs and it is not even possible to prove that it converges, but in [Yedidia et al., 2003] it was shown, that this algorithm leads to good results.

The weights of the CRFs are trained using a gradient descent for a fixed number of iterations, since this approach leads already to quite good results. The default criteria to train CRFs is to maximize the log-likelihood of the correct solution, which is given by a manually created gold standard alignment. However, in our experiments, the tuning process consists on a two step optimization: first we optimize towards Maximum likelihood, then towards AER. This sequence has been observed to provide the best results [Niehues and Vogel, 2008].

3.1 Baseline Features

In the model we are using there are three different types of factored nodes corresponding to three groups of features. The first group of features are those that depend only on the source and target words and are called local features. For instance, the lexical features, which represent the lexical translation probability of the words, belong to this group. In addition, there are source and target normalized lexical features for every lexicon. The source normalized feature, for instance, ensures that all translation probabilities of one source word to target words in the sentences sum up to one. The next group of features are the fertility features. They model the probability that a word translates into one, two, three or more words, or does not have any translation at all. Third, the first-order features model the first-order dependencies between the different links. For example, the link between source word i and target word j may depend on whether word $i+x$ and $j+y$ were linked. We denote the feature as (x,y), and any arbitrary number of first order features can be supplied. For further reference on the detail of these features, please see [Niehues and Vogel, 2008]

For our experiments the features for the baseline are set as follows: IBM4 lexica both directions, IBM4 fertilities, IBM4 viterbi alignments as well as source/target normalization features, identity feature, relative position feature, as well as the following directions for the first order features: (1,1),(1,2),(2,1),(1,-1),(0,1),(1,0).

3.2 New Discriminative Features

One of the purposes of performing the extensive analysis we did in Section 2 was to recognize the weaknesses of the generative models and therefore incorporate new knowledge into the DWA in the form of new features. Such features, along with the existing ones are expected to refine the alignment models and to result in an improvement of alignment quality. To facilitate the integration of such information, we have devised a new class-based feature, which would score the GIZA++ alignments that are used for discriminative training, according to the class of the source and target words.

This class-based feature is a local feature and is defined as the conditional probability of a link between words f_j and e_i given the classes of the words $C(f_j)$ and $C(e_i)$; and the alignments provided by GIZA++ in both directions. Using this feature, we can couple any kind of information regarding the class of words (whether they are POS tags, frequency based classes, etc.) with the alignment information from GIZA++. Then, we use this probability to provide a degree of confidence for the GIZA++ alignments.

This the degrees of confidence for the GIZA++ alignments have to be estimated according to labeled data, during a feature training phase. Then, this information is integrated into the DWA framework as a weighted link feature. The estimation for class-based feature is obtained through a simple MLE:

$$\begin{aligned} p(y_{i,j}|...) &= p(y_{j,i}|C(f_i), C(e_j), L_{s2t}(j,i), L_{t2s}(j,i)) \\ &= \frac{|C(f_j), C(e_i), L_{s2t}(j,i), L_{t2s}(j,i), L_{ha}(j,i)=1|}{|C(f_j), C(e_i), L_{s2t}(j,i), L_{t2s}(j,i)|} \end{aligned}$$

Where $C(f_j)$ is the class of the source word, $C(e_i)$ is the class of the target word, $L_X(j,i)$ a binary function which tell us if there is a link between positions the words f_j and e_i in a certain X alignment set (source to target, target to source, hand aligned).

Currently, we have tested this kind of feature using a frequency-based classification of the source and target words. That is, split the words into different classes according to their frequency. These classes are computed in such a way that they have similar number of word-counts. In the current implementation, the number of frequency based classes is set to 10 for source and 10 for target words.

In the next section, we will briefly describe our most recent experiments using these features to improve alignment and translation quality.

4 Experiments and Results

In our experiments we wanted to compare the output of the Discriminative Word Aligner (DWA) to that of the Viterbi alignments from GIZA++, as well as the combined alignment typically used for phrase extraction. As we stated before, the DWA has allows combining several sources of information. Among them are the Viterbi alignments, the IBM-4 lexicons and IBM4 fertilities, and relative position features. For the experiments described in this section, we also tested the frequency-class feature described in Section 3.2. This feature was trained using the data set previously defined for analysis (Table 1). Table 8 shows the corpus statistics for the respective data sets, for Arabic-English and Chinese-English. In Table 9 the alignment results are shown.

Table 8: Data Statistics for Experimental Data Sets

	# Sentences	# Words
Arabic - English		
Arabic dev	100	2,518
English dev	100	2,964
Arabic test	2,552	62,970
English test	2,552	76,652
Chinese - English		
Chinese dev	500	10,285
English dev	500	12,632
Chinese test	2,000	39,052
English test	2,000	48,655

Table 9: Alignment quality results for the different aligners

Aligner	**Alignment Quality**			
	AER (dev)	Precision (test)	Recall (test)	AER (test)
Arabic - English				
GIZA S2T	30.59	71.89	67.43	30.41
GIZA T2S	32.53	64.49	72.74	31.60
Combined	27.84	67.98	**76.52**	28.00
DWA-Baseline	**24.25**	80.51	72.48	**23.72**
DWA+Freq-class	24.41	**83.71**	69.64	23.97
Chinese - English				
GIZA S2T	59.50	51.67	34.91	58.33
GIZA T2S	40.17	66.48	56.92	38.67
Combined	40.06	67.98	56.29	38.41
DWA-Baseline	37.14	72.93	**57.40**	35.76
DWA+Freq-class	**36.53**	**75.79**	56.71	**35.12**

From these results we can note the following: First, discriminative word alignment improves over just heuristically combining the Viterbi alignments from the generative alignment models. This is more significant for Arabic, where the gain is of 4.28 points. whereas for Chinese is only of 2.65. Adding the frequency feature improves the precision for both Arabic and Chinese, however leads also to a drop in coverage, giving a modest improve AER only for the Chinese to English aligment.

5 Conclusions and Future Work

We analyzed word alignment resulting from the standard generative word alignment models. This analysis revealed fundamental problems with those alignment models. Most notably, while all source words are aligned to exactly one target word or the NULL word, a high percentage of the target words are not aligned. While symmetrizing the alignment using heuristics to combine the Viterbi alignments for both directions reduces the number of unaligned words, it does not improve the alignment all that much. Discriminative word alignment leads to significant improvement over the generative alignment models. However, discriminative alignments, being more precision based, can lead to larger and sparser phrase-tables. In [Guzman et al., 2009] they addressed this possible drawback by using the number of unaligned features in a phrase pair as features in the phrase table, which led to good results. It is clear that in order to fully benefit from improvements in word alignment, it is imperative to understand the role of succesive stages in the SMT pipeline, such as phrase extraction and scoring, and not to focus only in translation results.

References

Y. Al-Onaizan, J. Cuřin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, N. A. Smith, F.-J. Och, D. Purdy, and D. Yarowsky. Statistical Machine Translation. CLSP Research Notes No. 42, Johns Hopkins University, 1999.

P. Blunsom and T. Cohn. Discriminative word alignment with conditional random fields. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

D. Chiang. Hierarchical phrase based translation. *Computational Linguistics*, 33(2):201–228, 2007.

A. Fraser and D. Marcu. Marcu. measuring word alignment quality for statistical machine translation. In *Technical report, ISI-University of Southern California*, 2006a.

Alexander Fraser and Daniel Marcu. Semi-supervised training for statistical word alignment. In *ACL-44:*

Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 769–776, Morristown, NJ, USA, 2006b. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Francisco Guzman, Qin Gao, and Stephan Vogel. Reassessment of phrase extraction for pbsmt. In *Proceedings of MT Summit XII*, Ottawa, Canada, August 2009.

P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. of HLT-NAACL*, pages 127–133, 2003.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL (demo session)*, 2007.

P. Liang, B. Taskar, and D. Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June 2006. Association for Computational Linguistics.

D. Marcu, W. Wang, A. Echihabi, and K. Knight. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proc. of EMNLP*, 2006.

E. Matusov, R. Zens, and H. Ney. Symmetric word alignments for statistical machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 219–227, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

R. C. Moore. A discriminative framework for bilingual word alignment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 81–88, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

J. Niehues and S. Vogel. Discriminative word alignment via alignment matrix modeling. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 18–25, Columbus, Ohio, June 2008. Association for Computational Linguistics.

F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

B. Taskar, S. Lacoste-Julien, and D. Klein. A discriminative matching approach to word alignment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 73–80, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

D. Vilar, M. Popović, and H. Ney. AER: Do we need to "improve" our alignments? In *International Workshop on Spoken Language Translation*, pages 205–212, Kyoto, Japan, November 2006.

S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. pages 239–269, 2003.

A. Zollmann and A. Venugopal. Syntax augmented machine translation via chart parsing. In *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL*, 2006.