Reassessment of the Role of Phrase Extraction in PBSMT

Francisco Guzman
Centro de Sistemas Inteligentes
Tecnológico de Monterrey
Monterrey, N.L., Mexico
guzmanhe@gmail.com

Qin Gao and **Stephan Vogel**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{qing, stephan.vogel}@cs.cmu.edu

Abstract

In this paper we study in detail the relation between word alignment and phrase extraction. First, we analyze different word alignments according to several characteristics and compare them to hand-aligned data. Secondly, we analyzed the phrase-pairs generated by these alignments. We observed that the number of unaligned words has a large impact on the characteristics of the phrase table. A manual evaluation of phrase pair quality showed that the increase in the number of unaligned words results in a lower quality. Finally, we present translation results from using the number of unaligned words as features from which we obtain up to 2BP of improvement.

1 Introduction

Statistical word alignments serve as a starting point for the Statistical Machine Translation (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 (e.g. phrase extraction), the quality of word alignments is not necessarily related to the quality of translation.

This poses the question of whether alignment quality metrics (such as AER) might not be as good predicting translation quality as other metrics. As a result, the role of the quality of word alignments in machine translation remains rather unclear. Furthermore, there are several processing steps that follow the word alignment which rarely are taken into account. Most notably the algorithm used to extract phrase pairs consistent with the word alignment (Och and Ney, 2004). The goal of better understanding the relationship between the alignment metrics (AER, precision, recall) and translation quality, is to make improvements in word alignment carry over to improvements in the end-to-end system performance. This is especially important in the case of discriminative word alignment (Niehues and Vogel, 2008), where optimization towards a given manual alignment is used.

In this paper we study in more detail the dependencies between the word alignment and the phrase extraction, as an effort to better understand the role of word alignments in phrase extraction. We explore characteristics of the alignment such as link density and number of unaligned words, and their implications on the phrase-pairs extracted from them. We also make a first attempt to include the findings of our analysis as new features of the translation model. The remainder of this paper is organized as follows: in Section 2 we give an overview of the related work. In Section 3 we evaluate different word alignments according to different characteristics. In Section 4 we analyze the impact of such characteristics (unaligned words in particular) in the extraction of phrase-pairs. In Section 5 a human evaluation is performed to assess the quality of extracted phrase pairs. Finally in Section 6 translation experiments are performed and their results discussed.

2 Related Work

There have been several attempts to clarify the role of word alignment quality in machine translation.

For instance in (Fraser and Marcu, 2006), the correlation between BLEU and AER is evaluated. They identify several flaws of AER when Possible links are used. Furthermore, they propose a variation of the F-measure which uses the coefficient α to modify the balance between precision and recall. Then they find the optimal index that better correlates to BLEU depending on the corpus. On the other hand, shows that (Vilar et al., 2006) better BLEU scores can be obtained by degrading the quality of alignments with their "alignment adaptation". However they argue that the mismatch is due to the inconsistency between the alignment and translation models, and support the use of AER as an alignment quality metric.

Ayan and Dorr (2006) go one step further and analyze the quality of the alignments as well as the resulting phrase tables. Their analysis compares several alignments' quality, and the translation from the resulting phrase tables using different lexical weightings. They also do an extensive analysis on the length of phrases used by the decoder and the phrase-table coverage. However, they do not analyze other characteristics of the alignments that impact the quality of the phrase-table such as the number of unaligned words. In their study (Ayan and Dorr, 2006) also propose the Consistent Phrase Error Rate Metric (CPER) which is similar to AER but operates at the phrase level. CPER compares the phrase table extracted from an alignment to the one generated by a hand alignment. However, the underlying assumption, that the extracted phrases from the hand aligned data using the current phrase extraction algorithms is perfect, is yet to be challenged. In this paper, we make an effort to detangle the intricate relationships between the word alignment and the phrase extraction. In the following section we analyze different characteristics of the alignment that have an impact on the phrase-table generation.

3 From Sentences to Alignments

For the following analysis, we used a small set of hand-aligned data to compare the characteristics of different alignments. We used two types of alignments: generative and discriminative, to align a Chinese(f) English (e) corpus. For the generative alignment, we used 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 used the modified GIZA toolkit (Gao and Vogel, 2008). In addition, we generated the symmetrized alignment, using the grow-diag-final heuristic implemented and used in the MOSES package (Koehn et al., 2007). For the discriminative alignments, we used the approach described in (Niehues and Vogel, 2008), because the output alignment matrix generated by such a system is composed of continuous values representing the alignment strength between source and target word. Therefore it allows to easily control the density of the alignment matrix, by using different intensity thresholds, without having to recalculate the alignment. The different thresholds used throughout this paper are $p = \{0.1, 0.2, ..., 0.9\}$.

In the following experiments, the discriminative word aligner (DWA) uses the models from the GIZA training (lexicon, fertility) as well as the GIZA S2T and T2S Viterbi alignments as features. It is tuned to minimize AER on the hand-aligned data using the alignment with threshold $p = 0.5$ as output. In Table 1 we show the sizes of the training sets for each of the aligners. We also show the size of our testing set. Notice that for the tuning and the evaluation test sets the number of English words is about 20% higher than the number of Chinese words. For the training data the ratio is closer to $1 : 1.13$.

	Corpus Statistics	
	#Sentences	#Words
GIZA Training		
Chinese	11.0 M	273M
English	11.0 M	309M
DWA Tuning		
Chinese	500	10,285
English	500	12,632
Alignment Test		
Chinese	2,000	39,052
English	2,000	48,655

Table 1: Data Statistics for the Data set used in Word Alignment

3.1 Comparison of Outputs

When describing an alignment, there are two types of measurements we can use. On one hand, there are

the alignment quality measures like AER, precision and recall, which describe how close our output is to a gold standard in terms of the number of common links in the alignment. On the other hand, we have different statistics that can be computed over alignments, i.e. number of aligned words, number of links, etc., which allow us to better understand the inner structure of an aligner's output.

In Table 2 we display the quality measurements for the different word alignment approaches. First, the one-sided GIZA alignments and the symmetrized (grow-diag-final) alignment are listed. For the discriminative word alignments the results for different thresholds are shown. Notice that the lowest AER is achieved using the DWA-0.5. Changing the threshold allows us to cover a wide variety of alignments, from high precision to high recall. We also observe that the discriminative alignment gives a lower AER than the symmetrized alignment from the generative models.

	Quality Measurements		
Aligner	Precision	Recall	AER
GIZA S2T	51.67	34.91	58.33
GIZA T2S	66.48	56.92	38.67
Symmetrized	67.98	56.29	38.41
DWA-0.1	45.16	**71.96**	44.51
DWA-0.2	57.35	65.26	38.95
DWA-0.3	64.59	61.64	36.92
DWA-0.4	69.47	59.22	36.06
DWA-0.5	72.99	57.38	**35.75**
DWA-0.6	76.04	55.22	36.02
DWA-0.7	79.26	52.33	36.96
DWA-0.8	83.26	47.91	39.18
DWA-0.9	**89.19**	38.58	46.13

Table 2: Precision, Recall and AER for the different alignments

In addition to quality, other statistics related to the nature of the alignment were also computed. We also included the hand aligned data to have a better sense of which alignments are closer to the human generated data. These measures are:

- the number of links $L_a = |a_{ij}|$ in an alignment;
- the average density of an alignment $\bar{\delta}$, i.e. the average number of links by number of words in a sentence (source J or target I) in the set of all alignments A. For instance,

$$\bar{\delta}_f = \frac{1}{|A|} \sum_{a \in A} L_a / J_a$$

- the number of unaligned words, source and target u_{fa} , u_{ea} in the alignment a;
- the average unalignment rate $\bar{\omega}$, i.e. the average number of unaligned words per sentence in the set of all alignments A. For instance,

$$\bar{\omega}_f = \frac{1}{|A|} \sum_{a \in A} u_{fa} / J_a$$

Table 3 displays the source and target densities of the alignments resulting from human and different automatic alignments. For the hand aligned data we see that on average, one English word is aligned to 1.13 Chinese words, while the reverse case is almost one and a half. This discrepancy can be explained due to the difference in sentence lengths of Chinese and English test data.

	Alignment Statistics: density		
Aligner	$\sum L_a$	$\bar{\delta}_f$	$\bar{\delta}_e$
Hand Aligned	**55,322**	**1.41**	**1.13**
GIZA S2T	37,377	0.96	0.81
GIZA T2S	47,362	1.24	0.98
Symmetrized	45,805	1.25	1.01
DWA-0.1	88,164	2.26	1.82
DWA-0.2	62,951	1.66	1.33
DWA-0.3	**52,796**	**1.41**	**1.12**
DWA-0.4	47,160	1.27	1.01
DWA-0.5	43,493	1.17	0.93
DWA-0.6	40,176	1.09	0.86
DWA-0.7	36,523	1.00	0.79
DWA-0.8	31,833	0.89	0.70
DWA-0.9	23,931	0.69	0.55

Table 3: Total number of links $\sum L_a$, average number links per source word $\bar{\delta}_f$ and average number of links per target word $\bar{\delta}_e$, for different alignments

The GIZA alignments have the characteristic that in one direction each word is aligned exactly to one word in the other language (source and target change their role in different directions). Since some words, in our case 2-4%, are explicitly aligned to the NULL word, the density of links per proper word is slightly below 1. For the discriminative aligner the number

of links decreases when the threshold is increased. The threshold DWA-0.3 gives a density closest to the hand-aligned data. Nevertheless as shown in Table 2 its AER its not the best. This makes evident that human alignments are denser than our best alignments, suggesting that there are many "good links" that we are not getting; leaving plenty of room for improvement in quality. More interesting yet, is to look at

	Alignment Statistics: unalignments			
Aligner	$\sum_a u_{fa}$	$\bar{\omega_f}$	$\sum_a u_{ea}$	$\bar{\omega_e}$
Hand Aligned	**4629**	**0.11**	**3739**	**0.08**
GIZA S2T	1675	0.04	26597	0.51
GIZA T2S	9309	0.22	1293	0.02
Symmetrized	**4905**	**0.11**	9675	0.16
DWA-0.1	1241	0.03	240	0.00
DWA-0.2	3642	0.07	1180	0.02
DWA-0.3	5676	0.12	**2988**	0.04
DWA-0.4	7418	0.16	5095	**0.08**
DWA-0.5	8882	0.20	7137	0.12
DWA-0.6	10368	0.24	9531	0.16
DWA-0.7	11987	0.27	12623	0.22
DWA-0.8	14154	0.32	17002	0.31
DWA-0.9	18591	0.43	24768	0.45

Table 4: Total number of unaligned words for source u_{fa} and target u_{ea}. Also percentage of unaligned words for source ω_f and target ω_e.

the summary of unaligned words in Table 4. In general, there is a tendency to leave more Chinese words unaligned. However, the Hand Alignment tends to be more balanced in both sides (about 10%). The disparity is more acute for GIZA alignments, given their asymmetry. Therefore, while S2T leaves more than half of the English words unaligned, T2S leaves many Chinese words unaligned

Comparing the different alignments to the human generated alignment, we find that for the source side the symmetrized alignment is very similar to the gold standard on the number of words left unaligned. On the target side, the closest match is given by DWA-0.3 (totals) and DWA-0.4 (percentage). This shows that even our best quality alignments (AER-wise) are leaving too many words unaligned.

So far we have discussed several different statistics that can be used to describe alignments. They give us different perspectives on the nature of an alignment. In the following section, we will analyze the output of the phrase-extraction algorithm having the analyzed alignments as input. Our objective is to determine which of the characteristics in the alignment might have more impact on the generated phrase-pairs.

4 From Alignments to Phrases

After generating symmetrized word alignments, the usual step in the pipeline is to extract phrase pairs. In the experiments described in this section, we used phrase-extract heuristic (Och and Ney, 2004) as implemented in the Moses package (Koehn et al., 2007), with a maximum phrase length of 7.

As opposed to word alignments, there is no gold standard human generated phrase table. While some metrics as CPER (Ayan and Dorr, 2006) have been proposed, they rely heavily in the phrase-extraction algorithm to generate gold-standard phrase-pairs from a hand alignment. By doing so, the metric obscures the effect that the phrase-extraction heuristic may have on the quality of the phrase-table. In practice, measurements such as coverage, number of generated phrase pairs, size of the phrase table (i.e. unique phrase pairs), are used. In this study, we analyzed the following characteristics of the output of the extraction heuristic:

- The number of phrase-pair instances π_a generated by an alignment a.
- The percentage of singletons (S), i.e. unique phrase-pairs.
- The source l_e and target l_f phrase lengths per phrase pair.
- The number of source g_{ek} and target g_{fk} "gaps" or unaligned words inside a phrase-pair p_k.
- The average number of unaligned words $\bar{\gamma}$ of the phrases extracted from an alignment.

$$\bar{\gamma}_e = \frac{1}{|A|} \sum_{a \in A} \frac{1}{\pi_a} \sum_{k \in P_a} \frac{g_{ek}}{l_{ek}}$$

In Table 4 we summarize the statistics from the phrases according to their length, and number. The first piece of information that we observe is that as the DWA alignments gets sparser, the number of phrase-pairs increases steadily. Furthermore, the

Aligner	Phrase Statistics: number & length $\sum \pi_a$	S(%)	$\bar{l_f}$	$\bar{l_e}$
Hand Aligned	**111K**	**78.5**	**2.90**	**3.29**
GIZA S2T	112K	87.6	2.13	3.38
GIZA T2S	90K	77.3	2.74	2.60
Symmetrized	136K	84.8	2.80	3.35
DWA-0.1	19K	65.4	2.28	2.38
DWA-0.2	51K	74.4	2.63	2.70
DWA-0.3	86K	**79.0**	2.81	2.92
DWA-0.4	**126K**	82.7	**2.96**	3.13
DWA-0.5	171K	85.8	3.13	**3.33**
DWA-0.6	231K	88.4	3.28	3.56
DWA-0.7	321K	90.8	3.45	3.79
DWA-0.8	467K	93.3	3.64	4.01
DWA-0.9	829K	95.7	3.91	4.31

Table 5: Different statistics for the phrase-pairs according to their length and number. We have the total number of phrase-pair instances $\sum \pi_a$, the percentage of singletons (S) and the average source $\bar{l_{fk}}$ and target $\bar{l_{ek}}$ phrase length

percentage of singletons also increases. This a result of the behavior of the phrase extraction algorithm. As the DWA alignments become less dense, the number of phrase-pairs that are consistent with the alignment increases. This is similar with the results reported by (Ayan and Dorr, 2006), where they found that the size of the phrase table increases dramatically as the number of links in the initial alignment gets smaller. However not all the alignments exhibit the same behavior. For instance take DWA-0.7 and GIZA-S2T alignments. They have about the same number of links. Nonetheless, the number of generated phrase-pairs is almost three times larger for DWA-0.7 than for the S2T. Instead, the number of phrases generated it seems to be an interaction between the number of links and the number of unaligned words.

Another interesting piece of information is the distribution of lengths of the extracted phrase-pairs. As an alignment gets sparser the phrase extraction algorithm is able of finding longer phrases. Nonetheless, many of those phrases are achieved include a larger number of unaligned words.This is more evident in Table 4 where we summarize the gap statistics for the phrase-pairs extracted from different alignments. Observe that the number of expected gaps $\bar{g}$ in a phrase increases as the sparsity of an alignment increases. For instance, most of the phrase-pairs of most-dense alignment (DWA-0.1) are gap-less (90% for source and 98.9% for target). In contrast for the DWA-0.9, the gap-less phrase pairs for source and target side account for 16.4% and 13.6% respectively.

Aligner	Phrase Statistics: gaps $\bar{g_f}$	$g_f^0(\%)$	$\bar{g_e}$	$g_e^0(\%)$	$\bar{\gamma_f}$	$\bar{\gamma_e}$
Hand Aligned	**0.50**	**66.6**	**0.39**	**71.7**	**0.11**	**0.08**
GIZA S2T	0.15	85.6	2.01	20.2	0.04	0.44
GIZA T2S	0.79	52.2	0.07	93.2	0.19	0.02
Symmetrized	0.59	62.3	0.95	51.2	**0.11**	0.16
DWA-0.1	0.12	90.0	0.02	98.9	0.02	0.00
DWA-0.2	0.37	74.8	0.11	92.4	0.07	0.02
DWA-0.3	**0.56**	**64.0**	0.26	80.3	**0.11**	0.05
DWA-0.4	0.80	53.7	**0.51**	**66.6**	0.15	**0.08**
DWA-0.5	1.06	44.8	0.79	54.7	0.19	0.12
DWA-0.6	1.30	37.6	1.09	43.8	0.22	0.16
DWA-0.7	1.56	31.0	1.47	33.2	0.26	0.22
DWA-0.8	1.84	24.8	1.88	23.7	0.31	0.29
DWA-0.9	2.34	16.4	2.56	13.6	0.39	0.41

Table 6: Different statistics for the phrases according to their length and number. We average number of gaps found in source $\bar{g_f}$ and target $\bar{g_e}$ phrases , the percentage of phrases extracted without growing into any gap g_{fk}^0, g_{ek}^0 , the unsupported word rates $\bar{\gamma_f}$, $\bar{\gamma_e}$

Notice also that the number of gap-less phrase-pairs tends to be higher in the English side, than in the Chinese side (not taking into account GIZA alignments). This is a direct consequence of the number of unaligned words in each side as shown in Table 4. In fact, when we look at the unsupported word rate γ, which is an average of the gaps per word in a phrase-pair, we observe that is extremely close to the percentage of unaligned words ω of the alignments. In fact, the correlation between these two statistics is very high (0.93 for source and 0.94 for target) suggesting that the distribution of unaligned words in our alignment carries into the phrase-pairs even after phrase-extraction.

In summary, we observe that the number of unaligned words in an alignment have a large impact on the phrases generated. They affect the number of phrases generated, the number of unique phrases, the length of these phrases, the number of gaps inside a phrase, etc. In the next section, we show how these characteristics even affect the perceived human quality of the extracted phrases.

5 Human Evaluation of Phrase Pairs

As we have seen in the previous analysis, the number of unaligned words of an alignment has a huge impact on the number of phrase pairs extracted. As we observed, the phrase extraction heuristic allows to generate phrase pairs by growing into gaps. Some of these gappy phrase pairs will be useful. They will increase coverage, and actually might be accurate phrase pairs. However, many other phrase pairs will be partially, if not completely wrong. To investigate how the quality of the extracted phrase pairs depends on the type of the underlying word alignment, and on the gaps of the phrases extracted from these alignments, a small-scale human evaluation was conducted. Several native Chinese speakers participated in this evaluation.

The procedure was the following: Each subject was presented with a set of Chinese-English phrase pairs. For each phrase-pair they judged if source and target phrase were adequate[1] translations of each other. This was done without any other contextual information (i.e. the surrounding words, or the sentence pairs from where these phrases were extracted). This was also done blindly, as the evaluators did not have any knowledge of the origin of the phrase-pairs. Furthermore, we included a noisy set, which were pairs of randomly selected source and randomly selected target phrases. Such noisy set would help us to determine how likely is to obtain a good score by just having a random pair of source and target phrases.

The phrases included are the ones generated by the alignments from the DWA with thresholds 0.1 to 0.9, the symmetric alignment and the hand-aligned data. Each set was generated by randomly selecting unique phrases generated by the alignments. We split the phrase pairs extracted from the hand-aligned data into two groups, Gaps and No-Gaps, which stand for phrase pairs generated from the hand alignment by growing into gaps, and phrase pairs generated without growing into gaps. We used this configuration because we wanted to highlight the effect of having gaps in the phrases generated from a perfect alignment. The sample sizes were of 100 for the HA-Gaps, HA-No-Gaps, symmetric alignment and noisy sets and 50 for the DWA-0.1 to DWA-0.9 sets.

After the results were collected, an ANOVA was done, considering the independent variable: system, the random variables: evaluator, number of gaps in source phrase and target phrases; and the dependent variable: adequacy, with non-repeated measurements. As we can see in Table 5, only the system (alignment) is a significant factor, i.e. the means of the evaluation by system are not equal. This, as we expected, means that there are differences in quality across systems. The interaction between system and evaluator is not significant, which means that there is no evidence that show that evaluators were biased towards any specific system, which is expected in a blind experiment. Also note that while the effect of the number of source gaps is almost significative (at $\alpha = 0.01$, there is strong evidence that suggests that there is an interaction between source and target gaps. In other words, looking at the gaps in one side of a phrase-pair may not tell us much about its quality. However, the combination of source and target gaps might be a good indicator.

Source	SS	df	MS	F	p-val
EV	0.04	2	0.02	0.15	0.8556
SYS	13.92	8	1.74	11.48	**0.0000**
SG	1.76	2	0.88	4.42	0.0138
TG	0.74	2	0.37	1.77	0.1745
EV*SYS	6.13	40	0.15	1.01	0.4518
SG*TG	9.24	29	0.31	2.10	**0.0007**
Error	113.99	752	0.15		
Total	196.66	849			

Table 7: ANOVA table showing the effects in the experiment: Evaluator (EV), System (SYS), Number of Source Gaps (SG), Number of target Gaps (TG). Also two-way interactions are shown for Evaluator*System, Source Gaps*Target Gaps.

In Figure 1 we show the mean of the evaluation by system[2]. As expected, random phrase-pairs perform poorly. This verifies the consistency of the judges evaluation as good scores could not have been achieved randomly. Surprisingly, the phrase pairs extracted from DWA-0.1 achieved the highest

[1] By adequate, we mean that a source phrase could be used as translation of a target phrase in at least one situation, without loss of meaning

[2] The confidence intervals are merely informational. To determine statistical differences, one must perform unplanned pairwise comparisons such as Scheffé tests

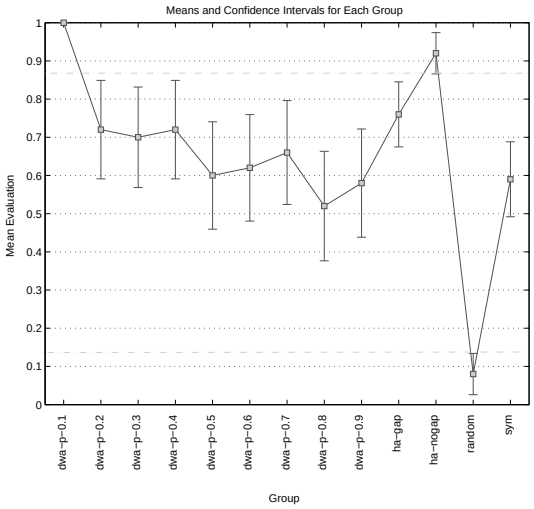

Figure 1: Mean of evaluation by systems, plotted with a confidence interval of 95%

score. While DWA-0.1. phrase are nearly all gappy-less, the underlying word alignment is far from perfect. Furthermore, comparing phrase pairs extracted from human word alignment, shows that a perfect word alignment does not lead to perfect phrase-pairs given the current extraction heuristic. However, the HA-no-gaps set performs better than the HA-gaps set. This suggests that the quality of a phrase-pair extracted from a gold standard hand alignment deteriorates when it has gaps on either its source or target phrases. Overall, we do observe the tendency that less number of unaligned words in the word alignment leads to better quality of the extracted phrase pairs. In other words low precision/ high recall alignment results in fewer but higher quality phrase pairs. To balance the trade-off between higher quality phrases and coverage, we conducted a series of translation experiments were the number of unaligned words was taken as a feature. In next section, we describe them thoroughly.

6 From Phrases to Translations

After the phrases are extracted, they are scored according to the MLE estimation described in (Koehn et al., 2003). Also, the reordering models and the lexical weighting are estimated. Then, these models (along with the language model) are used during decoding. For this study, we wanted to analyze the impact of the quality and other characteristics of the phrase-pairs that could affect the translation result. In particular, we want to pay special attention to the number of unaligned words for the phrase-pairs in our phrase table. Therefore, we performed translation experiments with the different alignments previously analyzed. In addition, we introduced two new features to the phrase table, account for the number of unaligned words.

6.1 Setup

For this experiment, we used a training data set consisting of the GALE P3 Data[3]. The data was filtered to have maximum sentence length 30. The final training set contains one million sentences. The different systems that were used, were built upon the alignments from the DWA with $p = \{0.1..0.0\}$, and the symmetrized alignment (grow-diag-final). The DWA Tuning remained the same as for Sec 3. We use the MT05 test set for tuning, and used a subset of the development dataset of GALE07 Evaluation (DEV07) as the blind testing data. The sub-sampling is done by third-party. The data set consists subsets from different sources: Newswire (NW) and Weblog (Web) with 427 and 358 sentences respectively. In Table 8 we display the BLEU Scores for these sets. First, notice how in our baseline the best results are

	TUNE		**NW**		**WEB**	
Alignment	base	unal	base	unal	base	unal
DWA-0.1	24.73	24.1	21.20	22.25	18.70	18.76
DWA-0.2	26.42	24.9	22.97	23.3	20.06	20.11
DWA-0.3	26.90	25.93	23.11	23.35	20.18	20.3
DWA-0.4	27.41	26.15	24.19	24.81	20.50	21.81
DWA-0.5	27.65	26.29	**24.56**	24.72	**20.78**	21.57
DWA-0.6	27.52	26.45	24.05	**24.97**	20.53	**22.57**
DWA-0.7	27.24	26.55	23.05	*24.62*	19.54	*21.86*
DWA-0.8	27.49	26.38	23.83	24.74	20.52	21.88
DWA-0.9	27.82	26.68	23.32	24.26	20.26	21.62
SYM	27.32	26.12	23.15	24.18	20.22	21.11

Table 8: Translation results for the baseline systems (base) and unaligned features enhanced systems (unal) built upon different alignments.

[3]FOUO data (LDC2006G05), HKnews (LDC2004T08), XinhuaNews (LDC2003T05), and parallel data from GALE (LDC2008E40, LDC2007E101, LDC2007E86, LDC2007E45, LDC2006E92, LDC2006E34, LDC2006E26, and LDC2005E83).

obtained by the system that previously achieved the highest AER (DWA-0.5). From there on, the systems trend to have lower quality as we shift the balance from precision/recall in our alignment. However, the alignments with higher recall (DWA-0.1) trend to perform more poorly than the high precision ones (DWA-0.9). This is not surprising, as this phenomenon has been observed previously. For the systems that use the number of unaligned words as a feature, we observe that the best results are found with a higher precision alignment (DWA-0.6). This can be explained as the result of penalizing the phrases that include a lot of gaps, which as shown before have lower human-perceived quality. The improvements of using unalignment features are more striking for the Web test set, where we obtain up to 2BP of improvement (for DWA-0.7). Finally, notice how the tuning results are significantly lower for the systems that use unalignment features, yet they achieve better results on unseen data. This suggests that using unalignment features might be preventing from over-fitting the tuning set.

7 Conclusions

In this paper we studied in detail the relation between word alignment and phrase extraction. First, we analyzed word alignment according to several characteristics and compared them to hand-aligned data. We observed that there is a lot of room of improvement for our alignment models. Second, we analyzed the phrase-pairs generated by these alignments. We observed that sparser word alignments lead to a larger phrase tables. While these larger phrase tables contain longer phrases, many of the phrases contain unaligned words. Also, the number of unaligned words in the alignment has a large impact on the characteristics of the extracted phrase table. The unaligned words in the extracted phrase pairs follow the distribution of unaligned words in the alignment from where they were extracted. Third, a manual evaluation of phrase pair quality showed that the more unaligned words (gaps) result in a lower human perceived quality. Finally, when we include the number of unaligned words as a feature in our phrase-table we are able of better discriminate good phrase pairs from bad phrase pairs. By doing so, we obtained up to 2BP of improvements.

Acknowledgments

This work is in part supported by the US DARPA GALE programs[4].

References

Necip F. Ayan and Bonnie J. Dorr. 2006. Going beyond aer: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney, Australia, July. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2006. Measuring word alignment quality for statistical machine translation. In *Technical report, ISI-University of Southern California*.

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

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL'07*, pages 177–180, Prague, Czech Republic, June.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

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

[4] Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.