All Links are not the Same: Evaluating Word Alignments for Statistical Machine Translation

Paul C. Davis*, Zhuli Xie*, and Kevin Small†‡

*Human Interaction Research Lab
Motorola
1295 E. Algonquin Road
Schaumburg, Illinois 60196, USA
{pdavis, zhuli.xie}@motorola.com

†Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801, USA
ksmall@uiuc.edu

Abstract

Word alignments, the mappings between source and target language words for two languages, are a critical component of statistical machine translation. A long-standing issue in statistical machine translation is that the quality of word alignments does not correlate as well as would be expected with measures of translation quality. A number of recent papers have shed light on this issue by improving on existing metrics such as Alignment Error Rate and examining the importance of word alignment quality in terms of phrase alignments. In this paper, we attempt to elucidate this situation further by first presenting a new word alignment evaluation metric, Word Alignment Agreement F1 ($\text{WAA}_{\mathbf{F1}}$), which improves upon existing alignment quality metrics. We then present experiments which demonstrate that $\text{WAA}_{\mathbf{F1}}$ also correlates better with measures of translation quality than do previous metrics.

Introduction

Statistical approaches to building machine translation systems typically includes the important step of aligning the words of translated sentence pairs, called *bitexts*. This alignment between source and target language words is called *word alignment*. A long-standing conundrum in statistical machine translation is that automatic measures of word alignment quality, such as Alignment Error Rate (AER) (Och and Ney, 2003), do not correlate as well as might be expected with automatic measures of translation quality, such as BLEU (Papineni et al., 2001). A number of recent papers have shed light on this issue by improving existing word alignment metrics like AER and examining the importance of word alignment quality in terms of phrase alignments (Fraser and Marcu, 2006; Lopez and Resnik, 2006).

In this paper, we attempt to elucidate this situation further, by extending an alternative word alignment evaluation metric called Word Alignment Agreement, first discussed in (Davis, 2002), which is a more appropriate measure for the word alignment task because it conserves the mass of each word. Word Alignment Agreement (WAA) is a symmetric measurement which treats words as the core unit when measuring alignment quality, rather than the links between words, so that the alignment quality for each word has an equal impact. We demonstrate that while WAA is limited to certain ideal alignment configurations, it can be straightforwardly extended to account for all types of word and phrase alignments. We next show through examples why the new metric, $\text{WAA}_{\mathbf{F1}}$, improves upon existing metrics. The remaining sections of the paper demonstrate empirically how this new metric yields good correlation with translation quality for individual types of word alignments, and, more importantly, how it yields significantly better correlation across different types of alignments than possible with previous metrics.

Word Alignment Evaluation Background and Related Work

While AER has been the most widely used word alignment evaluation (WAE) metric, Fraser and Marcu (2006) demonstrate that it is less than ideal because it does not appropriately penalize unbalanced precision and recall. In its place, they propose to use F-measure, as shown below, where the parameter α can be used to tune the relative importance of precision and recall. F-measure is said to be balanced when α equals 0.5.

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

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

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

‡Kevin Small participated in this research while working as a summer intern in Motorola's Human Interaction Research Lab.

$$(4)\ \text{F-measure}(A,P,S,\alpha) = \frac{1}{\frac{\alpha}{Prec(A,P)} + \frac{1-\alpha}{Recall(A,S)}}$$

Alignments, in particular those in *gold* (reference) sets, are often annotated with confidence levels, S (Sure) and P (Possible). We thus use S and P to indicate sets of Sure and Possible reference alignments, respectively, and A to indicate the set of predicted alignments. We will refer to F-measure often as $F1$, and in instances where Sure and Possible are used, as in (4), we will refer to it as $\text{F1}_{\mathbf{SP}}$. We use the term $\text{F1}_{\mathbf{S}}$ to indicate that we are using F-measure when there are only Sure alignments, or when we are only counting the Sure alignments, and thus precision becomes $\frac{|S \cap A|}{|A|}$.

Fraser and Marcu use F1, and in particular $\text{F1}_{\mathbf{S}}$, and demonstrate that by tuning α in different alignment environments, they can achieve an improved correlation with translation quality. They perform a series of illuminating experiments where they vary alignment and translation quality and measure correlation of F1 with various α values and BLEU across different types of alignments. One drawback to this tuning approach, however, is that it amounts to calculating a different metric (a different α) for each environment.

In a related study, Lopez and Resnik (2006) demonstrate that the quality of word alignment (measured with AER) and translation are correlated, but not very strongly, because the impact of word alignment quality is lessened when the word alignments are used to build phrase alignments, which are demonstrated to be more critical to translation quality. They do leave open the possibility in their experiment that the relationship between the quality of word alignment and translation is obscured by the use of a poor metric in either case. Similarly, Vilar et al. (2006) demonstrate that certain types of alignments when degraded according to metrics like AER and $\text{F1}_{\mathbf{SP}}$, still produce improved translation quality, when used with certain types of translation models.[1] Finally, Ayan and Dorr (2006) introduce a metric more oriented to phrases, consistent phrase error rate (CPER), which they find to be more informative than AER, but they are not able to find a direct correlation between their metric and BLEU.

A New Metric: Word Alignment Agreement F1

The most widely used word alignment metric, AER, counts the number of links, as does F1, as described above. There is, however, an alternative type of WAE metric that does not count links, but rather counts the number of words aligned. To our knowledge, only two metrics do this, WAA (Davis, 2002) and the metric in Melamed (1998). Metrics that count the number words improve on metrics which count the number of links because the latter have the tendency to overstate/understate the importance of words with a relatively large/small number of links. Since the number of words in a bitext is a constant, these approaches provide a better base for WAE metrics.[2]

Word Alignment Agreement Background

To describe and extend WAA, we must first define some terminology. Davis (2002) focused on *conservative* alignments, those defined as *complete* and *fully connected*.[3] *Completeness* simply means that every word is aligned with something, null or otherwise. *Connectedness* is the transitive closure of the link relation on word ids, thus a source word id can be *connected* to any other source or target id. For example, if source id s_1 is linked with target ids t_1 and t_2, and s_2 with t_2 and t_3, then s_1, s_2, t_1, t_2, and t_3 are connected.[4] We typically write links as (i:j), with source ids on the left and target ids on the right, and use 0 to represent null in null alignments:

(5) {(1:1)(1:2)(2:2)(2:3)}

We refer to such a maximal set of connected links as a link *group*. Finally, a group is *fully-connected* when every source id is linked to every target id (i.e., it forms a bipartite clique), thus to make the group in (5) fully-connected we would need to add the links (1:3) and (2:1), to yield:

(6) {(1:1)(1:2)(1:3)(2:1)(2:2)(2:3)}

Thus an alignment, the set of links in an aligned bitext, is fully-connected when all of its groups are fully-connected.

Given this terminology, the key then to WAA is a weighting of the links. Following the intuition of one-to-one alignments, each word in alignment can contribute exactly .5 units of weight, thus a one-to-one link intuitively would account for a weight of 1, if the words were only involved in that link, and a one-to-null link would account for .5 units of weight, since only one word is involved. A goal of WAA is to have every conservative alignment of a given bitext have the same total weight. This means there can be no under or over-counting of the correctness of a given bitext, and that a bitext will only contribute its appropriate share to a larger evaluation.

Suppose we have the following word alignment, for four word source and target sentences:

[1] Testing these models with $\text{WAA}_{\mathbf{F1}}$ is beyond the scope of this paper, as we use as our translation model what may be viewed as the current standard statistical model, but it seems clear that the translation models most likely to improve with improved alignment are those that use such alignments most directly.

[2] WAA and the metric from Melamed (1998) are quite similar in this respect; we extend WAA because it is always non-deficient and covers null alignments, but we could have used Melamed's metric as well.

[3] Conservative alignments are also not *null-deficient*, meaning a word aligned with null cannot be aligned with anything else.

[4] Connectedness is a property of ids, but we sometimes refer to links as being connected when their ids are.

(7) {(1:1) (2:2) (3:3) (4:4)}

There are 4 links, and 8 total words. The entire alignment has a weight of 4, equal to the number of source words (m) plus the number of target words (n) divided by 2.

When weighting link groups, the weight, w, for each link, l, is:

$$(8)\ l_w = \frac{total\ weight}{number\ of\ links} = \frac{\frac{m+n}{2}}{m \times n} = \frac{m+n}{2mn}$$

weight for word-with-null link$= 0.5$

Computing the links that exist for *many* alignments (those where the alignment does not commit to how ids are individually linked, as in (9)) is done by taking the cartesian product of the alignments to create fully-connected groups, and then calculating weights. This process is termed *flattening*.[5] In the 2:3 alignment case in (9), six links are produced in (10), each with $\frac{m+n}{2mn} = \frac{5}{12}$ weight.

(9) (1,2:2,3,4)

(10) {(1:2)(1:3)(1:4)(2:2)(2:3)(2:4)}

Comparing any two conservative word alignments of the same bitext is straightforward. For each alignment, flatten as needed, so that all the remaining links are either null alignments or 1:1, and count how much of the weight agrees. WAA is defined below, where *w* indicates weight, and one of the alignments is arbitrarily selected as a gold set to yield S, and the other as a predicted set to get A (as in definition of Recall (2)):

$$(11)\ WAA = \frac{A_w \cap S_w}{S_w}$$

The total agreeing weight is divided by the total weight. For an entire corpus, the sum of the agreeing weights for each bitext divided by the total weight of the corpus gives a global WAA score for two sets of word-aligned bitexts.

Improving on WAA

One shortcoming of WAA is that the formula for assigning weights to non-null grouped links in (8) assumes that the group is fully-connected. While it is true that human annotators often favor fully-connected alignments (they form the majority of the gold alignments used in the experiments) and that many automatic aligners, such as GIZA++ (Och and Ney, 2003) (when operated in a single alignment direction), perform fully-connected alignments, any alignment quality metric must also account for groups of alignments which are not fully-connected. Such alignments do occur from human annotators, arise as the result of heuristics to generate phrasal alignments, and certainly seem to be the correct alignments in some cases.

It turns out that it is quite easy to extend WAA's weighting scheme (given in (8)) to cover non-fully-connected link groups, as follows in (12). We refer to this as the WAA' weighting scheme, where W is the total number of words in the group of n source words and m target words, N is the number of null links, and F is the number of one-to-one word links:

$$(12)\ L = \frac{W}{N + 2F}$$

weight for word-to-word link$= L$

weight for word-with-null link$= \frac{L}{2}$

As with WAA, WAA′ distributes all the words' weight evenly among the links, and null alignments get half the weight of one-to-one word links from the same link group.[6] WAA′ simply generalizes WAA so that any complete alignment can be handled, and has the nice property of assigning the same score that WAA would for fully-connected alignments, and still remains a symmetric measure of alignment quality.

Calculating WAA′ is similar to calculating WAA. Alignments first need to be grouped (i.e., these are the groups of connected links defined earlier), then for each group, the number of regular links, F, and number of null links, N, are counted so that the weights can be calculated and assigned. The reweighted sets of links can then be compared to compute the matching weight divided by the total weight of the alignment, just as in (11).

WAA′ is still not a full solution, however. Like WAA, WAA′ only is appropriate for complete alignments. Complete alignments should arguably be required for any word alignment task. If a human or automatic aligner has not aligned every word of a bitext with another word or with null, one could argue that they have not finished the task. Nevertheless, word alignment metrics are most useful if they can also account for phrasal alignments, which whether done by people, heuristics, or other automatic methods, are quite often incomplete. WAA′ as defined amounts to a recall measure, but handles complete alignments because with such alignments and using the words' weight in the denominator in (11), recall and precision have the same value (since A_w=S_w, and ignoring questions of Sure and Possible confidences).

For incomplete alignments, WAA′ will tend to favor high recall metrics over high precision metrics. This leads us to the obvious and final extension, $\text{WAA}_{\mathbf{F1}}$.

[5] Note that we show *many* relationships by separating word ids with commas. This also shows how WAA handles alignments between larger sequences of words . They are treated as the set of relevant one-to-one links.

[6] One could imagine many other weighting schemes for non-fully-connected alignments, since the 'density' (number of links they participate in in the group) of the links may vary. Here we have chosen to distribute the weight equally. Note that this scheme also allows us to handle null-deficient alignments.

(13) $\text{WAA}_{Precision}(A,P) = \frac{P_w \cap A_w}{A_w}$

(14) $\text{WAA}_{Recall}(A,S) = \frac{S_w \cap A_w}{S_w}$

(15) $\text{WAA}_{F1}(\text{A,P,S},\alpha) = \frac{1}{\frac{\alpha}{WAA_{Prec}(A,P)} + \frac{1-\alpha}{WAA_{Recall}(A,S)}}$

$\text{WAA}_{\mathbf{F1}}$ balances (with α=0.5) precision and recall, so that certain types of alignments are not favored, and like F1, allows tuning of the metric via the α parameter. $\text{WAA}_{\mathbf{F1}}$ uses the same weighting scheme as defined for WAA′ in (12). As in our discussion with F1, we refer to $\text{WAA}_{\mathbf{F1}}$ as $\text{WAA}_{\mathbf{F1SP}}$ when S and P are used, and as $\text{WAA}_{\mathbf{F1S}}$ to indicate that there are only Sure alignments, or to make clear that we are only using the Sure alignments given, and thus $\text{WAA}_{Precision}$ becomes: $\frac{S_w \cap A_w}{A_w}$. We use $\text{WAA}_{\mathbf{F1}}$ when no such distinction is necessary.[7] $\text{WAA}_{\mathbf{F1}}$ has the property that it gives the same scores that WAA does for conservative alignments and the same scores for any complete alignment that WAA' would, thus it replaces these metrics in full.

$\text{WAA}_{\mathbf{F1}}$ does sacrifice one important property for an alignment metric, symmetry. Incomplete alignments are inherently unsymmetrical, because for precision and recall based measures, like AER, F1, and $\text{WAA}_{\mathbf{F1}}$, the counts or weights in the denominators of the measures will vary, depending on which corpus of aligned bitexts is taken to be the reference corpus. $\text{WAA}_{\mathbf{F1}}$ does, fortunately, give scores which scale well from group to bitext to aligned corpus, unlike F1 and AER, if the reference bitexts tend to be complete, as we will next show.

Comparing $\text{WAA}_{\mathbf{F1}}$ with AER and F1

As stated earlier, alignment metrics that count links have the tendency to disproportionately increase the importance of words that participate in many links (Melamed, 1998). This is true for both AER and F1. This tendency makes them worse measures of alignment quality, and the problem can be magnified as more and more bitexts are considered when comparing corpora of word-aligned bitexts.

A simple example serves to show this deficiency, one which $\text{WAA}_{\mathbf{F1}}$ does not share. Consider the following small correct and predicted alignments, where each corpus contains the same bitexts as the other, as is the normal case, and in addition the two bitexts happen to be identical three source word and three target word sentences.

(16)
Predicted 1:	$\{(1:2)(1:3)(2:1)(3:2)\}$
Correct 1:	$\{(1:1)(2:2)(3:3)\}$
Predicted 2:	$\{(1:1)(2:2)(3:3)\}$
Correct 2:	$\{(1:1)(2:2)(3:3)\}$

[7] All Sure links are also Possible, thus all three versions of $\text{WAA}_{\mathbf{F1}}$ mean the same thing when there are only Sure alignments.

In the example, the correct alignments are all one-to-one alignments of the first source word with the first target word, the second source with the second target, and the third with the third. Predicted 2 predicts perfectly. Predicted 1, on the other hand, does not get a single link correct. The alignments of all four bitexts are complete.

Given this situation, where we have an equal number of words in each reference and predicted bitext, the same bitext repeated, and where one prediction is completely right and another is completely wrong, it is reasonable that an alignment quality metric would yield a score of 0.50, i.e., half-correct.

The scores we get, however, for 1-AER (as the measure is typically used) and F1 diverge from this ideal:

(17) $\text{1-AER} = 1 - (1 - \frac{3+3}{6+7}) \approx .46$

(18) $\text{F1} = 1/(\frac{.5}{3/7} + \frac{.5}{3/6}) \approx .46$

(19) $\text{WAA}_{\mathbf{F1}} = 1/(\frac{.5}{3/6} + \frac{.5}{3/6}) = .50$

The problem arises simply from the fact that for Predicted 1, word 1 links with more words than do the other word ids in all the bitexts, thus overstating its importance, and thereby skewing the amount counted as incorrect. We could have easily made this example more extreme by aligning each word id in Predicted 1 with null as well, giving an F1 and 1-AER of approximately 0.32, which greatly overstates the importance of the first bitext.

One can imagine that given many alignments, each with many links, this sort of error could be magnified, and worse yet, although the scores would be incorrect, their total number of links for the bitexts might balance out, to give no indication that something is amiss. In addition, F1 and AER will similarly miscount null alignments when comparing alignments (in fact, we will see this when we look at the cross-aligner translation quality correlation for these metrics).

Thus, it should be clear even from this small example that $\text{WAA}_{\mathbf{F1}}$ is a better word alignment metric, when considering measuring alignment quality alone, than either AER or F1, because it correctly counts the importance of each word in an alignment. As such, $\text{WAA}_{\mathbf{F1}}$ scales well from links, to link groups, to phrases, to bitexts, to full word-aligned corpora, always basing its measurement on the number of words. AER and F1 are deficient in this regard.

Experiments: Correlation with Translation Accuracy

An additional important question that one may ask of a word alignment metric such as $\text{WAA}_{\mathbf{F1}}$ is how well does it correlate with translation quality. Put another way, given an improvement in alignment quality as measured by the metric, can we expect a commensurate improvement in translation quality? And further, if there is such a correlation,

does it hold across different alignment types? In this section we describe experiments we completed to answer these questions.

Experimental Methodology

To measure correlations of alignment and translation quality, it is necessary to have alignments of varying quality. We followed closely the methodologies of Lopez and Resnik (2006) and Fraser and Marcu (2006). The general idea is to take a training set of bitexts and divide the set into many pieces. These smaller pieces can then, using an automatic aligner like GIZA++, be aligned separately and then recombined. The automatic aligner will create better alignments the more data it is exposed to. Thus, each time the set is split, the alignment quality is expected to degrade. Repeatedly creating these small sets then recombining yields sets of varying quality, while keeping the overall data the same.

We ran experiments with two sets of data, French/English and Romanian/English. The training and test data for French/English are from the Canadian Hansards, as provided in the 2003 HLT/NAACL Workshop (Mihalcea and Pedersen, 2003),[8] where the test data was provided by Och and Ney (2000). We did some preliminary filtering of the training and test data using scripts provided with Moses, the open-source machine translation system replacement for Pharaoh (Koehn, 2004). The word alignment test data contained 447 bitexts with S and P confidence markings, and the training and development data consisted of just over 1 million bitexts. We randomized the this data, and reserved several thousand bitexts for minimum error rate training (MERT) (Och, 2003), and for translation testing (in our French/English tests, we used 1,000 sentences for MERT and 2,000 for translation testing).

We then created 10 data sets as follows. Taking 150,000 sentences as our base set, we created 7 additional sets by dividing the 150,000 into smaller parts, first dividing it in by 2, then by 4, by 8, and so on. We also created two larger sets using additional bitexts, which doubled and quadrupled the size of the base set. With each of these ten sets, we used GIZA++ to align each subpart in both directions, from French to English and English to French. After the subparts were aligned, they were put back together, so that each final set consisted of the same 150,00 original bitexts, plus the two larger sets which included this same data, thus creating sets of varying quality from one original data set.

Each of these now bidirectionally aligned data sets were processed to create phrase alignments using three symmetrization heuristics, intersection and union (Och and Ney, 2003) and grow-diag-final (Koehn et al., 2003).[9]

These resulting phrase alignments were then used with Moses, a representative of the current state of the art in statistical machine translation,[10] making 30 different machine translation systems (10 sets of data, each used with 3 symmetrization heuristics) trained with the phrases derived from the same bitexts, but of varying quality, which were used to evaluate the translation test sentences with BLEU. The word alignment test sentences were also evaluated for each of the data sets, and for data sets with subparts, we evaluated over each subpart and averaged the results. With this methodology, we had 10 data points to compare for each of the three heuristics, with values for alignment and translation accuracy.

We also wanted to test a different language pair, and one with only Sure confidences, so we selected the Romanian/English task also from Mihalcea and Pedersen (2003). This is a much smaller corpus, so we used 6 data sets, all divided into subparts as before, from an original 45,000 bitexts. We used 190 test sentences for word alignment evaluation, and 500 sentences for MERT training and 500 sentences for translation evaluation.

Experimental Results

To see how well translation and alignment results correlate, following Fraser and Marcu, we used r^2, the square of the Pearson product-moment correlation coefficient, where a positive correlation is a number between 0 (no correlation) and 1 (perfect correlation).

A first question we wondered about was how well $\text{WAA}_{\mathbf{F1}}$ would correlate with translation quality, for a single alignment method. Correlation scores for all the metrics were quite good, ranging from near .80 to just over .95 for the heuristics in different configurations. We did not find large differences in the various metrics.

A more interesting question we sought to answer is how well do alignment and translation quality correlate across different types of alignments. If correlation is good, it means that cross-alignment scores are more comparable in terms of effect on translation.

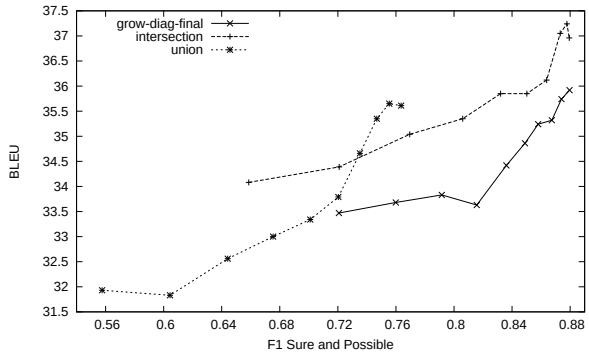
We see the results for French/English in Figure 1. A perfect correlation would look like a line from the lower left corner of the graph to the upper right. We see that $\text{WAA}_{\mathbf{F1}}$, with a score of 0.88, correlates better with BLEU than does a balanced F1, with an r^2 of .71. It also has better correlation than AER (see Table 1).

In Fraser and Marcu (2006), they demonstrate that correlation can be improved by testing and setting α. As men-

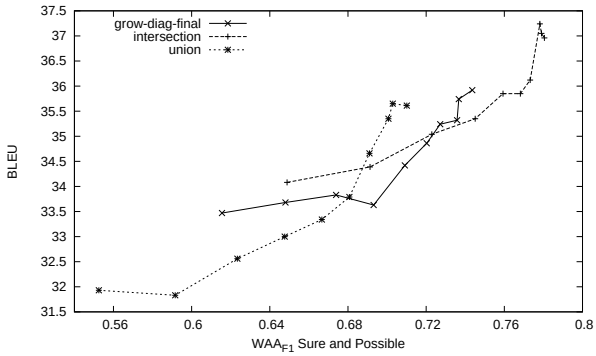
[8] Thank you to Ulrich Germann for creating the sentence aligned version of this data.

[9] This follows Fraser and Marcu (2006), where the refined heuristic was used rather than grow-diag-final, more closely than Lopez and Resnik (2006) where only one heuristic, grow-diag-final was used. The version of grow-diag-final we used, from Moses, did not produce null alignments, while our version of intersection and union do.

[10] See http://www.statmt.org/moses; thank you very much to the developers for making this code available, it was invaluable for this research.



(a) $\mathrm{F1}_{\mathbf{SP}}$ versus BLEU, $r^2 = 0.71$



(b) $\mathrm{WAA}_{\mathbf{F1SP}}$ versus BLEU, $r^2 = 0.88$

Figure 1: Correlation of Word Alignment Metrics versus BLEU, for French/English, showing stronger correlation for $\mathrm{WAA}_{\mathbf{F1SP}}$ than $\mathrm{F1}_{\mathbf{SP}}$

French/English Alignment Setting	1-AER	$\mathrm{F1}_{SP}$	$\mathrm{F1}_{S}$	$\mathrm{F1}_{P}$	WAA_{F1SP}	WAA_{F1S}	WAA_{F1P}	WAA_{F1Pw}
heuristics unchanged	0.72	0.71	0.64	0.01	0.88	0.71	0.46	0.84
heuristics no nulls	0.87	0.81	0.56	0.00	0.81	0.57	0.36	0.81
heuristics completed	0.08	0.06	0.31	0.00	0.19	0.45	0.10	0.13

Table 1: Summary of correlation (r^2) results for word alignment metrics and BLEU scores, for French/English

tioned earlier, a drawback of this approach is that it requires that α be calculated for each task. By providing a metric with better inherent correlation with translation, it can help to make alignment comparison more informative.[11]

We ran a number of other experiments with this same data (i.e., using the same translation results) and manipulated the predicted alignments before scoring them with the various metrics. These are shown in Table 1. We wondered what effect removing null alignments would have on the metrics.[12] Interestingly, this improved the other metrics' correlation scores (in particular AER) and lowered $\mathrm{WAA}_{\mathbf{F1}}$'s. This seems to support the notion that one of $\mathrm{WAA}_{\mathbf{F1}}$'s strengths is in its better accounting of null alignments. We also tried making the alignments complete, by adding null alignments for any unaligned words. This added noise to the alignments degrades the correlation as would be expected, but here $\mathrm{WAA}_{\mathbf{F1}}$ shows better but very low correlation, at 0.19, as compared to AER and F1.[13] Since this was an S and P gold alignment, we also investigated counting only the S alignments (see $\mathrm{F1}_{S}$ and $\mathrm{WAA}_{\mathbf{F1S}}$, as well as using P alignments in place of S in recall measures (see $\mathrm{F1}_{P}$ and $\mathrm{WAA}_{\mathbf{F1P}}$), and finally experimented with decreasing the weight applied to P links with $\mathrm{WAA}_{\mathbf{F1Pw}}$.[14]

We next take a look at the Romanian/English experiment, which, as mentioned earlier, had much less data (and thus lower translation and alignment quality scores), and used only Sure confidence on the gold alignments. The primary results for the Romanian/English task are shown in Figure 2. Again, in this experiment, $\mathrm{WAA}_{\mathbf{F1}}$ has a better correlation score, 0.87, than do AER and balanced F1, 0.73. These results lend additional support to $\mathrm{WAA}_{\mathbf{F1}}$ as a useful word alignment metric in terms of it relationship with translation quality.

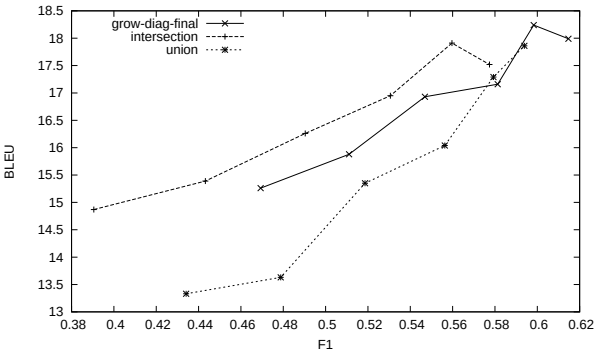
We also made the same manipulations to the predicted alignment data as we did in the French/English experiments, removing null alignments and, alternatively, making the alignments complete. Here the results are quite clear, with $\mathrm{WAA}_{\mathbf{F1}}$ showing better correlation with BLEU than do the other metrics. It also appears to indicate that in this data set, one with only Sure alignments, there was much less noise contributed by null alignments.

[11] Interestingly, for this result, $\mathrm{WAA}_{\mathbf{F1}}$ had higher correlation than did the best α value for F1, although this is not always true; of course, $\mathrm{WAA}_{\mathbf{F1}}$'s α can be tuned as well, although our preliminary take is that is shows less variation than in F1. We plan to investigate this further.

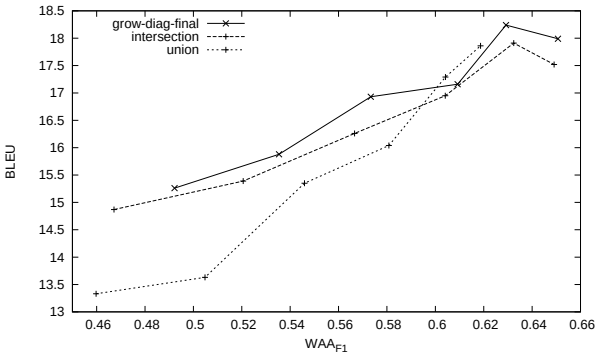
[12] Note that Moses as configured does not use the null alignments in its phrase tables, so adding or removing null alignments would have no effect on BLEU scores, thus our purpose in running these tests was to get a better idea of where the alignment metrics had difficulty.

[13] These results are also more in line with the first results with S and P in Fraser and Marcu (2006), suggesting perhaps that null alignments played a large role in those experiments as well, although there were a number of differences in the experimental setups.

[14] These additional experiments provide some potential insights, but will need to be investigated further. They do appear to tell us something about the gold set–that the P links are much lower quality, as would be expected. This is yet more evidence that Sure confidences should be used whenever possible.



(a) F1 versus BLEU, $r^2 = 0.73$



(b) WAA$_{\mathbf{F1}}$ versus BLEU, $r^2 = 0.87$

Figure 2: Correlation of Word Alignment Metrics versus BLEU, for Romanian/English, showing stronger correlation for WAA$_{\mathbf{F1}}$ than F1

Romanian/English Alignment Setting	1-AER	F1	WAA$_{F1}$
heuristics unchanged	0.73	0.73	0.87
heuristics no nulls	0.63	0.63	0.83
heuristics completed	0.57	0.57	0.77

Table 2: Summary of correlation (r^2) results for word alignment metrics and BLEU scores, for Romanian/English

Conclusions

We have presented a new word alignment metric called Word Alignment Agreement F1 (WAA$_{F1}$), which better accounts for word alignments than do current metrics such as AER and F1. We have demonstrated that WAA$_{F1}$ appears to better correlate with measures of translation accuracy, namely BLEU, for statistical machine translation systems. We are also interested in looking at this metric's relation with additional translation quality metrics, and in particular we would like to investigate its relation with human judgements of translation quality.

Of course, the amount of correlation for word alignment quality in general with translation quality is highly dependent on the way and degree to which translation systems employ alignments (Vilar et al., 2006). This effect will always be best measured by translation quality metrics themselves, and we suggest reporting translation quality results whenever claiming that improvements in alignments will lead to better translations.

WAA$_{F1}$ gains its strength from counting the words involved in alignments, instead of the links. As such, it should scale to different alignment environments, and yield well-defined results for words, phrases, and entire aligned corpora. Another benefit is that given its weighting scheme, it appears well-suited for use in a probabilistic framework, for example, when links are produced with numeric confidences.

References

Ayan, N. F., and Dorr, B. J. (2006). Going beyond AER: An extensive analysis of word alignments and their impact on mt. In *Proceedings of COLING-ACL '2006*, pages 9–16, Sydney, Australia.

Davis, P. C. (2002). *Stone Soup Translation*. Ph.D. thesis, Ohio State University.

Fraser, A., and Marcu, D. (2006). Measuring word alignment quality for statistical machine translation. Technical Report ISI-TR-616, University of Southern California, May.

Koehn, P.; Och, F. J.; and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 127–133, Edmonton, Canada, May.

Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA'2004*, pages 115–124, Washington, DC.

Lopez, A., and Resnik, P. (2006). Word-based alignment, phrase-based translation: What's the link? In *AMTA'2006*, pages 90–99, Cambridge, MA, August.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The Blinker project. Technical Report #98-07, IRCS, University of Pennsylvania.

Mihalcea, R., and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proc. of the HLT/NAACL workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May.

Och, F. J., and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China, October.

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

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual*

Meeting of the ACL, pages 160–167, Sapporo, Japan, July.

Papineni, K. A.; Roukos, S.; Ward, T.; and Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, September.

Vilar, D.; Popovic, M.; and Ney, H. (2006). AER: Do we need to ”improve” our alignments? In *Proceedings of the International Workshop on Spoken Language Translation*, pages 205–212, Kyoto, Japan.