Introducing a Word Alignment Dissimilarity Indicator: Alignment Links as Conceptualizations of a Focused Bilingual Lexicon

Devin Gilbert
CRITT dg@devrobgilb.com

Michael Carl
CRITT mcarl6@kent.edu

Abstract

Starting from the assumption that different word alignments of translations represent differing conceptualizations of cross-lingual equivalence, we assess the variation of six different alignment methods for English-to-Spanish translated and post-edited texts. We develop a word alignment dissimilarity indicator (WADI) and compare it to traditional segment-based alignment error rate (AER). We average the WADI scores over the possible 15 different pairings of the six alignment methods for each source token and correlate the averaged WADI scores with translation process and product measures, including production duration, number of insertions, and word translation entropy. Results reveal modest correlations between WADI and production duration and insertions, as well as a moderate correlation between WADI and word translation entropy. This shows that differences in alignment decisions reflect on variation in translation decisions and demonstrates that aggregate WADI score could be used as a word-level feature to estimate post-editing difficulty.

1 Introduction

Alignment error rate (AER) is a segment-based metric that compares one alignment (usually automatically generated) against another gold standard word alignment, assigning errors when the hypothesis alignment's links differ from those of the gold standard (Och and Ney, 2003). It is a normalized score with values between 0–1 for entire segments where a score of 0 indicates identical word alignments and a score of 1 indicates completely different sets of alignment links. When reported, AER scores are usually multiplied by 100 for readability. Usually an average AER score over many segments is reported, and automatic alignment systems have ranged between average AER scores of 3.7–50.6 (Liu et al., 2010) and 14.5–33.2 specifically for the English-to-Spanish language pair (Lambert, 2008). We have conducted alignment experiments with six different alignments of the same English-to-Spanish translations (a total of 1045 segments, 25936 tokens, translated by 31 participants), two manual alignments (M1, M2) and four automatic alignments (A1–A4).[1]

M1 is the original manual alignment (Mesa-Lao, 2014) which was later amended by another group of researchers. M2 is a realignment done by a group of researchers with very specific alignment criteria and, above all, the stipulation that only one aligner would sign off on the alignment of all translations for a given text in order to ensure consistency. A1 was aligned with GIZA++ (Och and Ney, 2003), trained on almost 2M en-es Europarl segments. A2 was aligned with SIMALIGN Match, A3 with SIMALIGN Argmax, and A4 with SIMALIGN Itermax (Sabet et al., 2020).

We obtain average AER scores between 8.8 and 26.3 (see Table 1). Perhaps not surprisingly, the lowest alignment scores (< 10.0, i.e., the most similar alignments) are between two automated alignment systems (A2-A4 and A3-A4) while the alignment scores between the two human alignments, M1 and M2, average out to 14.6. Also note that A4 is the automatic system that comes closest to human alignment M2 as well as human alignment M1. These scores give us a measuring stick with which to optimize word alignments, but we argue that word alignment links could be a much

[1] All data is publicly available on the CRITT website (Center for Research and Innovation in Translation and Translation Technology). For these alignments' study IDs in the CRITT Translation Process Research Database (TPR-DB), see Appendix A

Pairing	AER score	Pairing	AER score
M1-M2	14.6		
M1-A1	26.2	M2-A1	22.0
M1-A2	25.7	M2-A2	19.5
M1-A3	26.3	M2-A3	19.3
M1-A4	23.8	M2-A4	18.0
A1-A2	25.0	A2-A3	10.4
A1-A3	24.9	A2-A4	9.5
A1-A4	23.7	A3-A4	8.8

Table 1: Cumulative AER scores for six different alignments

richer source of information if we examine them on a more granular level than is afforded by AER. Instead of seeing dissimilarities between different alignment methods as errors, we suggest thinking of different word alignments as instantiations of a different contextualized and focused bilingual lexicon which may dynamically emerge in a translator's mind during the translation process.

We take it that alignment links are probabilistic in nature and that chances of two different alignment methods (human or machine) generating the exact same alignment links for any given segment are extremely slim. If we term a segment's set of alignment links an "alignment configuration," then a translation with m source words and n target words allows for 2^{m*n} unique alignment configurations. Think of a segment's alignment space as a grid where each source word is a row and each target word is a column. If a square of the grid is filled in, this represents an alignment link. The different possible patterns on this grid are an alignment configuration. A sentence with 10 source and 10 target words has 2^{100} (1.267e30) different possible alignment configurations[2]; finding the exact same alignment configuration on a segment level is not very likely.

Additionally, AER is usually reported for entire texts; an averaged AER score may be computed based on thousands of word alignments. While much effort has gone into developing systems to decrease global averaged AER (GIZA++, Och and Ney (2003); SIMALIGN, Sabet et al. (2020); FASTALIGN, Dyer et al. (2013); UALIGN, Hermjakob (2009); etc.), we posit that the agreement—as well as the disagreement—about alignment relations on the level of individual words carries crucial information about translation difficulties.

While some words may be 'easy' to align—i.e., with little or no discrepancies between different alignment methods—translational equivalents for other words may be disputable or 'controversial,' resulting in differences between different methods. In statistical MT, alignment links carry information about an underlying, contextualized bilingual dictionary. Along this line of thinking, differing alignments of the same translations represent differing conceptualizations of translational relations between words or phrases. Moreover, differing conceptualizations of translation equivalence point to potential discrepancies and difficulties, and therefore variation in alignment links could potentially be used as an indicator of ambiguity and translation difficulty.

In this paper we investigate differences among alignment links between individual source words and their translations as produced by different alignment methods. We posit that dissimilarities between different alignment methods are indicators for translational choice and difficulty, and we correlate variation in alignment links with other measures of translation difficulty such as production time. The next section discusses our method of calculating alignment error rate at the word level.

2 From segment to word alignment scores

Word alignments are commonly represented as sets of tuples, where each tuple represents one source-target alignment link. The first value in each tuple is the ordinal number of a token in the source segment; the second value in each tuple is the ordinal number of a token in the target segment to which the source word is linked. Figures 1 and 2 show two different alignment configurations of the same translation.

[2]This includes 'incomplete' phrase alignment with missing alignment links. Assume, for instance, the phrase translation {have bread ↔ Tengo pan} (see Figures 1 and 2). This should result in a set of alignment links {(1,0),(1,1),(2,0),(2,1)}. However, without further post-processing, MOSES' phrase-based system *grow-diag-final(-and)* may produce 'incomplete' phrase alignments in which one of the four alignment links may be missing, resulting in five possible alignment configurations for this phrase translation.

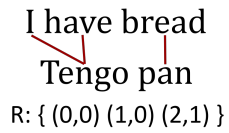

Figure 1: Reference

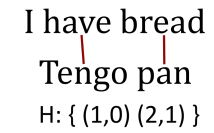

Figure 2: Hypothesis

Source sentence (S):	I	have	bread
$\text{WADI}_{\{0,1,2\}}$(R,H):	[1,	0,	0]

Table 2: WADI for 3-word source sentence

The first (reference) alignment configuration has two alignment links which connect the first and second source word ('I' and 'have') to the first word in the translation ('Tengo'), while the third source word ('bread') has a single alignment link that ties it to the second word in the target ('pan'). The set of hypothesis alignments consists of the same links with the exception to the first source word, which is unaligned. While the sets of reference and hypothesis alignments agree with respect to the translation equivalence of 'bread' and 'pan,' they differ on whether or not 'I' is conceptualized as being part of 'Tengo'—Spanish, as a pro-drop language, would also allow 'yo tengo'.[3]

$$\forall_{i \in S} \text{WADI}_i(\text{R,H}) = |\text{R}_i \cup \text{H}_i - \text{R}_i \cap \text{H}_i| \quad (1)$$

In order to assess the (dis)agreement between two alignment methods, we compute a word alignment dissimilarity indicator (WADI_i) that indicates the number of diverging alignment links for each source word position i. The WADI score (see Equation 1) takes as arguments the set of reference tuples (R) and hypothesis tuples (H) (see Figures 1 and 2) and produces a list that contains a WADI_i for every source word i which indicates the number of mismatches between the reference and the hypothesis. Table 2 shows the list of WADI results for the example in Figures 1 and 2 in which the first position corresponds to the first source word 'I' and a $\text{WADI}_i = 1$. For the two other positions ('have' and 'bread'), $\text{WADI}_i = 0$.

3 Examples of Alignment Dissimilarity

Here are some examples of high WADI scores that we have calculated between the M1 and M2 alignment methods.

Example 1:

Source
His withdrawal comes in the wake of fighting flaring up again in Darfur and *is set to* embarrass China ...
Target
Su retiro se produce a raíz de la lucha que surge de nuevo en Darfur y *tuvo lugar con el objetivo de* avergonzar a China..."

ST	W	M1	M2
is	4	tuvo lugar con el objetivo de	tuvo lugar
set	3	tuvo lugar con el objetivo de	con el objetivo
to	4	tuvo lugar con el objetivo de	de

Table 3: Alignment Dissimlarities in Example 1

One half of a segment from our English-to-Spanish data collection is shown in Example 1.[4] The respective alignments of M1 and M2 of the sub-segment "is set to" ↔ "tuvo lugar con el objetivo de" are shown in Table 3, together with their WADI scores (W in the Table). As the example shows, M1 aligns the ST and TT in a more compositional manner than M2. M1 linked 'is' as part of a three-word alignment group "is set to" and aligned it with a large target alignment group, "tuvo lugar con el objetivo de", which is repeated for each ST word in Table 3. M2, however, aligned more compositionally: {is ↔ tuvo lugar}; {set ↔ con el objetivo} and {to ↔ de}. WADI scores will be higher if one alignment method produces larger alignment groups than the other, as shown in Table 3.

Example 2 shows how alignments can have similarly long alignment groups yet high WADI scores because of the different conceptualizations of what these long alignment groups are equivalent to in translation.[5]

[3]Note that, according to our assumption above, this translation allows for $2^{3*2} = 64$ different alignment configurations, in which every ST word could or could not be paired with any TT word.

[4]Extracted from Participant 29's translation of segment 3 of multiLing Corpus Text 3. The text deals with Steven Spielberg not participating in the Beijing Olympics to protest China's backing of Sudan.

[5]Extracted from Participant 10's translation of segment 3 of multiLing Corpus Text 4. The text covers the topic of climate change and developing countries.

Example 2:

Source
Some of the most vulnerable countries of the world have contributed the least to climate change, *but are bearing the brunt of it*.
Target
Algunos de los países más vulnerables del mundo son precisamente los que menos han contribuido al cambio climático, a pesar de que *precisamente son algunos de los que más lo sufren*

The source word 'are' has a high WADI score of 5 because M1 aligns it by itself with the target word 'son'. This might seem to be a perfectly valid way to conceive of equivalence between the source and target, but when considering the other tokens in the surrounding phrase, we see that the M2 alignment also has a valid way of conceiving of the links of equivalence for this translation (see Table 4).

ST	W	M1	M2
but	1	a pesar de	pesar de que
are	5	son	precisamente son algunos de los que
bearing	6	algunos de los que más lo sufren	sufren
the	6	algunos de los que más lo sufren	más
brunt	6	algunos de los que más lo sufren	más
of	6	algunos de los que más lo sufren	lo
it	6	algunos de los que más lo sufren	lo

Table 4: Alignment Dissimlarities in Example 2

The source text in the last clause of this segment features a null subject due to a subtle bit of anaphora ('...but are bearing...'); the subject is inferred from the first part of the segment. The translator, in striving to create a target text that sounds natural in Spanish, has explicitated the subject by writing, 'son algunos de los que [are some of those that].' While M1 has this circumlocution aligned together with the verbal phrase 'bearing the brunt of it' yet separately from the verb 'are', M2 has treated this long stretch of text in Spanish as the argument of this clause and aligned the entire phrase together with 'are' while splitting the last verbal phrase into 'bearing', 'the brunt', and 'of it'. Additionally, M2 does not leave any target tokens unaligned, whereas M1 leaves 'que' and 'precisamente' unaligned. This example demonstrates how high-WADI tokens tend to "flock together" around longer alignment groups: all of the source tokens from this last verbal phrase, 'bearing the brunt of it,' have WADI scores of 6.

AER					
0	0–10	10–20	20–30	30–40	>40
% 13.4	28.5	27.4	18.4	8.4	3.9
WADI					
0	1	2	3	4	5+
% 76.1	11.8	7.7	2.5	1.1	0.8

Table 5: AER and WADI (M1 & M2) Distribution Pattern

4 WADI for different Alignment Methods

Table 5 compares the distribution patterns of WADI's word-level scores and AER's segment-level scores, as calculated between the M1 and M2 alignment methods. For WADI (M1-M2), the range of values is between 0 and 11. AER is a continuous variable whereas WADI is essentially categorical (ordinal) since it is only possible to have scores that are whole numbers. While calculated in a similar way, WADI and AER are quite different in how they are shaped. For example, WADI scores of zero are highly common in our data, while it is more rare to get AER scores of zero. Both have distributions that are skewed to right, but AER's distribution is much more even than WADI's. Let us compare what each metric shows us about our alignment methods. Figure 3 shows the relation between AER scores using M1 and M2 as references and A1 to A4 as hypotheses. Average AER scores show the automatic methods to be less similar to the manual methods than the manual methods are to each other (see Figure 3).

We can also see that when M1 is used as refer-

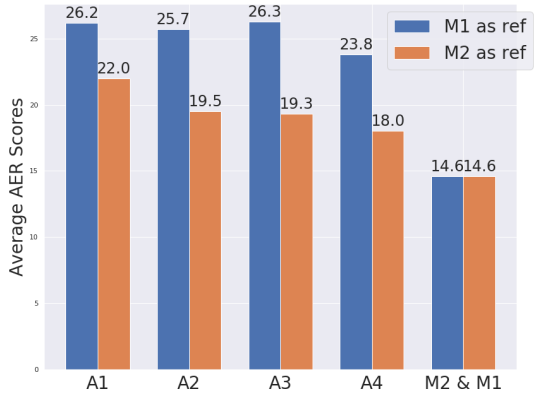

Figure 3: Average AER by Alignment Method

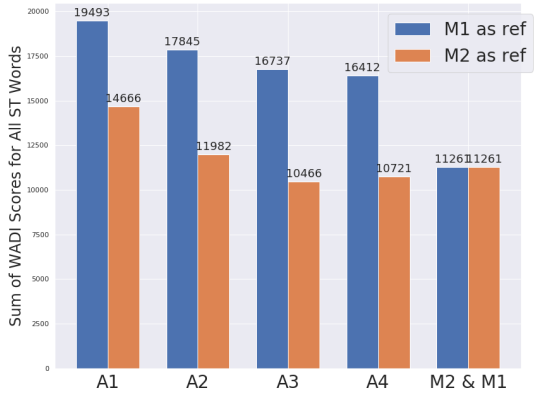

Figure 4: Sum WADI by Alignment Method

ence, all automatic alignment methods are nearly identical. However when M2 is used as reference, all automatic methods' average AER scores drop by 4.2–7.0, and A4 is 4.0 (rather than 2.4) lower than A1 and approaches the AER score of M1 (see Figure 3). These results would suggest that A4 would be the best automatic alignment method with respect to the gold standard.

When we consider the sum of WADI scores for each method instead of AER, we can see that the drop in dissimilarities persists when using M2 as reference instead of M1, but we can also see that both A3 and A4 (sum WADI scores of 10466 and 10721 respectively) are more similar to M2 than both manual methods are to each other (sum WADI score of 11261; see Figure 4). As opposed to the AER results, the WADI results suggest that A3 would be the best automatic alignment method with respect to the gold standard.[6]

[6]There are, of course, other considerations that go into what might be the "best" automatic alignment method for a particular use-case. For example, A3 leaves a lot more tokens unaligned than the other automatic alignment methods, which could make it less suitable for preparing data for translation process research or for using WADI scores as a quality estimation feature.

However, rather than thinking of the different sum WADI scores of each alignment method as an indication of the "best" alignment, we can also think of these WADI scores as measures of how much two alignment methods differ with regard to their conceptualization of a focused bilingual lexicon made up of all the words in the source and target text. If we take each source word to be the source-text component of one "entry" in a bilingual dictionary, then we can take the WADI scores as measures of agreement with respect to the target-text component(s) of that entry. Taking the example from Figures 1 and 2 in Section 2, the WADI score of 1 for 'I' shows the dissimilarity of the two alignments as to whether 'I' has a relationship of partial equivalence with the Spanish word 'tengo', and thus a different conceptualization of the bilingual lexicon.

Following this line of thinking, we can take the WADI data displayed in Figure 4 and conclude that methods A2–A4 all agree with M2 in their conceptualizations of our texts' bilingual lexicon about the same amount that M2 and M1 agree with each other. It is remarkable that automatic alignment methods agree with a human gold standard to the same degree that another human alignment agrees with this gold standard.

5 Examining WADI

Let us examine how WADI scores are distributed relatively by word class. Figure 5 shows a relative distribution for WADI scores of 0, 1, 2, and 3 or greater for the following word classes: adjective (Adj), adverb (Adv), function words (Func), nouns (N), numbers (Num), prepositions and conjunctions (PC), punctuation and other symbols (Sign), verbs (V), and wh-words (Wh). Figure 5 shows WADI scores for the two human alignments M1 and M2. It shows that adverbs and verbs are the least agreed-on (only 67% of adverbs and 71% of verbs have WADI scores of zero), suggesting that verbs and adverbs may be conceptualized differently in the bilingual translation lexicon more often than other word classes. On the other hand the alignment of punctuation and wh-words are the most agreed-on (91% and 88%, respectively, have WADI scores of zero), indicating that these items and their corresponding translations are less prone to dissimilar conceptualiza-

tions. Also, a surprisingly low share of function words have WADI scores of zero (73%).

We already observed some examples of verbs with high WADI scores with Examples 1 and 2 in Section 3. These examples showed how verbs in larger alignment groups had high WADI scores. The fact that verbs exhibit lower alignment agreement is consistent with research showing that verbs also tend to have significantly higher translation entropy values than other word classes (Ogawa et al., 2021), which indicates that translators tend to vary more when translating verbs. This could suggest that there is an association between variation in translation solutions and variation in how translations get aligned, which we test in Section 6.

Function words include determiners (e.g., 'the', 'a', 'this'), pronouns (e.g., 'they', 'he', 'their'), and the word 'to.' It makes sense that function words exhibit less agreement in alignment because the presence of these words across translations of the English-Spanish language pair is often asymmetrical, which would lead to function words tending to be aligned in larger alignment groups rather than by themselves, and this will tend to cause disagreements among aligners as to which neighboring words these function words get grouped with. This seems to be the case since the mean size (length in words) of target alignment group (TAGnbr) for M1 is 1.39, whereas mean TAGnbr for M2 is 1.06. Conversely, it makes sense for punctuation and wh-words to have high levels of agreement in alignment because there do tend to be clear-cut equivalents across languages for these two word classes, at least for the English-Spanish language pair.

Figure 6 plots relative shares of tokens belonging to different target alignment group sizes (TAGnbr), by word class. It shows TAGnbr figures from the M2 alignment method. Compared to most word classes, function words (Func) and punctuation/symbols (Sign) have a very high share of one-word target alignment groups (about 90% and 98%, respectively). This means that function words and symbols have less multi-word alignments. On the other hand, nouns have a large share of two-word alignment groups (over 30%; see Figure 6). It is also interesting to note that adverbs tend to be unaligned more often than other word classes (they have the highest share of target alignment groups of zero).

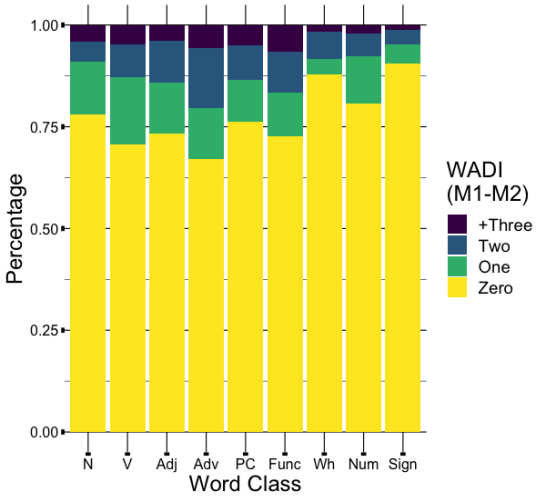

Figure 5: WADI Scores (M1-M2) by Word Class

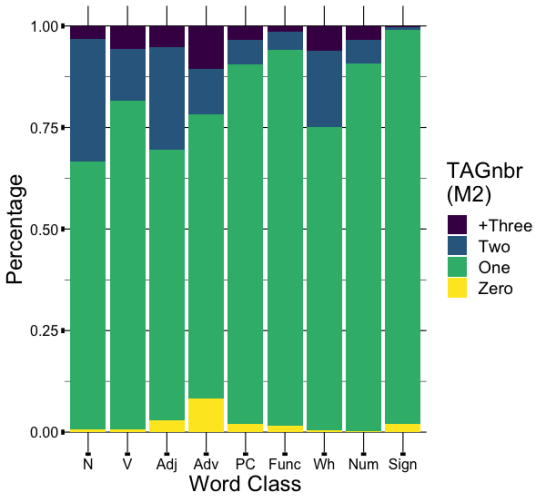

Figure 6: Target Alignment Group Size (M2) by Word Class

Comparing WADI scores (Figure 5) and the size of the target group by source word class (Figure 6), some interesting observations can be made. Function words exhibit alignment dissimilarity that is disproportionate to their high share of single-word alignment groups. This can be explained by the difference in mean TAGnbr between M1 and M2 that we discussed above. Another example: even though English nouns are more often linked to two Spanish target words, the alignment agreement seems to be a relatively uncontroversial; nouns have a higher share of zero WADI scores than verbs, adjectives, adverbs, function words, and prepositions/conjunctions (about 77%; see Figure 5). This could simply be due to the fact that many nouns occur in multi-word phrases yet are fairly straightforward to align because their trans-

lations have an easier-to-identify relationship of equivalence.

The correlation between WADI scores (M1-M2) and TAGnbr from the M2 alignments is significant yet extremely weak (Spearman $\rho(25934) = .08$, $p < .001$). However, the correlation between the same WADI scores and TAGnbr from the M1 alignments is remarkably stronger (Spearman $\rho(25934) = .55$, $p < .001$). This would seem to indicate that a great deal of the alignment differences that WADI (M1-M2) indicates are due to the discrepancies between TAGnbr for the M1 and M2 alignments.

6 Aggregating WADI across alignment methods

We calculate WADI scores for all 15 possible pairings of our six alignment methods and calculate the mean of these 15 different WADI scores for each source word. We investigate how this averaged value correlates with word-level translation process and product metrics such as production duration, insertions, and word translation entropy (HTra).

There is a positive, significant correlation between average WADI scores and log-transformed production duration per word, $r(25934) = .18$, $p < .001$ (see Figure 7), which is similar to the correlation between AER and production duration per segment (Spearman $\rho(1043) = -.11$, $p < .001$). There is also a positive, significant correlation between average WADI scores and number of insertions, $r(25934) = .28$, $p < .001$ (see Figure 8). Here we see a relationship between average WADI scores and behavioral indicators of translation effort which suggests that average WADI scores could be used as indicators for word-level quality estimation.

We also found there to be a positive, significant and moderate correlation between average WADI scores and HTra $r(25934) = .40$, $p < .001$ (see Figure 9). This demonstrates the relationship between the variation in alignment decisions (even among the four automatic alignment methods) and variation in translation. This evidence from production duration, insertions, and HTra leads us to conclude that aggregate WADI scores can be used as an indicator of translation (and post-editing) difficulty. That is, average WADI scores over several different alignment methods might be used to estimate post-editing difficulty on the word level.

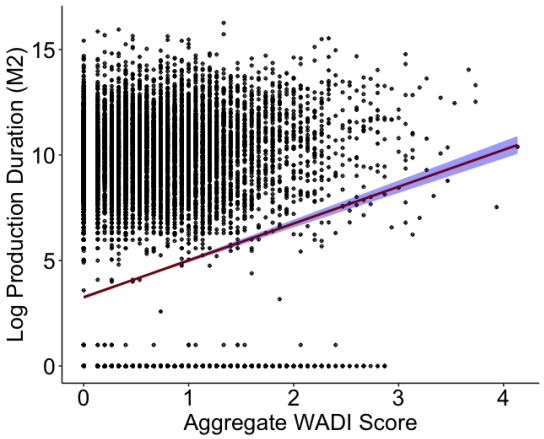

Figure 7: Scatterplot: Average WADI Scores and Log Production Duration

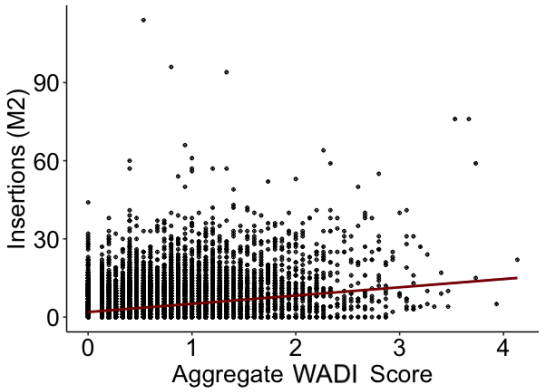

Figure 8: Scatterplot: Average WADI Scores and Number of Insertions

7 Conclusion

There are many ways to conceptualize equivalence in translation. We hypothesize that aligning translations is itself an act of declaring a bilingual focused dictionary, and different alignment relations represent differing possible conceptualizations of translation equivalents. We have developed a metric that, given two word alignments of the same translation, operationalizes dissimilar conceptualizations at the word level: word alignment dissimilarity indicator (WADI).

We observe that some word classes, such as verbs and adverbs, are more prone to dissimilar alignment conceptualizations while other word classes, such as wh-words, numbers, and punctuation/symbols are relatively uncontroversial in alignment. We also observe that size of alignment groups is related to word alignment dissimilarity,

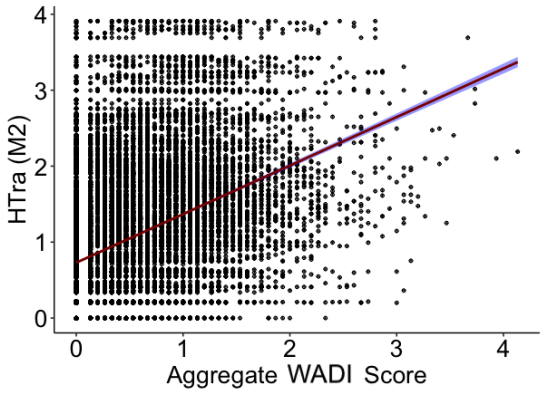

Figure 9: Scatterplot: Average WADI Scores and HTra

which shows that the fundamental conceptualization of what the source or target component of a unit of translation is could explain much of the observed variation in WADI.

Dissimilarities in word-to-word alignment between humans—but also between automatic alignment systems—of the same translations correlates with increased variation in the translation options produced by humans (i.e., aggregate WADI scores correlate with HTra), and we also observe a tendency for increased translation/post-editing effort—as indicated by production duration and number of insertions—to increase with WADI scores.

The observation that word alignments of different human annotators diverge substantially, and sometimes more than some automatic alignments differ from human alignments, suggests that there is no one gold standard for alignment relations. Rather, it stipulates that different conceptualizations of the same translation are possible and valid. Variation in translational conceptualization, however, has been shown to indicate translation difficulty and post-editing effort. The WADI score might captures some of these difficulties.

References

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Ulf Hermjakob. 2009. Improved Word Alignment with Statistics and Linguistic Heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 229–237, Singapore. Association for Computational Linguistics.

Patrik Lambert. 2008. *Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation*. Doctoral Dissertation, Universitat Politècnica de Catalunya, Barcelona.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Bartolomé Mesa-Lao. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors, *Post-editing of Machine Translation: Processes and Applications*, pages 219–245. Cambridge Scholars Publishing.

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

Haruka Ogawa, Devin Gilbert, and Samar A. Almazroei. 2021. redBird: Rendering Entropy Data and ST-Based Information Into a Rich Discourse on Translation: Investigating relationships between MT output and human translation. In Michael Carl, editor, *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications. Springer.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. In *EMNLP (Findings) 2020: arXiv:2004.08728 [cs]*, Online. ArXivLabs.

A Alignment Methods

Table 6 gives the CRITT TPR-DB study IDs for the six alignment methods used in this study.

M1	BML12	A2	BML12_SM
M2	BML12_re	A3	BML12_SA
A1	BML12_giza	A4	BML12_SI

Table 6: CRITT TPR-DB Study IDs for all Alignment Methods