On the Same Page? Comparing Inter-Annotator Agreement in Sentence and Document Level Human Machine Translation Evaluation

Sheila Castilho
School of Computing
Adapt Centre
Dublin City University
sheila.castilho@adaptcentre.ie

Abstract

Document-level evaluation of machine translation has raised interest in the community especially since responses to the claims of "human parity" (Toral et al., 2018; Läubli et al., 2018) with document-level human evaluations have been published. Yet, little is known about best practices regarding human evaluation of machine translation at the document-level. This paper presents a comparison of the differences in inter-annotator agreement between quality assessments using sentence and document-level set-ups. We report results of the agreement between professional translators for fluency and adequacy scales, error annotation, and pair-wise ranking, along with the effort needed to perform the different tasks. To best of our knowledge, this is the first study of its kind.

1 Introduction

Increasing efforts have been made in order to add discourse into neural machine translation (NMT) systems. However, the results reported for those attempts are somehow limited as the evaluation is still mostly performed at the sentence level, using single references, which are not able to recognise the improvements of those systems. The state-of-the-art automatic evaluation metrics have been shown to underestimate the quality of NMT systems (Shterionov et al., 2018), and the suitability of these metrics for document-level systems has also been criticised (Smith, 2017). For that reason, document-level human evaluation of machine translation (MT) has raised interest in the community recently as it enables the assessment of suprasentential context.

In a survey with native speakers, Castilho et al. (2020) tested the context span for the translation of 300 sentences in three different domains, namely reviews, subtitles, and literature. Over 33% of the sentences tested were found to require more context than the sentence itself to be translated, and from those, 23% required more than two previous sentences to be properly translated. Ambiguity, terminology, and gender agreement were the most common issues found to hinder translation. Moreover, differences in issues and context span were found between domains. This shows that document-level evaluation enables the assessment of textual cohesion and coherence types of errors which are impossible at times to recognise at sentence level.

Recent attempts to assess quality at the document-level were described in Toral et al. (2018) and Läubli et al. (2018) who independently reassessed the bold claims of MT 'achieving human parity' and found that the lack of extra-sentential context has a great effect on quality assessment, and pointed to a failure of the current best practices in MT evaluation. Toral et al. (2018) used consecutive single sentences to rank translations by two MT systems and a human reference. They found that the evaluators were able to better assess the translations when provided with more context, and moreover, inter-annotator agreement (IAA) between professional translators was higher than that between non-experts. However, this methodology does not discriminate sentence vs document-level set up as single sentences are shown consecutively.

Läubli et al. (2018) used pairwise rankings of fluency and adequacy to evaluate the quality of MT vs human translation (HT) for document-level texts. The methodology consists of translators choosing the 'best' translation in terms of i) adequacy and ii) fluency, that is, instead of choosing on a scale on how fluent or adequate the translations are, the raters just choose the 'best' one. Although not reporting IAA in the main paper, the authors report some IAA scores in the appendix of that work, showing that for fluency, document-level set up has

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1150–1159
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

higher IAA than for sentence set-up, but the opposite for adequacy. However, while this evaluation methodology seems appropriate when comparing different translations, it would not be feasible when evaluating a single MT system.

After these two papers were published, for the first time the WMT19 attempted a document-level human evaluation for the *news* domain. In that year, the direct-assessment (DA) task asked crowdworkers to give a score (0-100) regarding the accuracy of the translated sentence, where only one MT output is shown each time (no comparison with other MT system). However, conventional Kappa cannot be using with DA to measure IAA, and so consistency is measured instead, where raters have to pass some quality control criteria (Barrault et al., 2019).

In light of this, a comparison of IAA between quality assessments on sentence and document-level set-ups is needed in order to determine which set-up results in most reliable evaluation. This study presents a small-scale comparison on the differences in IAA between these two methodologies. To the best of our knowledge, this is the first paper to compare IAA for sentence vs document-level set-ups using the state-of-the-art MT evaluation metrics, namely fluency and adequacy scales,[1] error mark-up and pairwise ranking, along with reporting effort indicators.

We provide a detailed description of the experimental methodology in Section 2. Following, we report results in Section 3 on the agreement between professional translators for fluency and adequacy scales, error annotation, and pair-wise ranking (in 3.1), along with the effort needed to perform the different tasks (in 3.2). We discuss our results and draw conclusions in Section 4, and point out directions for future work.

2 Methodology

Five professional English (EN) to Brazilian Portuguese (PT-BR) translators were hired to perform the evaluation in terms of (1) fluency, adequacy, and error mark-up using the PET tool (Aziz et al., 2012); and (2) pairwise ranking using Google spreadsheet. The choice of PET was due to the fact that the tool is a free toolkit, easy to use, with its UI resembling translation tools, and it is able to handle different evaluations while logging time spent on tasks.

	Test Set 1	Test Set 2
Av. Sentence Length (WPD)	316	344
Av. Sentence Length (WPS)	20	21
Av. Sentence Count (SPD)	15	15
Total Words	10135	11019
Total Sentences	500	500
Total Docs	32	32

Table 1: Statistics for the test sets used, where average sentence length is calculated as words per document - WPD - (scenario B), words per sentence - WPS - (scenario A), and average sentence count is calculated as sentence per document - SPD.

	Test Set 1		Test Set 2	
	Source	Translation	Source	Translation
Flesch	47.9	57	50	55
TTR	0. 26	0.27	0.25	0.27

Table 2: Type-token Ratio (TTR) and Flesch Reading Ease Scores for both source and translated sides of Test Sets.

The evaluation was carried out in two scenarios: (A) evaluation at the sentence level, showing randomised sentences, one at a time, one score per sentence, and (B) evaluation at a document level, showing randomised documents, one document at a time, one score per document. After each scenario was complete, translators answered a post-task questionnaire about the evaluation and their perceived effort.

Corpus - We used the English corpora from WMT `newstest2019`, which has an average document length of 17 sentences (minimum 4 sentences, maximum 30 sentences). In total, 64 full documents were selected (32 per scenario) with 1000 sentences (500 per scenario). We made sure that both scenarios are comparable in terms of sentence and document length, as well as in terms of readability and lexical variation. Results on statistics of both data sets in Table 1, along with results for Type-Token Ratio and Flesch Reading Ease Scores in Table 2 (for both source and translated versions), show that, in fact, both test sets and translations are comparable. This suggests that any variation on the assessments is unlikely to be because the two test sets are overly distinctive.

The corpus was then translated from EN into PT-BR with Google Translate (for adequacy, fluency, and error mark-up) and also with DeepL for the ranking pairwise comparison.[2]

[1] It is important to notice that Läubli et al. (2018) used pairwise ranking of fluency and adequacy instead of the standard Likert scale.

[2] The online translators were used between 20-26 February 2020.

Translators - Five professional translators took part in the evaluation.[3] Their professional experience ranges from 6 to 14 years, and three of them have had previous experience with translation evaluation. A warm-up task with 20 sentences was set up so translators could get acquainted with the tasks, guidelines and tools, as well as clarify any doubts about the task. Each translator evaluated 1000 sentences, 500 in each scenario and Test Set. Table 3 shows the distribution of the task for each translator. No time limit was stipulated for the translators to finish the task, but they were asked to keep track of the time needed to complete the tasks.

Translators	T1,T5	T2	T3	T4
Test Set 1 (1-500 sent.)	S_1	S_2	D_1	D_2
Test Set 2 (501-1000 sent.)	D_2	D_1	S_2	S_1

Table 3: Distribution of tasks where S is sentence level and D is document level scenarios, and 1 and 2 is the order of the tasks.

Post-task Questionnaire - The post-task questionnaire consisted of 10 statements for each scenario, assessed on a scale from 1 to 6, where 1 is a negative answer (strongly disagree/very difficult/very tiring) and 6 is an affirmative answer (strongly agree/very easy/not tiring at all). The statements were the following:

1. I was *always* able to understand the meaning of the source [sentence/texts]
2. I was *always* able to understand the meaning of the translated [single sentence/full texts]
3. I was *always* able to recognise all the problems with the translation of [single sentence/full texts]
4. I would have preferred to evaluate [full texts/single sentences] than [single sentence/full texts]
5. I would have preferred to evaluate pair of sentences than [single sentence/full texts]
6. I would have preferred to evaluate full paragraphs than [single sentence/full texts]
7. I was satisfied with the evaluation I provided for the [single sentence/full texts] job
8. Spotting errors in the each translated [single sentence/full texts] was (difficulty level)
9. Assessing the translation quality on a [single sentence/full texts] level was (difficulty level)
10. Assessing the translation quality on a [single sentence/full level] was (fatigue):

3 Results

3.1 Inter-annotator agreement (IAA)

We compute IAA with Cohen's Kappa (Cohen, 1960) both weighted (W) and non-weighted (NW) as the most common statistics for IAA,

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ represents the proportion of times that the annotators agree, and $P(E)$ the proportion of times that the annotators are expected to agree by chance. While NW Kappa does not take into account the degree of agreement, W Kappa uses a predefined table of weights to measure the degree of disagreement between the two raters, the higher the disagreement the higher the weight. It is important to notice that in this case, W Kappa can only be calculated for adequacy and fluency as they are assessed using a Likert scale.

In addition to that, we also compute Inter-rater reliability (IRR) as the level of agreement between raters (percentage of matches), and Pearson correlation (r) between T1&T2 and T3&T4 (see Table 3). The comparison of the scenarios (sentence *vs* document) is calculated between the Test Sets (Test Set 1 & Test Set 2). We calculate IAA for all the tasks, namely adequacy, fluency, error and ranking. It is important to note that Fleiss Kappa is computed when analysing T1&T2&T5.

Due to the exploratory nature of this research, along with the small number of participants which is known to hinder the effectiveness of statistical analysis, we interpret the results gathered with these evaluations from a qualitative perspective.

3.1.1 Adequacy

Adequacy was assessed for each single sentence and full document (one score per document). Translators answered the question *"How much of the meaning expressed in the source appears in the translation?"* on a Likert scale from 1-4.[4] Table 4 shows the IAA scores for adequacy.

[3] At the time of submission of the paper, we had reported scores from 4 translators (T1, T2, T3 and T4). For the final submission we decided to add a 5th translator to be compared with T1&T2 in order to get a further understanding of issues observed with T1. Therefore, we keep the presentation of the scores between T1&T2 and, following, we present Fleiss Kappa scores with T1&T2&T5 as additional results.

[4] 1. None of it, 2. Little of it, 3. Most of it, 4. All of it

Adequacy		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	NW	0.13	0.01
	W	0.27	0.23
Pearson		0.5	0.64
p-value		0	0
IRR		47%	44%
Test Set 2		T3&T4	T1&T2
Kappa	NW	0.34	-0.06
	W	0.27	-0.12
Pearson		0.53	-0.37
p-value		0	0.03
IRR		63%	25%

Table 4: IAA for adequacy assessments for single sentences and full texts scenarios.

When looking at Test Set 1 (upper part of the table), we note that both W and NW Kappa show a higher score for single-sentence scenario. Interestingly, the difference between IAA scores for sentence and document for Test Set 2 (lower part) is very discrepant with IAA scores for document level reaching negative levels. This trend is supported by the negative correlation and the low IRR percentages. Since we have demonstrated that both test sets are comparable, we believe that translators T1 and T2 indeed disagreed on adequacy scores for the document scenario more than they did for the sentence scenario.

When adding T5 to the adequacy assessment, we see a decline in Kappa for both sentence-level and document-level scenarios, where *k*=0.04 and *k*=-0.12 respectively (Table 5) in contrast to *k*0.13 and *k*-0.06 (see Table 4). Conversely, we note an increase in IRR for both sentence and document-level scenarios, where IRR is 67% and 42% respectively (in contrast to 47% and 25%). These results draw near the results from T3&T4. Nevertheless, we note that IAA is higher when evaluations are performed in the sentence-level scenario.

Adequacy	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.04	0.01
IRR	67%	44%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.34	-0.12
IRR	63%	42%

Table 5: IAA for adequacy assessments for single sentences and full texts scenarios including T5.

3.1.2 Fluency

Fluency was also assessed for each single sentences and full documents (one score per document). Translators answered the question *"How fluent was the translation?"* on a Likert scale from 1-4.[5] Table 6 shows the IAA scores for fluency.

Fluency		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	NW	0.09	0.41
	W	0.06	0.25
Pearson		0.1	0.73
p-value		0.02	0
IRR		53%	56%
Test Set 2		T3&T4	T1&T2
Kappa	NW	0.27	0.05
	W	0.34	-0.02
Pearson		0.42	-0.11
p-value		0	0.53
IRR		57%	47%

Table 6: IAA for fluency assessments for single sentences and full texts scenarios.

When looking at Test Set 1, we note that IAA is higher in the document-level scenario for both W and NW Kappa when compared to the single-sentence scenario. This is confirmed by the linear relation expressed by Pearson. This might suggest that fluency is easier to assess with full texts rather than with non-contextual sentences. However, the same is not true when looking at Test Set 2, where W Kappa even reaches negative scores in the document-level set-up. Once again, we see that the IAA differences are bigger for T1&T2 who assessed Test Set 1 in the sentence-level scenario and Test Set 2 in the document-level scenario (see more discussion on this in Section 4), which is again confirmed by the negative correlation.

Fluency	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.88	0.41
IRR	63%	56%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.27	-0.12
IRR	57%	50%

Table 7: IAA for fluency assessment for single sentences and full texts scenarios including T5.

When adding T5 to the fluency assessment, we see a large increase in IAA for sentence-level scenario where *k*=0.88 and IRR=63% (Table 7) in contrast to *k*=0.09 and IRR=53% (Table 6). However, we note that apart from a slight increase in IRR for the document-level scenario (50% compare to 47%), Kappa shows a decrease reaching a negative value *k*=-0.12. With these new results, both

[5] 1. No fluency, 2. Little fluency, 3. Near native, 4. Native

Kappa and IRR are higher when evaluations are performed in the sentence-level scenario. However, by looking at the **translator pair** T3&T4, we can see that these two translators still agreed more when judging the document-level scenario (*k*=0.41) than when judging the sentence-level scenario (*k*=0.27).

3.1.3 Error

Error annotation was performed after translators assessed fluency and adequacy. Translators were asked to select from a drop-down menu which errors they found in the MT output. Because we are only interested in the agreement level between translators (as opposed to finding out the quality of the MT system), we decided to use a simple taxonomy that consisted of four error categories: Mistranslation, Untranslated, Word Form, and Word order. Translators could also select "No errors" in case the sentence/document did not contain any error. Each sentence or document could be annotated with more than one error category, but unfortunately because PET does not allow for word-level tagging, each error category could be assigned only once. Therefore, a segment or document could be tagged as containing all the errors, some of the errors, as well as no errors (no issues), but if the translator found that the segment contained *two* mistranslation errors, for example, the mistranslation category would be assigned only once to that segment. Yet again, we believe this set-up is enough to measure agreement levels.

Error mark-up results were divided into *binary*, when raters agree whether there was an error (any type) or no errors in the sentence/document,[6] and *type*, when raters agree on the exact error type found in the sentence/document. Table 8 shows the results for IAA for the error mark-up task.

The error annotation task shows higher IAA and IRR in document-level scenario for Test Set 1, however, the low Pearson correlation score does not indicate a strong linear relation. For Test Set 2, we see that sentence-level scenario shows higher Kappa for error *type* and higher IRR, confirmed with a positive correlation. It is important to note that Kappa for the *binary* classification in the document-level scenario is 1 (marked as n/a) as translators agreed that (almost) all documents contained **at least** one error. However, Kappa penalises it as all the ratings fall into a single category.

Error		SENTENCE	DOCUMENT
Test Set 1		T1&T2	T3&T4
Kappa	binary	0.28	n/a
	type	0.22	0.31
Pearson		0.21	0.08
p-value		0	0.49
IRR	binary	60%	100%
	type	50%	53%
Test Set 2		T3&T4	T1&T2
Kappa	binary	0.49	n/a
	type	0.38	0.20
Pearson		0.7	0.08
p-value		0	0.49
IRR	binary	76%	90%
	type	56%	33%

Table 8: IAA for error mark-up for single sentences and full text scenarios.

For this reason, we decided to also compute F-score for absolute error (disagreement) in the *binary* category (see Table 9).

ERROR	SENTENCE	DOCUMENT
Test Set 1	T1&T2	T3&T4
F-SCORE	60.4	100
Test Set 2	T3&T4	T1&T2
F-SCORE	76.6	93.75

Table 9: F-score for *binary* error mark-up evalaution.

F-scores show that indeed, *binary* classification is higher for the document-level scenario since we expect the full text to contain at least one error type. However, it is interesting to note that the document-level scenario for Test Set 2 presents only a 93.75 score and 90% (Table 8) since T1 and T2 disagreed in one document.

Error		SENTENCE	DOCUMENT
Test Set 1		T1&T2&T5	T3&T4
Kappa	binary	0.16	n/a
	type	0.02	0.31
IRR	binary	60%	100%
	type	56%	53%
Test Set 2		T3&T4	T1&T2&T5
Kappa	binary	0.49	-0.07
	type	0.38	-0.02
IRR	binary	76%	88%
	type	56%	50%

Table 10: IAA for error mark-up for single sentences and full text scenarios including T5.

By adding scores from T5 (Table 10), we note that IAA scores for Test Set 1 do not differ much, and document-level scenario shows higher Kappa and IRR as discussed previously. For Test Set 2, IAA scores decrease for Kappa, both for binary and error type categories. Interestingly, IRR scores for

[6] Intuitively, one might expect that at least one error will be found in a full document and so IAA will be high for document-level set-up in the binary category.

Ranking	SENTENCE	DOCUMENT
Test Set 1	T1&T2	T3&T4
Kappa	0.36	0.22
Pearson	0.41	0.36
p-value	0	0.04
IRR	59%	56%
Test Set 2	T3&T4	T1&T2
Kappa	0.29	0.19
Pearson	0.41	0.42
p-value	0	0.01
IRR	53%	47%

Table 11: IAA for Pair-wise ranking evaluation.

the binary category also slightly decreases. This is a bit surprising as we were expecting translators to assign at least one error type to full texts. The results with T5 indicate that, annotating error at a document-level is difficult as translators cannot tag exactly what the problematic parts are.

3.1.4 Ranking

Pairwise ranking was performed between translation from Google translate and DeepL. The systems' outputs (single sentences in scenario A, and full documents in scenario B) were randomly mixed so translators would see different outputs. Translators were asked to rate their preferred translation, and ties were allowed. Table 11 shows the IAA for the ranking task.

In Test Set 1, the ranking evaluation shows higher IAA for sentence-level scenario when compared to the document-level. Test Set 2 shows document-level scenario with lower agreement as seen in previous trend.

When adding scores from T5, we can see in Table 12 that IRR scores do not change. A 0.1 point decrease in Kappa scores can be observed for Test Set 1 for the sentence-level scenario (*k*0.36 to *k*0.26), and a slight decrease in Kappa scores for the document-level scenario fro test Set 2 (*k*0.19 to *k*0.14).

Rank	SENTENCE	DOCUMENT
Test Set 1	T1&T2&T5	T3&T4
Kappa	0.26	0.22
IRR	59%	56%
Test Set 2	T3&T4	T1&T2&T5
Kappa	0.29	0.14
IRR	53%	47%

Table 12: IAA for ranking assessment for single sentences and full texts scenarios including T5.

Interestingly, when looking at the output of both systems, Google seem to prefer to drop gender markers more than DeepL, which might make the sentence less adequate in terms of specifying who is speaking but the sentence can still be very fluent.

1) **Source**: *Her* decision to pull out left everyone involved absolutely stunned.
DeepL: A decisão *dela* de se retirar deixou todos os envolvidos absolutamente atordoados.
Google: *Sua* decisão de sair deixou todos os envolvidos absolutamente atordoados.

2) **Source**: To recover *it* is a duty."
DeepL: Recuperá-*lo* é um dever".
Google: Recuperar (x) é um dever."

This might suggest that translators' personal preferences play a role in document-level evaluation as well. For instance, translators might prefer adequacy over fluency, as in example 1, or in the case when there is not enough context in the source to specify the gender or solve ambiguity, translators might prefer the drop of the gender marker (as in example 2).

3.2 Effort

The effort spent on assessing the two scenarios was calculated in two ways: i) time translators spend assessing the sentences and full texts, and ii) self-report of effort required to perform the tasks via a post-task questionnaire.

Time - The time spent on evaluating Adequacy, Fluency and error mark-up could be drawn directly from PET logs. Unfortunately, it was not possible to count time for the ranking task because the pair-wise comparison was performed in Google Spreadsheet, and so no automatic log could be drawn. Although they were asked to keep track of their time while ranking the MT output, translators recorded this inconsistently. Therefore, we decide to use only the times logged in PET.

When performing the evaluation in PET, translators first had the chance to see the source and MT side by side in the post-editing window[7] and then to assess the MT output in another window. Therefore, PET logs two different times: one spent in the PE window, and one spent in the assessment window. Intuitively, one would believe that the translators would read the sentences/texts in the PE window and use the evaluation window only to

[7] It is worth noticing that the option of performing PE was disabled, so no time for any changes was counted.

Transl.		Reading	Assessing	Total
T1	Sent.	*09:29:33	*14:16:57	*23:46:30
	Doc	02:51:38	03:14:53	06:06:31
T2	Sent.	02:45:44	08:18:51	11:04:35
	Doc	03:25:39	02:08:26	05:34:05
T3	Sent.	05:42:25	03:07:27	08:49:52
	Doc	02:36:11	00:24:20	03:00:31
T4	Sent.	03:53:21	02:05:25	05:58:46
	Doc	02:41:15	01:13:46	03:55:01
T5	Sent.	00:35:22	05:43:11	6:18:33
	Doc	00:11:43	01:29:46	1:41:29

Table 13: Time spent on performing fluency, adequacy and error mark-up assessments in PET tool. (Note that T1 *times are compromised.)

give the scores. However, it is possible that some translators might have taken some time to read the source and MT in the assessment window.[8] Moreover, T1 reported that for the document-level scenario, they sometimes took a screenshot of the PE window "when the text was too long" and used it while evaluating the text in the assessment window (since the full text is not completely displayed in the PET assessment window).

For that reason, we decided to show both reading time (time spent on the PE window) and assessment time (time spent on the assessment window), and the total spent time. Table 13 shows the times spent for the task.

Unfortunately, T1 reported difficulties in carrying out the evaluation (due to COVID-19) and self-reported he was distracted while doing it, leaving the tool running mid-evaluation. For this reason, even when discarding obvious outlier times, there is a great discrepancy in the amount of time for T1 compared to the other translators: while translators had an average of 7-9 hours to complete the tasks, T1 took 23 hours to complete the task. Consequently, we decided to repeat T1's evaluation with T5 in order to see if patterns could be drawn from time spent on tasks.

Intuitively, one would expect translators to spend longer reading time for the document-level scenario compared to the sentence-level one, since full texts are longer. Furthermore, one would expect the assessing time to be longer for the sentence-level scenario since each sentence requires one assessment (500 assessments for 500 sentences), while in the document-level scenario, each text is assessed just once (32 assessments for 500 sentences). However, while T2 and T3 show longer reading time for document-level scenario, T1, T4 and T5 show lower reading time for that scenario.

Even though results for time do not seem to be a strong indicator of effort due to the PET's user interface limitation, it is interesting to note that while some translators do spend more time reading, some spend more time assessing. This might indicate that having the text available during the assessment of fluency/adequacy is essential for translators.

Post-Task Questionnaire - Translators answered the post-task questionnaire (see full statements in Section 2) after they finished all tasks in both scenarios. Table 14 show the average results for each statement (including T5's responses).

Statements	Sent.	Docs
1- understand source	5	5.4
2- understand translation	4.2	3.8
3- recognise problems	5.2	4.8
4- prefer (docs/single sent.) than (single sent./docs)	4	4.6
5- prefer pair of sentences than...	3.8	5
6- prefer full paragraphs than...	3.6	4.2
7- satisfied with evaluation	4.8	5
8- Spotting errors was (very easy - very difficult)	5.2	4.4
9- Assessing was (very easy - very difficult)	4.6	4.2
10- Assessing was (very tiring- not tiring)	3.2	1.8

Table 14: Post-questionnaire results (average). Scale range from 1 to 6 where 1 is strongly disagree/very difficult/very tiring and 6 is strongly agree/very easy/not tiring at all.

We observe a few interesting results for statements 1 and 2, where translators seem to be able to understand the meaning of the source better in the document-level scenario, but the meaning of the translation better in the sentence-level scenario. More interestingly, the average for statement 3 is slightly lower for document-level which might suggest that translators were less able to recognise all problems with the translation in the document-level scenario, likely due to the number of sentences.

Translators seem to prefer to judge single-sentences than full documents (statement 4), and, would rather evaluate sentence pairs (statement 5) or paragraphs (statement 6) than full documents.

Nevertheless, results for statements 8 indicate that translators found easier to spot errors in the full texts (which contradicts the results for statement 3). Previous work on evaluation of NMT systems (when compared to SMT) found translators find

[8] PET displays the MT and Source at the top of the assessment window.

NMT errors more difficult to identify due to its high fluency (Castilho et al., 2017) (at a sentence-level). This could be a good indication that, due to good levels of fluency in NMT systems, indeed the exhibition of full texts is more helpful for the assessment in general.

Finally, translators found the document-level scenario to be slightly easier to assess (statement 9) but much more tiring than assessing single sentences (statement 10).

4 Discussion

This paper attempts to shed light on the differences in IAA between sentence and document-level evaluation scenarios. The experiments performed with five professional translators have tested the state-of-the-art metrics commonly used to assess MT quality with humans, namely the assessment of fluency, adequacy, error mark-up and pairwise ranking (Castilho et al., 2018).

We note that when evaluating *adequacy* (Table 4) the scenario where single sentences are assessed show higher IAA for both test sets, and moreover, IAA for Test Set 2 presents the lowest IAA for the document-level scenario for all the metrics. Regarding *fluency* assessment (Table 6), document-level scenario for Test Set 1 has higher IAA but Test Set 2 the opposite is seen for Test Set 2.

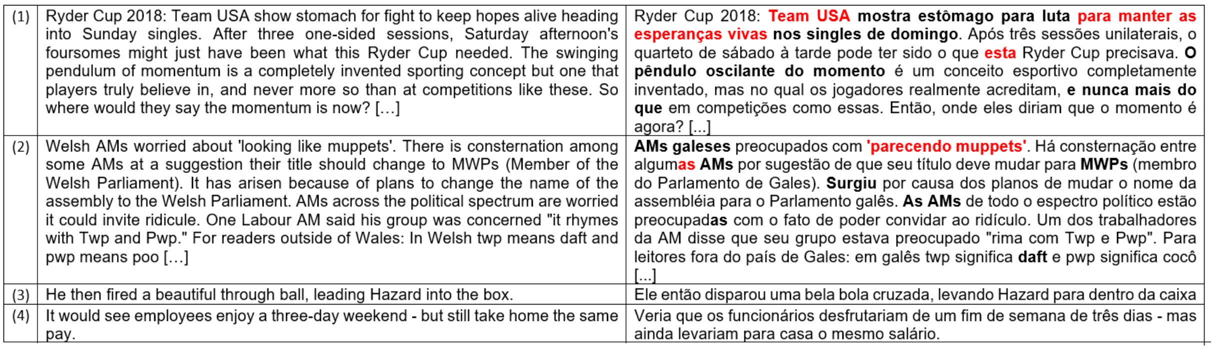
In addition to scores per test sets, it is interesting also to look at the IAA scores by **translator pairs**. We observe that there is a large difference between T1&T2 who evaluated Test Set 1 in the sentence-level scenario, and Test Set 2 in document-level scenario, against T3&T4 who evaluated the opposite. T1 and T2 tend to disagree more in both Test Sets for both fluency and adequacy assessments, while T3&T4 have closer IAA scores and higher Pearson correlation. The addition of T5's assessments, reported in terms of Fleiss kappa, indicate that for the majority of the case, IAA is indeed higher when evaluation is performed at a sentence level.

Figure 1 shows examples of disagreement between translators. In example (1), T1 assessed the text as containing "little" of the meaning of the source, T2 considered it to contain "all of it", and T5 assessed it as containing "most of it" (*adequacy*). T1 comments that "many mistranslations of golf/sport terms impaired meaning" and "some unstranslated terms found ('team USA', 'singles')", while T2 thinks that the text contains "minor issues, but the meaning isn't lost", and T5 says "the meaning is compromised by the word-by-word translation". While both T1&T2 agree that the text is "near native" regarding *fluency*, T5 assess it as having "little fluency", mentioning that fluency is also "compromised by the literal translation of some terms".

For T3&T4, the disagreement in the document-level (example (2)) is not as strong. While T3 assesses it as containing "most of the meaning", T4 thinks that it contains "little of it" because "there are a couple of plays on words in the source text, a big part of the translation is lost". However both agreed that regarding fluency, example (2) has "little fluency".

We speculate that the disagreements at the document-level scenario, especially for the adequacy evaluation, might be connected to the fact that because the texts are made up of "very good", "reasonably" and "poorly" translated sentences which, together, make the text understandable to a certain level, it is harder for translators to be consistent when assigning one single score for a full text. We estimate the percentages of *adequacy* scores for the document-level scenario as follows: T1&T2&T5 show 0% for score 1 (none of it), 7.29% for score 2, 61.46% for score 3, and 31.25% for score 4 (most of it); while T3&T4 show 4.69% for score 1, 17.19% for score 2, 64.06% for score 3, and 14.06% for score 4. These results show that a great number of scores fall into the middle category which makes it difficult for a consistent evaluation on a document-level scenario. Consequently, this type of problem will be persistent when evaluating at a document-level MT systems that operates at the sentence level, because a document translated with sentence-level NMT is still a sequence of translated sentences rather than an entire document translation.

We observe that disagreements in the sentence-level are more often related to ambiguity and lack of context. In example (3), while T1 commented that the translation "failed to use football terminology" and assessed it as containing "none of the meaning", T4 and T5 assessed it as containing "all of the meaning". We speculate that T4 and T5 were unaware that the sentence was about football due to the lack of context, and did not penalise mistranslations such as 'fired' which is better translated as 'chutar' (to kick) and 'box' which should be translated as 'pequena área'. T5 even mentioned that



(1)	Ryder Cup 2018: Team USA show stomach for fight to keep hopes alive heading into Sunday singles. After three one-sided sessions, Saturday afternoon's foursomes might just have been what this Ryder Cup needed. The swinging pendulum of momentum is a completely invented sporting concept but one that players truly believe in, and never more so than at competitions like these. So where would they say the momentum is now? […]	Ryder Cup 2018: **Team USA mostra estômago para luta para manter as esperanças vivas nos singles de domingo**. Após três sessões unilaterais, o quarteto de sábado à tarde pode ter sido o que **esta** Ryder Cup precisava. **O pêndulo oscilante do momento** é um conceito esportivo completamente inventado, mas no qual os jogadores realmente acreditam, **e nunca mais do que** em competições como essas. Então, onde eles diriam que o momento é agora? [...]
(2)	Welsh AMs worried about 'looking like muppets'. There is consternation among some AMs at a suggestion their title should change to MWPs (Member of the Welsh Parliament). It has arisen because of plans to change the name of the assembly to the Welsh Parliament. AMs across the political spectrum are worried it could invite ridicule. One Labour AM said his group was concerned "it rhymes with Twp and Pwp." For readers outside of Wales: In Welsh twp means daft and pwp means poo […]	**AMs galeses** preocupados com **'parecendo muppets'**. Há consternação entre algum**as AMs** por sugestão de que seu título deve mudar para **MWPs** (membro do Parlamento de Gales). **Surgiu** por causa dos planos de mudar o nome da assembléia para o Parlamento galês. **As AMs** de todo o espectro político estão preocupad**as** com o fato de poder convidar ao ridículo. Um dos trabalhadores da AM disse que seu grupo estava preocupado "rima com Twp e Pwp". Para leitores fora do país de Gales: em galês twp significa **daft** e pwp significa cocô [...]
(3)	He then fired a beautiful through ball, leading Hazard into the box.	Ele então disparou uma bela bola cruzada, levando Hazard para dentro da caixa
(4)	It would see employees enjoy a three-day weekend - but still take home the same pay.	Veria que os funcionários desfrutariam de um fim de semana de três dias - mas ainda levariam para casa o mesmo salário.

Figure 1: Examples of disagreement between translators.

they had problems with the word "Hazard" because eve thought "it seems to be a noun as it starts with a capital letter, I could not assess whether "Hazard" is a proper noun or just a noun, due to lack of context".

Example (4) is another example of lack of context, since the pronoun "It" is impossible to identify in the sentence. While T3 decides to rely only on the context given and assess *adequacy* as "all of it" and *fluency* as "native", T4 assesses the sentence as containing "little of the meaning" and "little fluency". According to T4, "the context was not enough to translate the pronoun 'it'". This is consistent with findings in Castilho et al. (2020) where authors found that over 33% of the surveyed sentences required more context than the sentence itself to be translated. Indeed, with the context of previous sentences it is possible to identify that "it" relates to "a radical plan" and therefore the addition of "O plano veria" (the plan would see) in the translation would make it more adequate:

(+2) Jeremy Corbyn's Labour Party is to consider **a radical plan** which will see Britons working a four day week - but getting paid for five.

(+1) The party reportedly wants company bosses to pass on savings made through the artificial intelligence (AI) revolution to workers by giving them an extra day off.

(S) ***It*** *would see employees enjoy a three-day weekend - but still take home the same pay.*

Interestingly, T1, T2 and T5 who assessed this sentence in context in the document-level scenario agreed that the text was "near native" and contained "most" and "all" of the meaning. This might be another indication that fluency is better assessed at a document level.

Regarding error mark-up assessment (Table 8), even though for Test Set 1 the document-level scenario has higher IAA than the sentence-level scenario, we note that when looking at **translator pairs**, the document-level scenarios has lower IAA scores for both test sets. Looking at T3&T4 both translators agree more on the error *types* found in the sentence-level than on the document-level scenario.

Finally regarding effort, unfortunately the logged time in PET tool was not decisive, even witht he addition of T5's assessments (due to discrepant results by T1 (Table 13)). Nevertheless, we believe that the results reported here show how difficult it is to run human evaluations, especially unsupervised ones. Additionally, the lack of proper tool able to handle different MT evaluation methodologies makes the assessment even more complex. We consider that time log gathered in PET can still be useful to draw specifications to develop a MT evaluation tool able to handle different methodologies. With respect to translators' self-assessment of their effort, the results from the post-task questionnaire showed that while translators prefer to see full texts than single sentences, they would rather see sentence pairs and paragraphs than having to assess full documents. This is not surprising since evaluating at a sentence-level is is what translators are used to already. Furthermore, they find assessing a full document more tiring than the alternative.

5 Conclusions and Future Work

The present work attempts to shed light at the differences in IAA when evaluating MT at the sentence and document levels with a small scale comparison. The main key findings of this comparison is that, a document-level evaluation methodology where translators assign one score per text leads to lower levels of IAA for adequacy, ranking, and error mark-up (when compared to methodologies where translators assign one score per sentence), but it might be useful for fluency assessments. This is consistent with (Läubli et al., 2018) findings on IAA for pairwise comparison, and previous work on NMT evaluation, where fluency proved to be harder to assess (than adequacy) in sentence-level scenarios.

Nevertheless, we also speculate that as Google Translate seems to operate on a sentence-level, a document-level evaluation of adequacy is penalised since a document can be constituted of sentences with different levels of quality. Moreover, we consider whether multiple scores per document (sentence pairs, paragraphs, and word-level error tagging) will yield higher levels of IAA when compared to the randomised sentence-level set-up for both sentence and document-levels MT systems.

Human-evaluation of MT in document-level setups is in its infancy, and therefore, it is essential to test which methodologies will be best suited for different tasks and domains. Future work will use more translators and different methodologies, as expressed in the post-task questionnaire and discussed above, with more specific guidelines for context-span issues found in previous works, and the development of test-sets, as well as using document-level MT systems' outputs.

Acknowledgments

I would like to thank the translators who participated in the experiment. Also a big thank you to Maja Popović and Joss Moorkens for the brainstorming sessions. This project was funded by the European Association for Machine Translation through its 2020 sponsorship of activities programme. The ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University is funded by the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded by the European Regional Development Fund.

References

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (WMT 19)*, pages 1–61, Florence, Italy.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to Human and Machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 9–38. Springer International Publishing.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of EMNLP*, pages 4791–4796, Brussels, Belgium.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus Automatic Quality Evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

Karin Sim Smith. 2017. On Integrating Discourse in Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT*, pages 113–123, Brussels, Belgium.