Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation

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

Abstract

Document-level human evaluation of machine translation (MT) has been raising interest in the community. However, little is known about the issues of using document-level methodologies to assess MT quality. In this article, we compare the inter-annotator agreement (IAA) scores, the effort to assess the quality in different document-level methodologies, and the issue of misevaluation when sentences are evaluated out of context.

1 Introduction

The use of machine translation (MT) has now become widespread in many areas thanks to improvements in neural modelling (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Accordingly, researchers have attempted to integrate discourse into neural machine translation (NMT) systems. As a consequence, document-level human evaluation of MT has raised interest in the community as it enables a more detailed assessment of suprasentential context. However, the definition of document-level, in terms of how much of the text needs to be shown, is still unclear. Moreover, although a few works have looked into document-level evaluation (Läubli et al., 2018; Toral et al., 2018; Barrault et al., 2019; Castilho et al., 2020), little is known about the issues of using document-level methodologies to assess MT quality.

The present research attempts to shed light on the differences in inter-annotator agreement (IAA) when evaluating MT with different methodologies, namely random single sentences, sentences in context, and full document scores. We also look into perceived effort from translators when evaluating the translations in the different methodologies. Results have shown a good level of IAA with a methodology where translators are able to assess individual sentences within the context of a document compared to a methodology with random sentence assessments, while a methodology where translators give a single score per document yields low IAA. Furthermore, we note that misevaluation cases recur in the random single sentences evaluation scenario.

2 Related Work

Document-level machine translation evaluation has been raising interest in the MT field, however, only a few works have attempted to use document-level boundaries for MT evaluation. Scarton et al. (2015) asked participants to post-edit and tag single sentences and full paragraphs in terms of cohesion and coherence. Their results showed that more post-editing was performed in paragraphs which suggests several issues could only be solved with paragraph-wide context. The authors reported Spearman's rank correlation for agreement which showed mixed to low agreement.

Toral et al. (2018) used consecutive single sentences to rank translations (in terms of preferred translation) of two MT systems and a human reference. They found that, when provided with more context, evaluators were better able to assess the translations, and moreover, IAA between professional translators was higher than that between non-experts. However, this methodology did not allow access to the full documents, as sentences were given one by one in order shown in the document.

Läubli et al. (2018) used pairwise rankings of fluency and adequacy to evaluate the quality of MT against human translation (HT) for document-level texts. The methodology consisted of translators choosing the 'best' translated documents in terms of i) adequacy and ii) fluency, that is, instead of choosing on a scale of how fluent or adequate the translations are, the raters just chose the 'best' one. The authors reported some IAA scores in the appendix of that work, showing that for fluency,

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 34–45
Online, April 19, 2021.

document-level set-up had higher IAA than sentence set-up, but that the opposite was the case for adequacy. However, this evaluation methodology can only be used when comparing two different translations.

Castilho et al. (2020) tested the context span for the translation of 300 sentences in three different domains (reviews, subtitles, and literature) in a survey with native speakers. The results showed that over 33% of the sentences tested were found to require more context than the sentence itself to be translated or evaluated, and from those, 23% required more than two previous sentences to be properly evaluated. The most common issues found to hinder translation were ambiguity, terminology, and gender agreement. Their results show that cohesion and coherence errors types cannot be recognised at sentence-level at times.

In 2019, the Fourth Conference for Machine Translation (WMT19)[1] attempted document-level human evaluation for the *news* domain for the first time (Barrault et al., 2019). Their direct-assessment (DA)[2] task asked crowdworkers to give a score (0-100) regarding the accuracy of the translated sentence, for one MT output. They asked raters to rate i) full documents, ii) single consecutive segments in original sequential order and iii) single random sentences. WMT20 (Barrault et al., 2020) modified the methodology and extended the context span to entire documents, asking raters to score individual segments whilst seeing the entire document, and also to judge the translation of the entire document. However, conventional Kappa cannot be used with DA to measure IAA, and so consistency is measured instead, where raters have to pass some quality control criteria.

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 the most reliable evaluation. This study is a follow-up of results presented in Castilho (2020) where we present a small-scale comparison on the differences in IAA between judgements given in isolated random sentences and entire documents. In the present study, we compare the IAA in evaluation of i) random single sentences, ii) evaluation of individual sentences while translators have access to the full source and MT output, and iii) evaluation of full documents. To the best of our knowledge, this is the first paper to compare IAA for random single sentence vs individual sentences in a document-level set-up using the state-of-the-art MT evaluation metrics, namely fluency and adequacy scales,[3] error mark-up and pairwise ranking, (Castilho et al., 2018) along with reporting effort indicators.

3 Methodology

3.1 Evaluation Design

Professional English (EN) to Brazilian Portuguese (PT-BR) translators were hired to perform the evaluation in terms of fluency, adequacy, error mark-up, and pairwise ranking using a spreadsheet. The evaluation was carried out in two scenarios:

1. **Sentence-level**: where translators give one score per random single sentence, henceforth *Random-Sentence score* - RSs.

2. **Document-level**: where translators give:

    - **A**: one score per individual sentence while having access to the full text. Henceforth, *Sentence-in-Context score* - SCs;
    - **B**: one general score for the full document, henceforth *Document score* - Ds. This evaluation was performed immediately after 2A.

This methodology is used to reflect the results of the first stage of this work (Castilho, 2020) and the context-span necessary for translation as seen in Castilho et al. (2020).

3.2 Corpus

Fourteen short documents (513 sentences) from various sources were selected: News from the WMT newstest 2019, Ted Talk from OPUS Corpus (Tiedemann, 2012), excerpts from two books, and product reviews.[4] These texts were selected because they consist of relatively short documents so it was possible to display the whole documents to translators. The two books were chosen because they were both

[1]WMT is running since 2006 and had always performed evaluation solely at the sentence level until 2019 (http://www.statmt.org/wmt19/).

[2]Direct assessment started in 2016 and was performed solely on single sentences until 2019.

[3]It is important to notice that Läubli et al. (2018) used pairwise ranking of fluency and adequacy instead of the standard Likert scale, while WMT uses direct assessments.

[4]The excerpts from both books were found freely available online: The Girl on the Train (www.bookbrowse.com) and The Fault in Our Stars (www.penguin.com). Product reviews were collect on the Amazon.com website.

narrated by female characters, which is important for translation of gender. Regarding user reviews, some were chosen because they do not contain information about the reviewer's gender, or about the product. Moreover, a few documents in the WMT News had the gender modified, for example, in a document where the politician was male, it was changed to female. These characteristics were selected to add challenging gender translations to the test set (Castilho et al., 2020).

3.3 Tools and MT systems

The collected corpus was translated from EN into PT-BR using Google Translate and DeepL. This language pair was selected because, as it is the researcher's mother tongue, it makes it possible to analyse the results more carefully and see possible patterns in the process. Additionally, as Portuguese is a romance language, it is possible that the results of this pilot can be extended to the language family.

While Google Translate was used for all the tasks, DeepL was used for a second translation for the ranking task. As we are mainly interested in finding out the best document-level methodology and annotator agreement as opposed to the quality of the translation, we believe that these two freely available MT system were adequate.

The tasks were set up on a spreadsheet since it proved to be the best tool where translator can see the full text at once (or most of it) and be able to judge fluency, adequacy and error at the same time.

3.4 Human Evaluation Metrics

We used the state-of-the-art MT evaluation metrics for this comparison, namely fluency and adequacy scales, error mark-up and pairwise ranking.

Adequacy was assessed for each scenario, RSs, SCs and Ds. Translators answered the question *"How much of the meaning expressed in the source appears in the translation?"* on a Likert scale from 1 to 4, where 1. None of it, 2. Little of it, 3. Most of it, 4. All of it.

Fluency was also assessed for each scenario, RSs, SCs and Ds. Translators answered the question *"How fluent was the translation?"* on a Likert scale from 1-4, where 1. No fluency, 2. Little fluency, 3. Near native, 4. Native.

Error mark-up - Translators were asked to select from a drop-down menu the types of errors found in the MT output. As 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" where the sentence/document did not contain any errors. Each sentence or document could be annotated with more than one error category, and each error category could be assigned more than once.

Translators	**Group 1**		**Group 2**	
	T1/T5	T2/T6	T3/T7	T4/T8
Test Set 1	S_1	S_2	D_1	D_2
Test Set 2	D_2	D_1	S_2	S_1

Table 1: Distribution of tasks where S is sentence-level scenario (RSs) and D is document-level scenarios (SCs and Ds), and 1 and 2 are the order of the tasks.

Pairwise Ranking was performed with translation from Google Translate and DeepL online MT systems. The systems' outputs were randomly mixed in each scenario so translators would see different outputs while ranking the translations. Translators were asked to rate their preferred translation, and ties were allowed.

3.5 Translators

Eight professional translators took part in the evaluation.[5] Their professional experiences range from 4 to 10+ years, and half of them have had previous experience with translation evaluation. Detailed guidelines on how to rate adequacy and fluency, tag errors and rank translations were made available and translators could ask for clarification for any doubts about the tasks. In order to avoid translators evaluating the same source twice, documents and scenarios were randomised. Each translator evaluated 513 sentences, 258 in scenario 1 (test set1- TS1) and 254 in scenario 2 (test set2 -TS2). Table 1 shows the distribution of the tasks for each translator, where Group 1 is made up of translators T1/T2/T5/T6, and Group 2, translators T3/T4/T7/T8.

3.6 Post-task Questionnaire

The post-task questionnaire consisted of 10 statements for the RSs and SCs scenarios. These were assessed on a scale from 1 to 6, where 1 is a negative answer (very difficult (statements 1-7) / very tiring (statement 8) / strongly disagree (statements 9-10)) and 6 is an affirmative answer (very easy/not tiring at all/strongly agree). Two additional statements for the assessment of fluency,

[5]Ethical approval has been obtained from the Dublin City University Research Ethics Committee.

Coeficients	Chance Correction	Weighted	# Raters	Measurement
Inter-rater reliability (IRR)	no	no	any	percentage
Cohen's Kappa	yes	no	2	interval 0-1
Weighted Cohen's Kappa	yes	yes	2	interval 0-1
Fleiss' Kappa (version of Scott's)	yes	no	any	interval 0-1
Krippendorff's Alpha	yes	yes	any	interval 0-1

Table 2: Inter-annotator coefficients comparison

adequacy and ranking were displayed for the Ds scenario (shown immediately after the statements for the SCs scenario). The statements for scenarios RSs and SCs were the following:

1. Understanding the meaning of the source [in the random sentences/in each sentence, with access to the full document] in general was
2. Understanding the meaning of the translated [in the random sentences/in each sentence, with access to the full document] in general was
3. Recognising the adequacy problems [in the random sentences/in each sentence, with access to the full document] in general was
4. Recognising fluency problems [in the random sentences/in each sentence, with access to the full document] in general was
5. Spotting errors [in the random sentences/in each sentence, with access to the full document] in general was
6. Choosing the best of two translations [in the random sentences/in each sentence, with access to the full document] was
7. In general, assessing the translation quality on a [sentence/document] level was (difficulty)
8. For me, assessing the translation quality on a [sentence/document] level was (fatigue)
9. I was confident with every assessment I provided for the [sentence/document] level evaluation tasks
10. I could have done a more accurate assessment if I [had had access to the full text/was assessing random sentences]

The additional statements for the Ds scenario were the following (note that statements including 'best target' and 'worst target' were only displayed for the ranking assessment):

1 Giving a general (adequacy / fluency / ranking) score for the full text was: (1 very difficult - 6 very easy)
2 In order to give a general (adequacy / fluency / ranking) score for each text, I had to re-read the full text:
Yes, both source and target texts
Yes, but only the target text
Yes, but only the best target text
Yes, but only the worst target text
No, I haven't re-read the full text(s). I remember it so I gave a general score according to my feeling of the translation

3.7 Inter-annotator agreement (IAA)

We compute IAA with some of the most common statistics for IAA in the field of computational linguistics (Artstein and Poesio, 2008). We compute IAA with some of the most common statistics for IAA in the field of computational linguistics (Artstein and Poesio, 2008). We compute **Cohen's Kappa** (Cohen, 1960)[6] both non-weighted and weighted versions.[7] We also use **Fleiss' Kappa** (Fleiss, 1971) which accounts for more than two raters, and **Krippendorff's Alpha reliability** (Krippendorff, 2011) which also applies to multiple coders, and allows for different magnitudes of disagreement. Fleiss Kappa and Krippendorff's Alpha are also used for the aggregated judgements within each condition. In addition to that, we compute a simple measure of percentage of agreement (we call it **inter-rater agreement - IRR**) calculated as the number of agreements, divided by the total number of assessments.[8] Table 2 summarises the features of each agreement coefficient.

The purpose of using Kappa-like coefficients for this study is to determine whether the assessments capture some kind of observable reality (Artstein and Poesio, 2008). Moreover, it is important to note that a discussion on the interpretation of the value of Kappa-like coefficients is beyond the scope of

[6] As Cohen's Kappa is designed for measuring the agreement between only two raters, when computing it for multiple raters, one can report the average of the Kappa statistics computed from each possible pair of raters (Mitani et al., 2017).

[7] Weighted Kappa was computed for the Adequacy and Fluency scores as they are assessed using a Likert scale, while non-weighted Kappa was computed for ranking and error tasks.

[8] All metrics were computed with Kappa built-in in SPSS software

Adequacy	RSs	SCs	Ds
Test set 1	**Group1**	**Group2**	**Group2**
Weighted κ (av)	0.40	0.41	0.30
Fleiss κ	0.32	0.29	0.13
Krippendorff α	0.50	0.51	0.28
IRR	70%	55%	68%
Test set 2	**Group2**	**Group1**	**Group1**
Weighted κ (av)	0.31	0.31	0.31
Fleiss κ	0.23	0.22	0.06
Krippendorff α	0.38	0.36	0.18
IRR	59%	59%	47%

Table 3: IAA for adequacy assessments for random single sentences (RSs), individual sentences in document context (SCs), and one score per document (Ds) scenarios.

Adequacy	RSs	SCs	Ds
Fleiss κ	0.10	0.10	-0.02
Krippendorff α	0.16	0.18	0.19

Table 4: Aggregated IAA scores for adequacy assessments.

this paper.

The comparison of the scenarios (1 - sentence *vs* 2 - document) is calculated between the test sets (Test Set 1 & Test Set 2) for a more detailed evaluation of the IAA scores, and scores are also generalised for each methodology. 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.

4 Results

4.1 Adequacy

Results for adequacy in Table 3 show that, in general, the RSs scenario has higher IAA than the document-level scenarios (SCs and Ds) for both test sets. Interestingly, if we look at IAA within each group, we note that group 2 has higher κ and α in the SCs scenario, even though the IRR is lower. The Ds scenario has the lowest IAA scores (apart from weighted κ for group 1). The aggregated scores in Table 4 show that Rs and SCs have the same Fleiss κ, while Ds shows negative scores. Interestingly, higher α is shown for the Ds scenario, followed by the Sc scenario. Nonetheless, we observe from the IAA scores that RS and SC methodologies seem to yield similar IAA scores, higher than the Ds scenario.

4.2 Fluency

Results for the fluency assessment in Table 5 show that for Test set 1, the RSs scenario has higher

Fluency	RSs	SCs	Ds
Test set 1	**Group1**	**Group2**	**Group2**
Weighted κ (av)	0.40	0.41	0.00
Fleiss κ	0.28	0.16	-0.03
Krippendorff α	0.46	0.27	0.07
IRR	69%	49%	47%
Test set 2	**Group2**	**Group1**	**Group1**
Weighted κ (av)	0.31	0.31	0.31
Fleiss κ	0.16	0.19	-0.20
Krippendorff α	0.26	0.29	-0.19
IRR	56%	61%	27%

Table 5: IAA for adequacy assessments for RSs, SCs, and Ds scenarios.

Fluency	RSs	SCs	Ds
Fleiss κ	0.09	0.05	-0.05
Krippendorff α	0.14	0.10	-0.06

Table 6: Aggregated IAA scores for fluency assessments.

IAA than both-document level scenarios, SCs and Ds. However, the SCs scenario shows slight higher IAA for Test Set 2. Within each group, we observe that group 2 has higher weighted κ and α in the SCs than in the RSs scenario, even though IRR is lower for the SCs scenario. The Ds scenario, in general, show lower IAA scores than RSs and SCs methodologies. The aggregated scores in Table 6 also confirms that the Ds scenario yields a lower IAA, and the RSs scenario shows slight higher IAA than SCs.

4.3 Error

Error mark-up results were divided into *binary*, when raters agree there was an error (any type) or no errors in the sentence/document, and *type*, when raters agree on the exact error type found in the sentence/document. Note that for the error mark-up task we decided not to ask translators to tag errors per document (Ds), for two main reasons: i) as it was proven to be hard for translators in our previous work (Castilho, 2020) and ii) as Ds scenario was evaluated immediately after RSs scenario, translators could just copy and paste the errors they have found in RSs into Ds.

Results in Table 7 show that IAA is higher for all assessments in the RSs scenario. However, we note that IAA scores for SCs, especially in Group 2, are closer to the ones in the RSs scenario. Moreover, the aggregated results in Table 8 show IAA for the SCs is similar to the RSs for the binary category, suggesting that a document-level methodology where translators can tag errors for each sentence with access to the full document can lead to

Error		RSs	SCs
Test Set 1		Group1	Group2
Cohen κ	binary	0.29	0.27
	type	0.28	0.25
Fleiss κ	binary	0.28	0.22
	type	0.27	0.24
α	binary	0.28	0.22
	type	0.27	0.24
IRR	binary	68%	63%
	type	65%	55%
Test Set 2		Group 2	Group 1
Cohen κ	binary	0.22	0.21
	type	0.25	0.21
Fleiss κ	binary	0.27	0.15
	type	0.25	0.16
α	binary	0.26	0.15
	type	0.24	0.16
IRR	binary	62%	60%
	type	58%	55%

Table 7: IAA for error mark-up assessments for RSs and SCs scenarios.

Error		RSs	SCs
Fleiss κ	binary	0.09	0.08
	type	0.10	0.86
α	binary	0.09	0.09
	type	0.10	0.08

Table 8: Aggregated IAA scores for error mark-up assessments.

better IAA.

4.4 Ranking

Results in Table 9 show that the RS scenario presents higher IAA compared to both document-level scenarios, while in test Set 2, it is the SC scenario which shows higher IAA. Interestingly, when looking within each group, we can see that the IAA scores are very close in the RSs and SCs scenarios. Moreover, the IAA scores when full texts were ranked (Ds) are largely lower compared to IAA scores where translators rank individual sentences with access to full texts (SCs). The aggregated scores in Table 10 confirm the low IAA when the Ds scenario is used, and close IAA for RSs and SCs.

4.5 Effort

The effort spent on assessment was calculated via a post-task questionnaire. Translators answered the questions (see full statements in Section 3) after they finished all tasks in all scenarios. Table 11 shows the average results for each statement for RS and SC scenarios.

We observe positive answers for the SC scenario for all statements which indicates that translators found it easier to understand both source (statement

Ranking	RSs	SCs	Ds
Test Set 1	Group 1	Group 2	Group 2
Cohen κ	0.41	0.37	-0.03
Fleiss κ	0.40	0.36	-0.12
Krippendorff α	0.45	0.39	-0.13
IRR	61%	58%	35%
Test Set 2	Group 2	Group 1	Group 1
Cohen κ	0.38	0.43	0.14
Fleiss κ	0.38	0.42	0.09
Krippendorff α	0.43	0.47	0.19
IRR	60%	62%	44%

Table 9: IAA for pair-wise ranking evaluation assessments for RSs, SCs, Ds scenarios.

Ranking	RSs	SCs	Ds
Test Set 1	Group 1	Group 2	Group 2
Fleiss κ	0.18	0.17	0.02
Krippendorff α	0.23	0.19	0.02

Table 10: Aggregated IAA for pair-wise ranking evaluation assessments.

1) and translation (2) when assessing sentences in context. Translators found it easier to recognise adequacy (3) and fluency (4) problems, as well as spotting errors (5) and choosing the best translation (6) when having access to full texts. Moreover, they found it easier to assess the quality in general (7) and less tiring (8) when having full texts, being more confident with their assessment (9). Overwhelmingly, translators think they give more accurate assessments when having access to full texts (10).

We also asked translators about the effort of giving one single score to the full texts (Ds). Table 12 shows the result for the statement "*Giving a general (adequacy / fluency / ranking) score for the full text was (1 very difficult - 6 very easy)*", while Table 13 show the result for the statement "*In order to give a general (adequacy / fluency / ranking) score for each text, I had to re-read the full text*".

We note that adequacy was the hardest assessment to be performed when translators are asked to give one score per document. One translator mentioned that both texts "*had lots of mistakes so I had to score based on quantity and quality of the mistakes, it took some calculations*. Another translator mentioned that "*Occasionally, a text would have some great individual sentences translation, but then would have missed some key words with mistranslations. So it was hard to think which factor should play a bigger role into the score*".

Regarding the question about re-reading the texts in order to assign one score for a full document, we see in Table 13 that for adequacy and fluency, while

Statements	RSs	SCs
1- Understand SOURCE	4.37	5.75
2- Understand TARGET	3.87	5.12
3- Recognise ADEQUACY	4.12	5.25
4- Recognise FLUENCY	4.62	4.87
5- Spot ERRORs	4.5	5.12
6- Choose BEST translation	4.12	4.87
7- Difficulty in assessing	4	5
8- Tiredness	3.75	4.62
9- Confidence	4.12	4.62
10- Preference	5.12	1.37

Table 11: Post-questionnaire results (average) for RSs and SCs scenarios. Scale range from 1 to 6, where 1 is very difficult/very tiring/strongly disagree and 6 is very easy/not tiring at all/strongly agree.

Statement	Adequacy	Fluency	Ranking
Difficulty level	4	4.37	4.5

Table 12: Average scores for assessing sentences in the Ds scenario, where 1 is "very difficult" and 6 is "very easy".

two translators re-read both source and target, 3 re-read the target only and 3 did not re-read the texts. For the ranking task, the majority of translators did not need to re-read any of the texts.

5 Towards a better human evaluation methodology for document-level

The recent interest in document-level MT evaluation has raised a few questions in the area. For example, it is still not clear how much context needs to be shown in a document-level evaluation set-up. This is important as we need to understand whether there is a pattern regarding how much context is required (in cases when full texts cannot be fully displayed or when they are not available) in order to have a reliable quality assessment and to avoid misevaluation issues. Some studies have used consecutive sentences (showing one at time) (Toral et al., 2018), and a few have used full short texts(Läubli et al., 2018; Barrault et al., 2019).

Castilho et al. (2020) have shown that for a great number of sentences, their successful translation and their MT evaluation requires more than sentence pairs and sometimes even full texts. Corroborating these findings, we also observe the need for a wider context in order to solve ambiguities in the evaluation. Figure 1 shows examples of context span needed when evaluating translations from EN→PT-BR.[9]

To evaluate sentence 105, the translators need

[9]The full speech can be found in the appendix I

Re-read (Y/N)	Adequacy	Fluency	Ranking
Source and target	2	2	2
Target only	3	3	1
Best target only	-	-	0
Worst target only	-	-	0
No	3	3	5

Table 13: Responses for the statement: *In order to give a general score for each text, I had to re-read the full text* displayed in the Ds scenario. Note that Best and Worst target only was only shown for the ranking assessment.

to identify the gender of the speaker in order to know whether "thank you" will be translated into the feminine (*obrigad**a***) or masculine (*obrigad**o***). For sentences 103 and 104, the translators need to know whether the pronoun "you" refers to singular or plural (*você/vocês*). Moreover, in sentence 104, because of the verb "love", some syntax constructions would need to have the gender of the pronoun "you" determined (***as** amamos -f, **os** amamos- m*).[10]

The issue of gender in 104 might be solved with sentence 102 with the use of the term "young women", as they are the ones who need "to stand up and take the reins", and the speaker knows that "you can do it" ("to stand up and take the reins"). This might imply that the "you" is also female and unlikely to be male, i.e.:

> *And we need strong, smart, confident young women to stand up and take the reins.*
> *We know you can do it, Paul.*×
> *We know you can do it, Mary.*✓

Before sentence 102, it is only in sentence 52 that "you" is clearly identified as "women". Regarding number (singular/plural), however, it is still not possible to affirm whether "you" in sentence 104 and 103 refers to singular or plural, with the context of sentence 102, because the one being talked to could still be singular, i.e.:

> *And we need strong, smart, confident young women to stand up and take the reins.*
> *We know you can do it, girls.*✓
> *We know you can do it, Mary.*✓

It is only with Sentence 99 that the number of "you" is solved with the term "every single one of you", which indicates that the speaker is talking to more than one person. Before that, it is only with sentences 52 and 54 that "you" is again identified as plural.

Regarding the gender issue in sentence 105, one can claim that sentence 95 "My husband works in

[10]It is also possible to translate the sentence "we love you" with the gender-neutral pronoun "vos" (nós vos amamos). However, as this is an old Portuguese construction, it is not considered by any of the translators.

1	Speaking at a London **girls' school**, **Michelle Obama** makes a passionate, personal case for each student to take education seriously.
2-25	…
26	I am an example of what's possible when girls from the very beginning of their lives are loved and nurtured by the people around them.
27-35	…
36	And these were the same qualities that I looked for in my own husband, Barack Obama.
37-51	…
52	You are the women who will build the world as it should be.
54	Not just for yourselves, but for your generation and generations to come.
55-90	…
91	My husband works in this big office.
92-97	…
98	Because we are counting on you.
99	We are counting on every single one of you to be the very best that you can be.
100	Because the world is big.
101	And it's full of challenges.
102	And we need strong, smart, confident young women to stand up and take the reins.
103	We know you can do it.
104	We love you (you).
105	Thank you so much.

Figure 1: Examples of context span needed to solve gender and number issues in sentence 104 and 105. The parts in pink relate to the gender of the speaker, red parts relate to the number of "you" (singular/plural), orange parts relate to the gender of "you", and the green parts relate to the resolution of "it".

this big office" would indicate that the speaker is feminine, however, men can also have husbands. Even if the following sentence identifies that the "big office" is the "Oval office", it does not clearly identify that the husband that works there is the actual president. It is only with sentence 36 that we see that the speaker is Michele Obama as she names the husband as "Barack Obama", however, that requires world knowledge. Sentence 26 is the closest one to 105 that clearly identifies the speaker as a "girl".

These problems with context span show that it is still uncertain how much context translators need to see in order to identify the issues in the translation and assess translation quality accordingly. We have previously shown (Castilho, 2020) that there is a risk of misevaluation when random single sentences are used in evaluation because of the lack of context. We also observe misevaluation issues in the present study, where disagreements in the RS scenario are more often related to ambiguity and lack of context. Figure 2 shows two examples of misevaluation for sentences 104 and 105 when assessments were performed in the RS scenario.

(104) Source: We love you.
MT: Nós te amamos. (no gender/singular)
HT1: Nós **as** amamos. (feminine/plural)
HT2: Nós amamos **vocês**. (no gender/plural)

When comparing the scores assigned to sentence 104 ("We love you") in the RSs and SCs scenarios, we note that in the RS evaluation, as expected, all translators assessed the MT output as having all the meaning of the source, to be native, and free of errors. None of the translators commented on the fact that there are four possible translation for the pronoun "you" in the source (singular/masculine, singular/feminine, plural/masculine, plural/feminine) and only a wider context would determine gender and number in the sentence. Translators who assessed sentence 104 in the SC scenario were able to find that the MT was not able to keep the gender agreement in the translation,[11] even when they erroneously did not consider the whole context in order to assess the sentence, as it is the case of T5. It is interesting that the scores for adequacy in the SC are quite divergent, while the scores for fluency are more homogeneous. This corroborates findings from our previous work Castilho (2020) where we note that disagreements at the document-level are more related to adequacy errors. Misevaluation was also observed in sentence 105:

(105) Source: Thank you so much.
MT: Muito obrigad**o**. (masculine)
HT: Muito obrigad**a**. (feminine)

Similar to the previous sentence, translators assessed the MT output of sentence 105 in the RS scenario as having all the meaning of the source, to be native, and free of errors. However, this time one translator (T4) commented on the fact that only a wider context would determine the gender in the sentence. Translators who assessed sentence 105 in the SC scenario were able to find errors in the MT output. Again, we note that T5 erroneously does

[11] Note that T2 considered (erroneously) the mistranslation to be a word form error

Sentence 104 S= We love you MT= Nós te amamos	Translator	Adequacy	Fluency	Errors	Comments
Random Sentence (RS)	T3	4. All of it	4. Native	No errors	
	T4	4. All of it	4. Native	No errors	
	T7	4. All of it	4. Native	No errors	
	T8	4. All of it	4. Native	No errors	
Sentence-in-Context (SC)	T1	3. Most of	4. Native	Mistranslation	amamos vocês
	T2	2. Little of	2. Little	Word Form	
	T5	4. All of it	4. Native	No errors	should be plural from context, but nothing in the sentence by itself leads us to that
	T6	3. Most of	4. Native	Mistranslation	

Sentence 105 S= Thank you so much. MT= Muito obrigado.	Trans.	Adequacy	Fluency	Errors	Comments
Random Sentence (RS)	T3	4. All of it	4. Native	No errors	
	T4	4. All of it	4. Native	No errors	without the context we cannot be sure about the gender (obrigado/obrigada)
	T7	4. All of it	4. Native	No errors	
	T8	4. All of it	4. Native	No errors	
Sentence-in-Context (SC)	T1	3. Most of	4. Native	Mistranslation	obrigada
	T2	3. Most of	2. Little	Word Form	
	T5	4. All of it	4. Native	No errors	should be feminine from context, but nothing in the sentence by itself leads us to that.
	T6	4. All of it	4. Native	Mistranslation	

Figure 2: Examples of misevaluation and (dis)agreement among translators in the RSs and SCs scenarios.

not consider the whole context in order to assess the sentence. Similar to sentence 104, the scores for adequacy in the document-level scenario are divergent, while the scores for fluency are more homogeneous.

We speculate that the reason methodologies with random single sentences show higher IAA agreement is because raters tend to accept the translation when adequacy is ambiguous but the translation is correct, especially if it is fluent. Thus, sentences like 104 and 105 are judged as correct in a scenario where there is no context to tell the evaluator *why* the translation should be different. Therefore, higher IAA scores in RS methodologies do not necessarily mean translators agreed more because the MT output was in fact better. Moreover, since NMT systems are known to have improved fluency, these types of misevaluation as shown previously are more likely to happen in a RSs set-up.

Our results have shown that a methodology where translators are able to assess individual sentences within the context of a document (SC scenario) yields good level of IAA compared to RS scenario, while a methodology where translators give one score per document (Ds) shows very low level of IAA. Moreover, the SCs methodology avoids the misevaluation cases which proved to be quite common in the RS evaluation set-ups.

6 Conclusions and Future Work

The present work attempts to shed light on the differences in IAA when evaluating MT with different methodologies, namely random single sentences, sentences in context, and full document scores.

The main finding of this comparison is that, an evaluation methodology where translators judge single random sentences might yield a better annotator agreement at times but with a high cost of misevaluation cases. Moreover, a methodology where translators assign one score per text leads to lower IAA and a great level of effort. This corroborates the results seen in Castilho (2020) where IAA scores for document-level reaches negative levels, and the level of satisfaction of translators with that methodology is also very low. In turn, evaluating the quality of MT output with individual sentences showed in the context of a document yields not only good IAA scores but avoids the issue of misevaluation which is extremely common in random single sentence evaluation set-up. We believe that a translator will be more inclined to accept as correct an ambiguous but fluent translation. This is problematic for an accurate evaluation of MT quality since it might lead to misevaluation especially when assessing the quality of NMT systems which are known to have an improved fluency level. Therefore, we suggest that evaluation set-ups using random single sentences should be avoided.

For future work, we will investigate the differences in context span needed for different domains, as well as whether the state-of-the-art metrics for human evaluation of MT (fluency, adequacy, error, ranking) must be modified in order to capture more realistically the quality level of the systems.

Acknowledgements

This project was funded by the Irish Research Council (GOIPD/2020/69) and partially by the

European Association for Machine Translation through its 2019 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

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, San Diego, CA.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

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, 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.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

Aya A. Mitani, Phoebe E. Freer, and Kerrie P. Nelson. 2017. Summary measures of agreement and association between many raters' ordinal classifications. *Annals of epidemiology*, 27:677–685.e4.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. 2015. Searching for context: a study on document-level labels for translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA.

A Appendices

Michelle Obama Ted Talk (Opus Corpus – Tiedemann, 2012)

1	Speaking at a London girls' school, Michelle Obama makes a passionate, personal case for each student to take education seriously.
2	It is this new, brilliant generation, she says, that will close the gap between the world as it is and the world as it should be.
3	culture,education,global issues,leadership,politics
4	Michelle Obama's plea for education
5	This is my first trip, my first foreign trip as a first lady.
6	Can you believe that?
7	And while this is not my first visit to the U.K., I have to say that I am glad that this is my first official visit.
8	The special relationship between the United States and the U.K. is based not only on the relationship between governments, but the common language and the values that we share, and I'm reminded of that by watching you all today.
9	During my visit I've been especially honored to meet some of Britain's most extraordinary women – women who are paving the way for all of you.
10	And I'm honored to meet you, the future leaders of Great Britain and this world.
11	And although the circumstances of our lives may seem very distant, with me standing here as the First Lady of the United States of America, and you, just getting through school, I want you to know that we have very much in common.
12	For nothing in my life's path would have predicted that I'd be standing here as the first African-American First Lady of the United States of America.
13	There is nothing in my story that would land me here.
14	I wasn't raised with wealth or resources or any social standing to speak of.
15	I was raised on the South Side of Chicago.
16	That's the real part of Chicago.
17	And I was the product of a working-class community.
18	My father was a city worker all of his life, and my mother was a stay-at-home mom.
19	And she stayed at home to take care of me and my older brother.
20	Neither of them attended university.
21	My dad was diagnosed with multiple sclerosis in the prime of his life.
22	But even as it got harder for him to walk and get dressed in the morning – I saw him struggle more and more – my father never complained about his struggle.
23	He was grateful for what he had.
24	He just woke up a little earlier and worked a little harder.
25	And my brother and I were raised with all that you really need: love, strong values and a belief that with a good education and a whole lot of hard work, that there was nothing that we could not do.
26	I am an example of what's possible when girls from the very beginning of their lives are loved and nurtured by the people around them.
27	I was surrounded by extraordinary women in my life: grandmothers, teachers, aunts, cousins, neighbors, who taught me about quiet strength and dignity.
28	And my mother, the most important role model in my life, who lives with us at the White House and helps to care for our two little daughters, Malia and Sasha.
29	She's an active presence in their lives, as well as mine, and is instilling in them the same values that she taught me and my brother: things like compassion, and integrity, and confidence, and perseverance – all of that wrapped up in an unconditional love that only a grandmother can give.
30	I was also fortunate enough to be cherished and encouraged by some strong male role models as well, including my father, my brother, uncles and grandfathers.
31	The men in my life taught me some important things, as well.
32	They taught me about what a respectful relationship should look like between men and women.
33	They taught me about what a strong marriage feels like: that it's built on faith and commitment and an admiration for each other's unique gifts.
34	They taught me about what it means to be a father and to raise a family.
35	And not only to invest in your own home but to reach out and help raise kids in the broader community.
36	And these were the same qualities that I looked for in my own husband, Barack Obama.
37	And when we first met, one of the things that I remember is that he took me out on a date.
38	And his date was to go with him to a community meeting.
39	I know, how romantic.
40	But when we met, Barack was a community organizer.
41	He worked, helping people to find jobs and to try to bring resources into struggling neighborhoods.
42	As he talked to the residents in that community center, he talked about two concepts.
43	He talked about the world as it is and the world as it should be.
44	And I talked about this throughout the entire campaign.
45	What he said, that all too often, is that we accept the distance between those two ideas.
46	And sometimes we settle for the world as it is, even when it doesn't reflect our values and aspirations.
47	But Barack reminded us on that day, all of us in that room, that we all know what our world should look like.
48	We know what fairness and justice and opportunity look like.
49	We all know.

Table A: Full speech by Michelle Obama's - see Figure 1

50	And he urged the people in that meeting, in that community, to devote themselves to closing the gap between those two ideas, to work together to try to make the world as it is and the world as it should be, one and the same.
51	And I think about that today because I am reminded and convinced that all of you in this school are very important parts of closing that gap.
52	You are the women who will build the world as it should be.
53	You're going to write the next chapter in history.
54	Not just for yourselves, but for your generation and generations to come.
55	And that's why getting a good education is so important.
56	That's why all of this that you're going through – the ups and the downs, the teachers that you love and the teachers that you don't – why it's so important.
57	Because communities and countries and ultimately the world are only as strong as the health of their women.
58	And that's important to keep in mind.
59	Part of that health includes an outstanding education.
60	The difference between a struggling family and a healthy one is often the presence of an empowered woman or women at the center of that family.
61	The difference between a broken community and a thriving one is often the healthy respect between men and women who appreciate the contributions each other makes to society.
62	The difference between a languishing nation and one that will flourish is the recognition that we need equal access to education for both boys and girls.
63	And this school, named after the U.K.'s first female doctor, and the surrounding buildings named for Mexican artist Frida Kahlo, Mary Seacole, the Jamaican nurse known as the black Florence Nightingale, and the English author, Emily Bronte, honor women who fought sexism, racism and ignorance, to pursue their passions to feed their own souls.
64	They allowed for no obstacles.
65	As the sign said back there, without limitations.
66	They knew no other way to live than to follow their dreams.
67	And having done so, these women moved many obstacles.
68	And they opened many new doors for millions of female doctors and nurses and artists and authors, all of whom have followed them.
69	And by getting a good education, you too can control your own destiny.
70	Please remember that.
71	If you want to know the reason why I'm standing here, it's because of education.
72	I never cut class.
73	Sorry, I don't know if anybody is cutting class.
74	I never did it.
75	I loved getting As.
76	I liked being smart.
77	I liked being on time.
78	I liked getting my work done.
79	I thought being smart was cooler than anything in the world.
80	And you too, with these same values, can control your own destiny.
81	You too can pave the way.
82	You too can realize your dreams, and then your job is to reach back and to help someone just like you do the same thing.
83	History proves that it doesn't matter whether you come from a council estate or a country estate.
84	Your success will be determined by your own fortitude, your own confidence, your own individual hard work.
85	That is true.
86	That is the reality of the world that we live in.
87	You now have control over your own destiny.
88	And it won't be easy – that's for sure.
89	But you have everything you need.
90	Everything you need to succeed, you already have, right here.
91	My husband works in this big office.
92	They call it the Oval Office.
93	In the White House, there's the desk that he sits at – it's called the Resolute desk.
94	It was built by the timber of Her Majesty's Ship Resolute and given by Queen Victoria.
95	It's an enduring symbol of the friendship between our two nations.
96	And its name, Resolute, is a reminder of the strength of character that's required not only to lead a country, but to live a life of purpose, as well.
97	And I hope in pursuing your dreams, you all remain resolute, that you go forward without limits, and that you use your talents – because there are many; we've seen them; it's there that you use them to create the world as it should be.
98	Because we are counting on you.
99	We are counting on every single one of you to be the very best that you can be.
100	Because the world is big.
101	And it's full of challenges.
102	And we need strong, smart, confident young women to stand up and take the reins.
103	We know you can do it.
104	We love you.
105	Thank you so much.

Table A: Cont. - Full speech by Michelle Obama's - see Figure 1