Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3735–3742
Marseille, 11–16 May 2020

On Context Span Needed for Machine Translation Evaluation

Sheila Castilho, Maja Popović, Andy Way
ADAPT Centre, School of Computing
Dublin City University
{sheila.castilho, maja.popovic, andy.way}@adaptcentre.ie

Abstract

Despite increasing efforts to improve evaluation of machine translation (MT) by going beyond the sentence level to the document level, the definition of what exactly constitutes a "document level" is still not clear. This work deals with the context span necessary for a more reliable MT evaluation. We report results from a series of surveys involving three domains and 18 target languages designed to identify the necessary context span as well as issues related to it. Our findings indicate that, despite the fact that some issues and spans are strongly dependent on domain and on the target language, a number of common patterns can be observed so that general guidelines for context-aware MT evaluation can be drawn.

Keywords: MT evaluation, document-level MT evaluation, human evaluation

1. Introduction

One of the biggest challenges for MT is the ability to handle discourse dependencies and the wider context of a document. Although currently an active community is working on developing discourse-level MT systems, the improvements reported for those systems are still limited. One of the main reasons for this is that the evaluation of document-level systems (both automatic and human) has primarily been performed at the sentence level and is, therefore, unable to recognise the real improvements of document-level systems.

Furthermore, two recent studies (Toral et al., 2018; Läubli et al., 2018) independently reassessed the bold claims of MT "achieving human parity" (Hassan et al., 2018) and found that the lack of extra-sentential context has a great effect on quality assessment. After these two papers were published, WMT19 considered their criticisms and attempted, for the first time, a document-level human evaluation for some language pairs.

Despite the increased in the field, both for extending MT systems to operate on the document level, as well as for improving the evaluation methodology by expanding it to the document level, the definition of what exactly constitutes a "document level" evaluation is still not clear. At this point, it is uncertain whether it refers to pairs of consecutive sentences, to a paragraph, or even to whole chapter. Our paper tests the context span necessary for human evaluation on three different domains, guided by the two research questions:

- RQ1 - Are two previous sentences enough for reliable evaluation?

- RQ2 - What are the linguistic issues related to the necessary context span?

Despite certain limitations of our experiments (small corpora, unbalanced number of participants for different target languages and domains, as well as lack of specific guidelines), we believe that our findings represent a good starting point towards reliable context-aware MT evaluation.

In the following section (section 2.), we present the related work in MT evaluation with document-level set-ups, and our rationale that guides this experiment is given in section 3. Section 4. describes in detail our methodology, followed by section 5. where we present the results and discuss our findings. In section 6., we present our general observations and general guidelines for MT evaluation context, with suggestions for future research.

2. Related Work

Although the term "document level" has been used freely to refer to MT systems handling context beyond the sentence level, the definition of what exactly constitutes a document-level is not yet well defined. Work on document-level MT show that those systems mostly use a context span of sentence pairs (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Müller et al., 2018) and very few have attempted to go beyond that span (Voita et al., 2019). For some of the developed context-aware MT models, test suites have been designed to better evaluate translation of the addressed discourse-level phenomena (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019).

As for overall MT evaluation, a few attempts have been made to perform human evaluation with document-level set-ups. To reassess claims of "human parity" in MT, Toral et al. (2018) used consecutive single sentences (opposed to randomised single sentences in Hassan et al. (2018)) to rank translations by two MT systems (Microsoft and Google) and a human reference from the WMT 2017. They found that the evaluators were able to better assess the translations when provided with more context. The authors suggest that human evaluation should take the whole document into account instead of single sentences in isolation.

Another work on reassessing the claims of human parity is that of Läubli et al. (2018), who used the test set from WMT 2017 to compare sentence-level evaluation versus document-level evaluation. Professional translators evaluated entire documents as well as single isolated sentences, and ranked the MT and human translation (HT) in terms of adequacy and fluency. Their results show that, while the sentence-level raters found it harder to discriminate between HT and MT, document-level raters clearly preferred HT than MT, especially in terms of adequacy. The authors argue that document-level evaluation enables identifying certain types of errors which are impossible to spot in a sentence-level evaluation, such as ambiguous words or errors related to textual cohesion and coherence.

After these two papers were published, WMT19 organisers attempted for the first time a document-level human evaluation for some language pairs (Barrault et al., 2019). Two evaluation configurations were set up: 1) a sentence-score + document level evaluation, i.e. assigning sentence level scores where sentences are shown in their natural order as they appear in the document, and 2) a document-score + document-level evaluation, i.e. assigning a score for the whole document. However, assigning a score for the whole document has shown to be problematic in terms of small sample size and inconclusive ties, so that assigning a sentence-level score with available context is recommended.

Nevertheless, none of the papers report any information about the definition or the size of the so-called "documents". Läubli et al. (2018) and Barrault et al. (2019) evaluated news articles which consisted of short texts – therefore the evaluators could assess the whole news article (referred to as "document") without concerns about the portion of the document needed for reliable evaluation. For evaluating different types of texts, the ratings might change depending on the context span shown to evaluators because there are many sentences which can be correct in isolation but incorrect in a number of contexts.

This paper attempts to shed some light on this question by conducting an extensive survey involving three domains and various languages in order to identify how much context span is needed to correctly evaluate specific sentences and which linguistic factors have influence on this span. Our aim is to aid both human and automatic evaluation, and our findings can also indicate research directions for context-aware MT systems.

3. Rationale

An MT system can often produce correct translations of isolated sentences which end up being incorrect when put in a certain context. For such sentences, a human evaluator cannot be completely sure whether the MT output in isolation is indeed correct. Our method is based on the fact that the same applies to human translation: for such sentences, a translator seeing them in isolation is not completely sure how the parts depending on the context should be exactly translated.

sent1	I put it in my car.	what is "it"?
+1 pr.	What did you do with the suitcase? I put it in my car.	it=SUITCASE
sent2	Yes, she did.	main verb?
+1 pr.	Did she give you any? Yes, she did.	main verb=GIVE what is "any"?
+2 pr.	So you went to your wife for money. Did she give you any? Yes, she did.	main verb=GIVE any=MONEY
sent3	Are you sure?	number of "you"? number and gender of "sure"?
+ 1 f.	Are you sure? I certainly am.	you,sure=SING
+ 2 f.	Are you sure? I certainly am. Thank you, Ms. Jones.	sure=FEMALE

Table 1: Examples of potential problems for evaluation of isolated sentences and resolving them by adding context: the first sentence is resolved by one preceding sentence, the second one with two, whereas the third one is resolved by two following sentences.

Three examples of problematic sentences are shown in Table 1. For many target languages with grammatical gender (for objects, animals, etc.), it would not be possible to translate the first sentence in isolation because it is not known what "it" refers to. When the preceding sentence is available, too, everything becomes clear and the sentence can be translated using the corresponding gender of "suitcase" in the given target language. Therefore, the necessary context span is +1pr (previous). The second sentence cannot be translated into (a number of) target languages which require the main verb missing in the English source. When the previous sentence is added, the problem with the verb is resolved (the main verb is "give"), but another problem is introduced – it is not clear what "any" refers to. This information might be necessary for gender, as well as for the correct pronoun. The problem is finally resolved with the second preceding sentence and no additional problems are introduced, so that the necessary context span is +2pr. For the third sentence, the number of the English pronoun "you" is not clear, neither are number and gender for the adjective "sure". The following sentence resolves the number both for the pronoun and for adjective (singular), but the gender is still not clear. Finally, the second following sentence contains the information about gender, so the necessary context span is +2f (following). In addition, for some target languages, the information about formal vs. informal "you" is necessary, too, which can be resolved by the second sentence: the fact that "Ms. Jones" is used instead of the first name implies a formal register.

The span and exact problems related to it can vary depending on the language pair and domain. Therefore, we perform our experiments on three different domains and several language pairs always including English as the source language.

4. Experimental set-up

Firstly, we prepared forms to be filled by native speakers of the target language consisting of two main questions: is it possible to translate the given sentence, and if not, what are the problems preventing the translation.

The experiment consists of three parts: in the first part, the *isolated sentences* were given to the participants. For each of the sentences found to be globally problematic, *two consecutive preceding sentences* were given to the participants and the same question was asked in part two. The problematic sentences were defined as those where over 40% of participants agreed that the sentence could not be translated without the context. Then the sentences which remained problematic even with two preceding sentences were analysed in order to identify the most prominent factors related to the larger context span. These sentences were then analysed in depth for two target languages – Portuguese and Serbian – in order to estimate the maximal span needed for reliable evaluation.

4.1. Data sets

Three data sets from three different domains/genres were chosen for this experiment:

Literature Domain: 250 sentences from *Alice in Wonderland* with 5920 running words were selected from the OPUS[1] data (Tiedemann, 2012).

Subtitles: Subtitles from three movies ("Joan of Arc", "Accused", "Phantom") and four TED talks ("sleep", "philanthropy", "garden", "astrophysics") were selected from the OPUS data. In total, 1345 segments and 18828 running words were available.

User reviews: User reviews about Amazon products, hotels and restaurants from Trip Advisor, as well as IMDb movie reviews (176 segments and 13219 running words in total) were collected from the web.

Text statistics can be seen in Table 2.

corpus		# seg.	# words
literature	"Alice"	250	5920
subtitles	movies	933	9537
	TED	412	9291
	total	1345	18828
user reviews	Amazon	24	2939
	TripAdvisor	71	3103
	IMDb	81	7177
	total	176	13219

Table 2: Text statistics for the three used data sets: literature, subtitles and user reviews.

4.2. Participants

Participants from 18 different languages responded to the survey: Spanish speakers (14), Greek (12), Portuguese (10), Chinese, Finish, French and German (4 respondents each), Italian and Irish (3 respondents each), Amharic, Arabic, Bulgarian, Catalan, Croatian, Dutch, Indonesian, Russian and Serbian (1 respondent each).

When asked about their English level, 85% self assessed as having a C2 or C1 English level, and 62% holds an English language certificate. In addition, six bilingual participants were native English speakers. When asked about their experience with translation, 81% responded they had at least *some* experience with translation, varying from a few months to over five years. Their experience with domains include general, technical, scientific, literary, tourism, audio-visual, legal, certified, creative, poetry, etc. When asked what tools they used for their translation, 77% answered they still use a word processor (i.e. Word, notepad), while 19% stated they use some CAT (computer-aided translation) tool only. Finally, when asked if they make use of any MT system for their translation, 32% stated they *never* use or used any MT system.

4.3. Survey

In order to collect answers from the participants, a Google Form survey was set up. The survey was split into two phases: in Phase I, single isolated random sentences were shown to the participants, while Phase II showed sentences together with their two preceding sentences (see Figure 1). A website[2] was set up to provide participants with a detailed explanation of the survey, its goals, as well as to provide with the Consent Forms and Plain Language statement.[3] Participants were asked to fill in a background questionnaire before they started answering the questions in the survey.

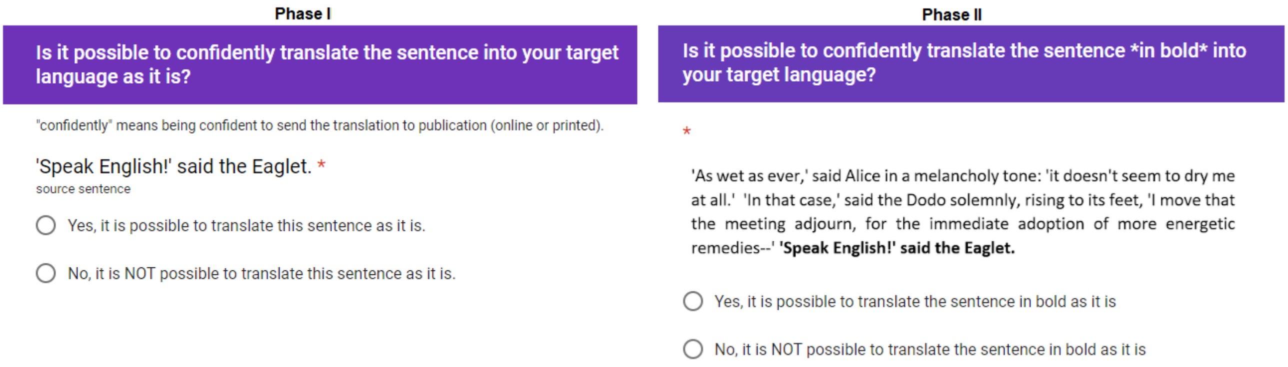
For **Phase I**, 100 *isolated sentences* were selected randomly from each of the three data sets described in Section 4.1. (300 sentences in total). Participants were shown a random English sentence and were asked: "Is it possible to confidently translate the sentence into your target language as it is?". The question was followed by a brief explanation as to what "confidently" meant: "*confidently* means being confident to send the translation to publication (online or printed)". When answering "yes", participants were taken to the next sentence, while when answering "no", a new window with the question "Why it is not possible to translate the sentence as it is?" showed up and participants saw a list of possible reasons they could select from:

1. Source problems
   There are problems with the English sentence that prevents me from understanding the meaning of it (typos, grammar, missing words, sentence ends abruptly, etc).
2. Unknown words
   There are terms/words/expressions I do not know in English and cannot understand with the given context.
3. No equivalent term
   There are terms/words/expressions with no equivalent in my target language.
4. Terminology
   It requires more context to be able to translate terminology appropriately.
5. Ambiguity
   It requires more context to be able to translate ambiguity appropriately.
6. Gender
   It requires more context to be able to determine gender appropriately.

[1] http://opus.nlpl.eu/

[2] https://sites.google.com/view/translation-survey/home

[3] Ethical approval for this survey has been obtained from the Dublin City University Research Ethics Committee.



Phase I

Is it possible to confidently translate the sentence into your target language as it is?

"confidently" means being confident to send the translation to publication (online or printed).

'Speak English!' said the Eaglet. *

source sentence

○ Yes, it is possible to translate this sentence as it is.

○ No, it is NOT possible to translate this sentence as it is.

Phase II

Is it possible to confidently translate the sentence *in bold* into your target language?

*

'As wet as ever,' said Alice in a melancholy tone: 'it doesn't seem to dry me at all.' 'In that case,' said the Dodo solemnly, rising to its feet, 'I move that the meeting adjourn, for the immediate adoption of more energetic remedies--' **'Speak English!' said the Eaglet.**

○ Yes, it is possible to translate the sentence in bold as it is

○ No, it is NOT possible to translate the sentence in bold as it is

Figure 1: Survey set-up. Phase I shows a single random sentence while Phase II shows the sentence in context.

7. Other

The selection of options is based on frequently reported context problems in the literature. The option "other" offered the participants the possibility to report other context issues.

Phase I was online from 26 March to 29 April 2019 and collected 192 responses in total.

From the 300 sentences, 107 sentences presented over 40% agreement on "no" answers (42 from reviews, 30 from subtitles and 25 from literature). These sentences were presented to the participants in Phase II together with the context consisting of two preceding sentences.

In **Phase II**, participants were given an English text consisting of 3 *consecutive sentences*, the third one being the one selected as problematic in Phase I, and were asked to answer the same questions as in Phase I. The only difference was that the question was formulated as "the sentence (in bold)" instead of "the given sentence". As in Phase I, for every "no" answer, a new window with a list of reasons appeared.

Phase II ran from 20 May till 24 June 2019. From the 107 selected sentences, 95 sentences were actually used – 12 sentences were discarded due to problems with the English source (i.e. sentences were incomplete), or not having a precedent sentence in the corpus. In total, 70 responses were collected in this phase.

5. Results and Discussion

5.1. Isolated sentences

As mentioned in Section 4.3., in total 300 source sentences from the three domains were judged in isolation in Phase I. Participants were asked to judge if the source sentence was possible to be translated as it was, and in case of a negative response, to identify why.

The main issues found to hinder the translation in these 300 sentences were:

- Ambiguity – 28.6%
- Gender – 20.9%
- Source problems – 16.6%
- Terminology – 15.5%
- Unknown words – 9.9%
- No equivalent term – 4.5%
- Other - 4.0% (most mentioned were case, register and number)

When comparing the issues for the different domains in Table 3, several differences can be noticed.

issues (%)	reviews	subtitles	book
Ambiguity	**26.3**	**30.0**	**32.0**
Gender	15.7	**22.6**	**30.1**
No equivalent term	5.7	3.3	3.0
Source problems	**21.5**	12.2	11.0
Terminology	18.4	13.8	11.0
Unknown words	11.0	7.6	10.0
Other	1.4	10.6	2.9

Table 3: Percentage of issues hindering translation of 300 English isolated sentences for each domain.

We can see that ambiguity is the most common issue found in all three domains. However, the distribution of other issues is different. The review domain has source problems as the second most common issue, which is far from being a surprise – reviews belong to the user-generated content genre known to contain poorly formed sentences in the source. For the literature domain, gender is the second frequent issue, being very close to ambiguity. The subtitles domain show the highest number of "other" issues, almost 11% of the sentences. As the most mentioned "other" issues relate to register, case and number, we hypothesise that the dialogue nature of the subtitles require more context to identify those.

From these single 300 sentences, we gathered those which participants agreed that they could not translate. Taking into account that 18 different target languages from distinct language families were involved, high agreements were hard to expect – as seen in 2, only 1-2% of the sentences show agreement over 80%. We decided to use the cut off of 40% agreement, hoping to get both the most reliable as well as the most heterogeneous set of issues. The selected sentences were then used for a second round of judgements, henceforth, Phase II, where they were displayed together with two preceding sentences (section below).

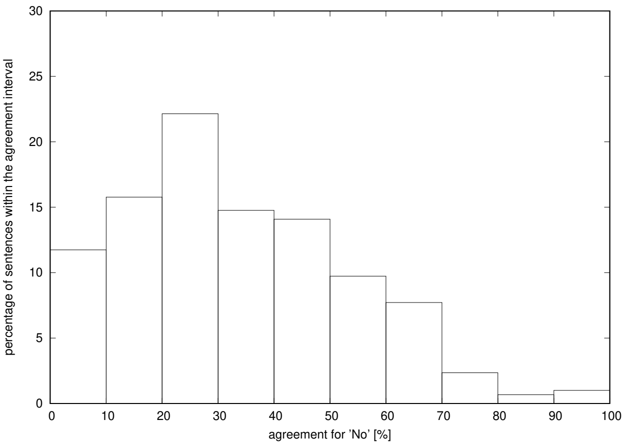


Figure 2: Percentage of isolated sentences within the agreement interval for the answer“No”.

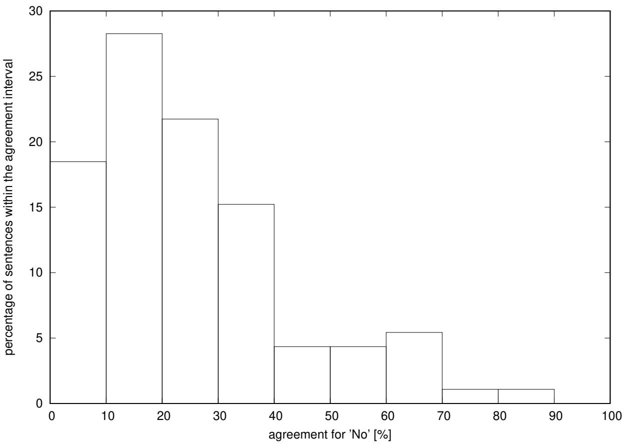


Figure 3: Percentage of agreement for “No” with two preceding sentences

5.2. Sentences in context of two preceding ones

From the 300 sentences presented in Phase I, 33% (107) presented an agreement of over 40% that they could not be translated in isolation. From those, 12 sentences were discarded because additional context was not expected to help: problems with the English source or being the first sentence in the corpus. The main issues found to hinder the translation in those 95 sentences were:

- Terminology – 19.9%
- Source problems – 18.6%
- Ambiguity – 18.6%
- Unknown words – 16.3%
- Gender – 13.1%
- No equivalent term – 10.8%
- Other – 2.8% (again, the most mentioned issues were case, register and number)

It can be noticed that the percentage of terminology and source problems increased whereas the percentage of gender, ambiguity and other problems decreased – a tendency towards more target language independent issues can be observed.

issues (%)	reviews	subtitles	literature
Ambiguity	12.7	**25.5**	**23.7**
Gender	7.8	9.8	18.1
No equivalent term	14.6	8.2	12.2
Source problems	**25.6**	8.2	10.4
Terminology	15.1	**27.8**	16.0
Unknown words	**24.2**	18.6	**19.2**
Other	0.0	2.0	0.4

Table 4: Percentage of issues found to mostly hinder translation of the 95 sentences in Phase II per domain.

When looking more closely into the three domains in Table 4, we notice that the review domain unsurprisingly shows that the most frequent issues are related to the source language, namely source problems and unknown source words. While ambiguity decreased for over 14% and gender for 9% compared to the values for single sentences, terminology remained within the same range. For the subtitle domain, terminology has an increase of 14%, gender decreases for 13%, while ambiguity remains in the same range (30-26 respectively). Gender also decreases for the literature domain (30% to 18%), as well as ambiguity (32% to 24%), while there is an increase for terminology (+5%), unknown words (Cannot understand) and No equivalent in target language. The other issues are still predominant in subtitles, although much less than for isolated sentences. All in all, for each domain, the overall tendency of a decrease in target language specific problems can be observed.

From these 95 sentences, 23% of them (22 sentences) were globally agreed (over 40% agreement) that even two preceding sentences were not enough to translate them accurately. Once more, the decision of the cut off of 40% is due to the distribution of agreement as shown in Figure 3.

In order to estimate issues requiring a larger context for a large number of target languages, we analysed those 22 sentences and the main issues were:

- Terminology – 23.9%
- Unknown words – 21.3%
- Source problems – 20.9%
- Ambiguity – 19.3%
- No equivalent term – 7.1%
- Gender – 7.0%
- Other – 0.6% (the most mentioned issues were again case, register and number)

It can be seen that the target language-independent issues are further increasing whereas the language-dependent ones are further decreasing.

When looking more closely into the three domains in Table 5, we notice that the differences between the review domain and the other two are not so prominent anymore. However, it is worth noticing that while source issues increase, terminology and gender decrease.

Little variation is found in the subtitle domain for ambiguity, source, unknown words and no equivalent types of issues, but terminology has a considerable increase, while gender has a considerable decrease. Gender also decreases in the literature domain, which is the only domain to have ‘other’ issue types remaining.

issue (%)	reviews	subtitles	literature
Ambiguity	9.8	27.1	**27.2**
Gender	3.4	5.8	11.3
No equivalent term	9.8	3.9	6.7
Source	**32.8**	7.1	17.2
Terminology	19.6	**37.4**	16.6
Unknown words	24.5	18.7	18.5
Other	0.0	0.0	2.6

Table 5: Percentage of issues found to mostly hinder translation of the resulting 22 sentences from judgements in Phase II per domain.

The described analysis involving three domains and a number of distinct language pairs has shown that the majority of the sentences can be translated if the two preceding sentences are available. The analysis also has shown that the remaining problems are mainly language-independent (problems with the source, ambiguity/terminology in English). However, it is still not clear what the necessary span is, and whether it is language dependent. Therefore, we conducted the third part of the experiment, namely a detailed qualitative analysis of necessary context for two distinct target languages, Serbian and Portuguese, which is described in the following section.

5.3. Analysis of context span for two target languages

In the final part of our experiment, we conducted a detailed analysis of a set of the sentences for two target languages: Portuguese (PT) and Serbian (SR). The choice of these languages is due to two facts: they belong to two different European language families and therefore cover different types of linguistic issues, and they are the researchers' mother tongues, meaning that the conducted analysis is reliable. From the starting 300 sentences from Phase I (100 for each domain), we selected those that could not be translated in isolation, and carried out a qualitative analysis of the minimum context span needed to solve the issues.
Table 6 shows the number of sentences analysed for each language. It is worth noticing that for Portuguese, as there were five respondents, we selected the sentences where they agreed over 40% that sentences could not be translated, while for Serbian there was only one respondent.

% of "no"	reviews	subtitles	book
PT	46.0	32.0	51.0
SR	56.0	60.0	42.0

Table 6: Isolated sentences not possible to translate into PT and SR.

From the 100 sentences for each domain, 56%, 60% and 42% could not be translated in isolation for Serbian, in the review, subtitle and literature domain respectively. The numbers for Portuguese are lower for the review and subtitles domain (46% and 32% respectively) but higher for the literature domain (51%) compared to the Serbian language. Table 7 shows that there are some clear differences between the two target languages. For Serbian, the predominant issues are gender as well as "other" which mainly contains case and number, two very important morpho-syntactic features of this language. Gender also represents a frequent issue for translation into Portuguese, although less frequent than for Serbian. As for ambiguity, the percentages are comparable.

SR- issues (%)	reviews	subtitles	literature
Ambiguity	19.7	19.6	11.2
Gender	**34.4**	**35.6**	**42.2**
No equivalent term	0	0	0
Source problems	6.6	3.4	2.3
Terminology	0	1.1	0
Unknown words	4.9	2.4	4.4
Other	**34.4**	**37.9**	**39.9**

PT- issues (%)	reviews	subtitles	literature
Ambiguity	19.4	16.8	**21.8**
Gender	**24.7**	**25.2**	**47.3**
No equivalent term	2.7	0	0.6
Source	11.8	**21.0**	2.4
Terminology	**36.0**	13.7	15.2
Unknown words	5.4	13.6	4.8
Other	0	9.7	7.9

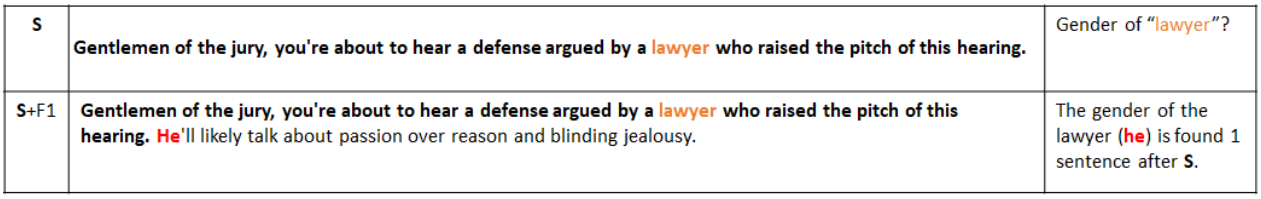
Table 7: Percentage of issues found to mostly hinder translation of isolated sentences into Serbian and into Portuguese.

EN	So I **ordered** a replacement
PT	Então **solicitei** a troca
SR	Pa sam **naručiO/LA** zamenu

Table 8: Example of different issues and context spans for Serbian and Portuguese.

Table 8 demonstrates one example when a different context span is needed for Portuguese and Serbian. The gender of the writer is not given in the source text, which is not problematic for Portuguese – the past participles do not need to agree with a gender, and so the sentence can be correctly translated in isolation. For Serbian, however, the gender of the writer is essential as you need that information to be able to translate "ordered" into the corresponding gender form (to choose between the suffixes "O" and "LA"). Nevertheless, this information cannot be found at all in the review – the only possibility is to ensure the gender is consistent throughout the review.
Regarding similarities between Serbian and Portuguese, figure 4 shows a sample where the gender of the noun "lawyer" is needed for both languages (PT=advogado/advogada; SR=advokat/advokatica). This information, which is also necessary for the past participle "raised" when translating into Serbian, is found in the first following sentence.

Qualitative analysis of context span The qualitative analysis of the necessary context span has shown that despite the described differences between the domains and languages, there is a number of common traits for both lan-



S	**Gentlemen of the jury, you're about to hear a defense argued by a** lawyer **who raised the pitch of this hearing.**	Gender of "lawyer"?
S+F1	**Gentlemen of the jury, you're about to hear a defense argued by a** lawyer **who raised the pitch of this hearing.** **He**'ll likely talk about passion over reason and blinding jealousy.	The gender of the lawyer (**he**) is found 1 sentence after **S**.

Figure 4: Example of the same issue and context span for Serbian and Portuguese.

guages and all domains:

- 30-60% of problems in isolated sentences can be resolved with one or two preceding sentences – however, that still leaves a large number of unresolved sentences;
- 5-15% sentences need up to 10 preceding sentences
- 10-20% of issues can be resolved only by global or visual context (such as character gender for movie subtitles and literature, speaker gender for TED talks, information about particular product or movie plot for terminology/ambiguity in reviews, etc.)
- 10-20% reviews would need the user gender, which is not available – in this case, a consistent use of gender in the target language should be required
- about 10% of literature sentences require character gender, which might not be explicitly available – again, the consistence is then needed
- certain sentences (5-25%) require up to 10 *following* sentences

Recommendations Taking into account all reported results, including qualitative analysis of the necessary context span for two target languages, a few points can be used to address language independent set ups. Therefore, our preliminary recommendations for a reliable translation evaluation are as follows:

- for reviews, news articles, and other relatively short texts (up to 15 sentences): show the whole document, including the title.
- for domains with longer texts, such as subtitles and literature, show at least 10 preceding and 10 following sentences;
- an even better solution would be to show shorter context (e.g. 5 preceding and 5 following sentences) but to enable a sliding window for cases which need more context
- make clear which type of text is being evaluated (subtitles from which movie, TED talk on which topic, reviews about movies, hotel or products (which products), etc.
- include a visual context if available (picture or description)

6. Conclusion and Future Work

This paper addresses the context span necessary for reliable human evaluation of MT. Regarding our research questions defined in Section 1., we have demonstrated that:

RQ1 For about 30-60% of sentences, across three domains and 18 target languages, two preceding sentences are indeed enough to resolve context-related issues. However, this still leaves a large number of sentences which require more context than this usual one. By analysing those sentences in depth for two target languages, Serbian and Portuguese, we found that the necessary context span can go much further into preceding sentences and also in the following sentences. For a number of sentences, broader (global) context is needed, too (such as title, gender of the speaker, etc.).

RQ2 Ambiguity is the most common issue across domains and languages, followed by gender and terminology. However, certain issues are highly target-language specific, such as case. In addition, some issues can be resolved only by providing a broader context, such as terminology for product reviews and gender for TED speakers, characters in a book or movie.

Despite certain limitations of our experiment, we believe that our findings represent a good base for reliable context-aware MT evaluation. Guidelines based on our findings are included, which mention the inclusion of whole context for texts shorter than 15 sentences, inclusion of titles and pictures if possible, as well as enabling a sliding window for cases where longer context is necessary.

In addition, our findings might represent a direction for further research on context-aware MT systems, such as how to treat different context-related issues, where and how to find the necessary information, etc.

The findings also open several directions for future work. First of all, more sentences should be annotated by a larger number of participants, preferably by professional translators. We plan to develop more strict guidelines for the participants regarding the options for problematic sentences (answer "no"), to include more options depending on the target language, and also to include other source languages.

7. Acknowledgements

We would like to thank all the participants for providing us with their responses. 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.

8. Bibliographical References

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguis tics: Human Language Technologies (NAACL-HLT 2018)*, pages 1304–1313, New Orleans, Louisiana, June.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

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

Müller, M., Rios Gonzales, A., Voita, E., and Sennrich, R. (2018). A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 61–72, Belgium, Brussels, October.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, September. Association for Computational Linguistics.

Tiedemann, J. (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, May. European Languages Resources Association (ELRA).

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT*, pages 113–123, Brussels, Belgium.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, e llipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy, August.