Document-Level Machine Translation Evaluation Project: Methodology, Effort and Inter-Annotator Agreement

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

Abstract

Recently, document-level (doc-level) human evaluation of machine translation (MT) has raised interest in the community after a few attempts have disproved claims of “human parity” (Toral et al., 2018; Läubli et al., 2018). However, little is still known about best practices regarding doc-level human evaluation. This project aims to identify methodologies to better cope with i) the current state-of-the-art (SOTA) human metrics, ii) a possible complexity when assigning a single score to a text consisted of ‘good’ and ‘bad’ sentences, iii) a possible tiredness bias in doc-level set-ups, and iv) the difference in inter-annotator agreement (IAA) between sentence and doc-level set-ups.

1 Introduction

Although currently an active community is working on developing document-level (doc-level) MT systems, their evaluation has primarily been performed at the sentence level. In 2019, for the first time, WMT19 attempted a doc-level human evaluation for the news domain, after considering criticisms by Toral et al. (2018) and Läubli et al. (2018) regarding the current best practices in MT evaluation. Both papers independently reassessed the claims of MT “achieving human parity” and found that the lack of extra-sentential context has a great effect on quality assessment.

In a recent survey with native speakers, Castilho et al. (2020) tested the context span for the translation of three different domains (reviews, subtitles, and literature). Results show that over 33% of the sentences tested (300 in total) required more content than the sentence itself to be translated, and from those, 23% required more than two previous sentences to be properly translated. Some of the issues which the participants found to most hinder the translation include word ambiguity, terminology, and gender agreement. Moreover, the authors found that there are differences in issues and context span between domains. This shows that doc-level evaluation enables to assess suprasentential context, textual cohesion and coherence types of errors.

In one of the few studies on doc-level evaluation, Läubli et al. (2018) use pairwise rankings of fluency and adequacy in which raters give one single score to the full document. For WMT19, the direct assessment task asked crowdworkers to give a single score (0–100) to full documents for accuracy, where only one MT output is shown each time (no comparison with other MT system).

With that in mind, this project aims at identifying methodologies to better cope with the SOTA human metrics, namely ratings of fluency and adequacy, error mark up and ranking evaluations (Castilho et al., 2018). We will gauge the complexity when assigning a single score to full texts, since they can consist of ‘good’ and ‘bad’ sentences, which could mean that instead of a single score, translators would prefer to give scores to different chunks of the texts while seeing the whole text. We will investigate the difference in IAA between sentence and document level set-ups. Furthermore, a possible tiredness bias in doc-level set ups will also be investigated, for example, the extend to which translators judge a long text on the quality of its first sentences. For that end, we will run a series of experiments with the WMT newstest2019, with professional translators.

2 Methodology

The evaluation setup is made up of two sequential stages: (1) Fluency/adequacy and error markup, and (2) Pairwise ranking. Four professional translators will carry out the tasks in two scenarios: (A) evaluation at the sentence level, showing randomised sentences, one at a time, and (B) evaluation at a document level. While Scenario A will be the baseline as it follows common practice in MT evaluation, Scenario B will show how translators will make the decisions and what influences them when they have to give one score for full texts. After each task, translators will answer a post-task questionnaire about the tasks. The documents and scenarios are randomised to avoid participants evaluating the same source twice. Table 1 shows how documents and sentences are randomised by participant.

Documents (groups)	P1	P2	P3	P4
A (1–500 sentences)	S_1	S_2	D_1	D_2
B (501–1000 sentences)	D_2	D_1	S_2	S_1

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

The corpus used is the WMT *newstest2019* English corpora, which has an average document length of 17 sentences (minimum 4 sentences, maximum 30 sentences). Full documents that amount to 1000 sentences are selected, totalling 64 documents. The English documents are translated into Brazilian Portuguese with Google Translate for stage 1, and with both Google Translate and DeepL for stage 2.

The choice of language is because as it is the principal researcher's mother tongue, this will make it possible to analyse it more carefully and see possible patterns in the process. Moreover, being Portuguese a Romance language, it is possible that the results of this pilot can be extended to the language family.

The tasks are set in two tools. For fluency, adequacy, and error mark up, PET tool (Aziz et al., 2012) is used as it allows time tracking. For the ranking tasks, an online spreadsheet is used and extension to track time is also implemented. Translators are also requested to keep track of their time while performing the evaluation.

After stages 1 and 2 are complete, a second round of evaluation will designed. This time, doc-level evaluation will be performed with translators giving one score per chunks/sentence in the text while having access to the full document. That way we will be able to compare effort and IAA between the two methodologies for doc-level.

3 Final Remarks

This project aims at shedding light at methodology, effort and IAA and systematically improve human evaluation of MT at the doc-level. Preliminary results will be available by June 2020, and data sets will be fully available at the end of the project.

Acknowledgements This project is funded by the EAMT under its programme "2019 Sponsorship of Activities". The ADAPT Centre for Digital Content Technology 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.

Thank you all the translators for providing us with annotations and feedback.

References

Aziz, Wilker, 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, may.

Castilho, Sheila, 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, July.

Castilho, Sheila, 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, may.

Läubli, Samuel, 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.

Toral, Antonio, 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.