DELA Project: Document-level Machine Translation Evaluation

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

Abstract

This paper presents the results of the Document-level Machine Translation Evaluation (DELA) Project, a two-year project which started in September 2020 funded by the Irish Research Council. This paper describes the results of the project to date, as well as its latest developments.

1 Introduction

The challenge of evaluating translations in context has been raising interest in the machine translation (MT) field. However, the definition of what constitutes a document-level (doc-level) MT evaluation, in terms of how much of the text needs to be shown, is still unclear (Castilho et al., 2020). Few works have taken into account doc-level human evaluation (Barrault et al., 2020), and one common practice is the usage of test suites with context-aware markers. However, test suites with document-level boundaries are still scarce (Rysová et al., 2019). The main objective of the DELA Project is to define best practices for doc-level MT evaluation, and test the existing human and automatic sentence-level evaluation metrics to the doc-level. We present here the results from the project to date, as well as the upcoming research to be carried out.

2 Context Span for MT

In Castilho et al. (2020), we tested the context span, that is, the length of context necessary, for the translation of 300 sentences in three different domains (reviews, subtitles, and literature) and showed that over 33% of the sentences tested required 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. Ambiguity, terminology, and gender agreement were the most common issues to hinder translation, and moreover, there were observable differences in issues and context span between domains.

3 Doc-Level Evaluation methodology

In Castilho (2020; 2021), we tested the differences in inter-annotator agreement (IAA) between single-sentence and doc-level setups. First, translators evaluated the MT output in terms of fluency, adequacy, ranking and error annotation in: (i) one score per single isolated sentence, and (ii) one score per document. Then, the doc-level setup was modified, and translators evaluated (i) random single sentences, (ii) individual sentences with access to the full source and MT output, and (iii) full documents. Results showed that assessing individual sentences within the context of a document yields a higher IAA compared to the random single-sentence methodology, while when translators give one score per document, IAA is much lower. Assigning one score per sentence in context avoids misevaluation cases, extremely common in the random sentences-based evaluation setups.[1] The higher IAA agreement in the random single-sentence setup is because raters tend to accept the translation when adequacy is ambiguous but the translation is correct, especially if it is fluent.

[1] Without context, the sentence *'I am satisfied'* translated into Portuguese in the masculine *'Eu estou satisfeito'* will get a perfect score even when the gender of the pronoun *I* is feminine (*'satisfeitA'*).

4 DELA Corpus

Using the issues found in Castilho et al. (2020), we developed the DELA corpus, a doc-level corpus annotated with context-aware issues when translating from English into Portuguese, namely gender, number, ellipsis, reference, lexical ambiguity, and terminology (Castilho et al., 2021). The corpus contains 60 full documents and was compiled with six different domains: subtitles, literary, news, reviews, medical, and legislation; and can be used as a challenge test set, training/testing corpus for MT and quality estimation, and for deep linguistic analysis of context issues.[2]

5 Examining Context-Related Issues

Using the DELA Corpus, we examine the shortest context span necessary to solve the issues annotated in the corpus, and categorise the types of contexts according to their position, and report the i) Context Position, and ii) Context Length We find that the shortest context span might appear in different positions in the document including preceding, following, global, world knowledge. The average length depends on the issue types as well as the domain. The results show that the standard approach of relying on only two preceding sentences as context might not be enough depending on the domain and issue types.

6 Latest Developments

The DELA Project, running until September 2022, will focus now on the human and automatic evaluation metrics for MT, testing and developing new ways to use them for doc-level evaluation.

Doc-level human and automatic evaluation metrics: The focus of the DELA Project is to answer the following research questions: i) Are the state-of-the-art (SOTA) human and automatic evaluation metrics able to capture the quality level of the doc-level systems realistically?; and ii) Can/should they be modified or do new ones are needed?

A series of experiments with the SOTA human metrics are being carried out, informed by the best methodologies found in previous results. With that, we will determine whether these metrics can be used in doc-level evaluations, or if new metrics should (and could) be developed. The doc-level human evaluation will inform automatic metrics to be used for document-level systems.

Doc-level evaluation tool: The DELA project will gather specification from translators to design a translation evaluation tool which will provide an environment to assess MT quality at a doc-level with human and automatic evaluation metrics scores specified as best suited for doc-level evaluation in the project. The tool will be made freely available.

Acknowledgements: This project is funded by the Irish Research Council (GOIPD/2020/69). ADAPT, the Science Foundation Ireland Research Centre for AI-Driven Digital Content Technology at Dublin City University, is funded by the Science Foundation Ireland through the SFI Research Centres Programme (Grant 13/RC/2106_P2).

References

Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, and et al. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Castilho, Sheila, Maja Popović, and Andy Way. 2020. On Context Span Needed for MT Evaluation. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, page 3735–3742, Marseille, France, May.

Castilho, Sheila, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. Dela corpus - a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 571–582. Association for Computational Linguistics.

Castilho, Sheila. 2020. On the same page? comparing IAA in sentence and document level human mt evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159. Association for Computational Linguistics, November.

Castilho, Sheila. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems*, pages 34–45. Association for Computational Linguistics, April.

Rysová, Kateřina, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy, August. Association for Computational Linguistics.

[2] The corpus and annotation guides can be found at: `https://github.com/SheilaCastilho/DELA-Project`