Editors' foreword

Castilho, Sheila; Knowles, Rebecca

Publisher's version / Version de l'éditeur:

https://doi.org/10.1017/nlp.2024.6

Natural Language Processing, 31, 4, pp. 983-985, 2025-06-23

Natural Language Processing (2025), **31**, pp. 983–985
doi:10.1017/nlp.2024.6

CAMBRIDGE UNIVERSITY PRESS

EDITORIAL NOTE

Editors' foreword

Sheila Castilho[1] and Rebecca Knowles[2]

[1]School of Applied Language & Intercultural Studies/ADAPT Center, Dublin City University, Dublin, Ireland and [2]National Research Council of Canada, Ottawa, ON, Canada
Corresponding author: Sheila Castilho; Email: sheila.castilho@dcu.ie

(Received 29 February 2024; accepted 29 February 2024)

Special Issue on '**The Role of Context in Neural Machine Translation Systems and its Evaluation**', guest-edited by Sheila Castilho and Rebecca Knowles

Abstract

This special issue focuses on the many roles that context plays in neural machine translation and the evaluation of machine translation. It includes a survey of the current state of research related to context in machine translation and machine translation evaluation, along with three papers that focus on a variety of topics related to evaluation. Together, these highlight the wide range of topics in this area, and its increasing relevance as high-quality machine translation becomes increasingly common in a variety of settings.

Keywords: machine translation; evaluation

In recent decades, the field of machine translation (MT) has undergone a major shift in approaches—from phrase-based statistical machine translation to neural machine translation—and a corresponding remarkable improvement in translation quality. Since the era of statistical MT systems and moving into the era of neural models, the role of MT in industry has grown, and research aimed at enhancing the models has only intensified. Recently, there have been increasing claims that neural machine translation (NMT) systems are reaching human parity (Hassan et al., 2018, i.a), followed by subsequent analyses that incorporate contextual information and show that there remains a gap between machine translation and human translation performance (Toral et al., 2018; Läubli et al., 2018). This has resulted in calls for changes to evaluation to distinguish high-performing sentence-level machine translation from human translation, as well as for improved approaches to incorporating context into machine translation systems and automatic evaluation metrics.

This special issue aims at tackling a variety of different questions about the roles that context in NMT and in its evaluation. The three research papers published here focus on components of the evaluation process. They each take different perspectives and examine different aspects of what it means to consider "context" in NMT. This highlights the broad range of work still to be undertaken on this topic, which is also summarized in the survey article (Castilho and Knowles, 2024). The survey article not only covers past and recent work in the field but also highlights the emergence of large language models and their applications in translation and evaluation. Furthermore, the authors provide perspectives on the future of the field.

Motivated by analyses of reference translations periodically turning up reference translations that are not of as high quality as expected, Zouhar et al. (2024) propose a consensus-based

approach to building high-quality document-level translations, which they call "optimal reference translations," along with manual evaluation of those translations in order to verify their quality. They examine annotator agreement, finding that meaning, style, and pragmatics were most influential in overall score, which also matched annotator questionnaire responses about the importance of these factors. Additionally, they consider annotator differences, including those between annotators with differing levels of translation expertise. As the field shifts towards document-level translation and evaluation and as NMT performance continues to improve, having high-quality document-level references will be vital for appropriate use of automatic reference-based metrics and human evaluation.

With the increase in the fluency of MT output brought about by the shift to NMT, the focus of do Campo Bayón and Sánchez-Gijón (2024) is on evaluating the naturalness and user acceptance of NMT output in a low-resource language (Galician) in the social media domain. They propose evaluating NMT using the non-inferiority principle, which is more commonly used in the realm of health and medicine. This type of evaluation for naturalness can serve as a complement to adequacy-based evaluations. Their work touches on context in two important ways. First, they highlight the importance of providing annotators with additional textual context (in this case tweet threads as opposed to tweets on their own) in the annotation process. Second, and more broadly, they consider how the genre, real-world context, and annotator characteristics may interact, such as how different age demographics may have different perceptions and expectations for social media translation. As NMT is used in a wider variety of everyday contexts, it may be increasingly important to design evaluations that take into consideration aspects specific to the context in which the NMT is being used and the interplay between that and users' expectations.

Knowles and Lo (2024) examine data from recent WMT (Conference on Machine Translation) shared tasks in order to explore the impacts of incorporating intersentential context into human evaluation of MT. They examine inter- and intra-annotator variation and discuss some best practices for balancing the challenges of handling document-level intersentential context in human evaluation, such as by using calibration sets to pre-screen annotators or standardize annotator scores. They also highlight the need for future work on better understanding how annotators interpret annotation tasks and how the shift towards document-level translation and evaluation necessitates a reexamination of assumptions about human evaluation protocols.

These papers cover a wide range of disparate context-related topics in NMT and also illustrate how interconnected many of these aspects are in the broader conversation about context. Zouhar et al. (2024) and do Campo Bayón and Sánchez-Gijón (2024) both consider differences between annotators (the former in terms of qualifications related to translation, the latter demographically) and how these influence annotation. Knowles and Lo (2024) used anonymized data from WMT, so did not have access to such information, but examined patterns of behaviour across annotators. For Knowles and Lo (2024), this included certain annotators who strongly preferred to give scores that lined up with the tick marks on a sliding scale annotation tool, comparable to the observation in Zouhar et al. (2024) that annotators tended towards scores of round numbers (typing scores in a spreadsheet as an annotation tool). The observation in Knowles and Lo (2024) that low-scoring segments observed in evaluations that include intersentential context have an influence on the scores given to other segments may relate to the analysis in Zouhar et al. (2024) that found that annotators focused on the lowest-rated segments when asked to produce document-level scores; there is clearly more to examine in this area. Both do Campo Bayón and Sánchez-Gijón (2024) and Knowles and Lo (2024) consider the issue of whether the task the annotators are performing is sufficiently well-defined, with do Campo Bayón and Sánchez-Gijón (2024) iterating on their survey questions and providing annotators with more explanations in order to better capture the desired information. Zouhar et al. (2024), do Campo Bayón and Sánchez-Gijón (2024), and Knowles and Lo (2024) all consider the effects of having access to additional within-document context while doing evaluation, which ties into the questions considered in prior work about how much intersentential context is necessary in order to do well-informed evaluation (Castilho et al.,

https://doi.org/10.1017/nlp.2024.6 Published online by Cambridge University Press

2020; Castilho, 2022, i.a). In turn, this question of which intersentential context is helpful is discussed in the survey (Castilho and Knowles, 2024), from the perspective of how it can impact the quality of translation output as well as how it can be involved in evaluation.

The many questions of context in NMT and NMT evaluation are deeply interconnected and remain an exciting area for future examination, especially as we see the rapid introduction of models such as large language models with extended intersentential context capabilities. The fact that perspectives on evaluation were the main focus of the special issue's submissions highlights the importance of appropriate evaluation in the development of the field of context-aware MT, as well as the fact that the field is still in the process of developing best practices for evaluation. Ensuring that we have sound and appropriate methods of evaluation will allow us to better explore their capabilities and in turn allow the field to make the most of a wide range of ways that many types of context may be useful for improving MT quality.

We hope that the insights shared in this special issue will prove valuable to our readers and that it will not only ignite further curiosity but also serve as a practical guide for the community, fostering a deeper understanding and robust exploration of the diverse ways in which context can enhance MT.

References

Castilho S. (2022). *How much context span is enough? examining context-related issues for document-level MT*. In Calzolari N., Béchet F., Blache P., Choukri K., Cieri C., Declerck T., Goggi S., Isahara H., Maegaard B., Mariani J., Mazo H., Odijk J. and Piperidis S. (eds), Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, European Language Resources Association, pp. 3017–3025.

Castilho S. and Knowles R. (2024). A survey of context in neural machine translation and its evaluation. *Journal of Natural Language Processing*.

Castilho S., **Popović M.**, **Way A.**, **Béchet N.**, **Blache F.**, **Choukri P.**, **Cieri K.**, **Declerck C.**, **Goggi T.**, **Isahara S.**, **Maegaard H.**, **Mariani B.**, **Mazo J.**, **Moreno H. and** A. (2020). *On context span needed for machine translation evaluation*In Odijk J. and Piperidis S. (eds), Proceedings of the Twelfth Language Resources and Evaluation Conference, Calzolari, Marseille, France. European Language Resources Association, pp. 3735–3742.

do Campo Bayón M. and Sánchez-Gijón P. (2024). Evaluating NMT using the non-inferiority principle. *Journal of Natural Language Processing*.

Hassan H., **Aue A.**, **Chen C.**, **Chowdhary V.**, **Clark J.**, **Federmann C.**, **Huang X.**, **Junczys-Dowmunt M.**, **Lewis W.**, **Li M.**, **Liu S.**, **Liu T.**, **Luo R.**, **Menezes A.**, **Qin T.**, **Seide F.**, **Tan X.**, **Tian F.**, **Wu L.**, **Wu S.**, **Xia Y.**, **Zhang D.**, **Zhang Z. and Zhou M.** (2018). Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Knowles R. and Lo C.-k. (2024). Calibration and context in human evaluation of machine translation. *Journal of Natural Language Processing*.

Läubli S., **Sennrich R. and Volk M.** (2018). *Has machine translation achieved human parity? A case for document-level evaluation*. In Riloff E., Chiang D., Hockenmaier J. and Tsujii J. (eds), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Association for Computational Linguistics, pp. 4791–4796.

Toral A., **Castilho S.**, **Hu K. and Way A.** (2018). *Attaining the unattainable? reassessing claims of human parity in neural machine translation*. In Bojar O., Chatterjee R., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Yepes A. J., Koehn P., Monz C., Negri M., Névéol A., Neves M., Post M., Specia L., Turchi M. and Verspoor K. (eds), Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium. Association for Computational Linguistics, pp. 113–123.

Zouhar V., **Kloudová V.**, **Popel M. and Bojar O.** (2024). Evaluating optimal reference translations. *Journal of Natural Language Processing*.

Cite this article: Castilho S and Knowles R (2025). Editors' foreword. *Natural Language Processing* **31**, 983–985. https://doi.org/10.1017/nlp.2024.6