

Article Reviews re. Opportunities and Limitations in Machine Translation Evaluation for Neural Machine Translation

Tom Keane

June 7, 2021

I have elected to review the following two articles with regards to Opportunities and Limitations in Machine Translation Evaluation for Neural Machine Translation

1. - Indices of cognitive effort in machine translation post-editing [7]
2. - Eye tracking as an MT evaluation technique [2]

The reviews will consist of a brief overview of the motivations, goals, the methodologies, results expressed in the articles, separately.

The findings and conclusions of both articles will then be discussed in relation to each other, and in relation to Opportunities and Limitations in MT Evaluation for NMT.

1 Summaries

1.1 Indices of cognitive effort in machine translation post-editing

The purpose expressed in this article is finding methods to assess the cognitive difficulty in the post-editing (PE) process of machine translations (MT), based on three textual variables and various individual factors relevant to the post-editor:

These were the source text (ST), the raw MT output, and the emerging target text (TT), as textual features, and individual features such as working

memory capacity (WMC) and source language (SL) proficiency. The features that were selected were based on Kittredge [3], which established the three textual features relevant to PE used in this study.

The Utility of the study is stated as that PE can increase productivity, improve translation quality, therefore establishing reliable estimations of MT quality and PE effort are of interest in the field, and are problems worth solving. Working Memory Capacity’s relevance as an individual’s post editing ability is considered in Lehmann et al. [4].

This study found that according three mixed effect models of: Meteor (a semantics-based automatic evaluation metric)[1], Sentence length, prepositional phrases and PC3 (a principal component model based on various established Linguistic and psycholinguistic features). Meteor was found to be highly statistically significant in three models, which implies that the textual features outlined above can be used to assess PE difficulty, which can be of great use when estimating MT quality and PE pricing schemes.

It was also found that more effort was required to process sentences of ST by those with low SL proficiency. This article suggests that this may allow PE to be done by those with low SL proficiency [7](5.1).

This claim is bolstered by Mitchell, Roturier, and O’Brien [6], which shows that blind PE can lead to promising gains in MT fluency, regardless of the professional status of the post-editors.

It was also shown that WMC is positively linked to PE productivity. This is a similar result to McCutchen [5], which found that WMC has a positive impact on monolingual reviewing.

1.2 Eye tracking as an MT evaluation technique

This paper details a preliminary study regarding the use of eye tracking as a method to evaluate the output from a machine translation. The primary research questions are established in the introduction.

”To what extent does eye tracking data reflect the quality of MT output as rated by human evaluators? And, related to this question, could eye tracking potentially be used as a tool for semi-automatically measuring MT quality?” [2] (introduction)

It is also stated that this research may give insights into unobtrusive recording of MT reading effort, which could "supplement or confirm automatic evaluation metrics." [2](introduction)

A number of MT-outputs were pre-rated as "good" or "bad" on a scale of 1 - 4 by human evaluators. This study found that on average, 45% more gaze-time was spent looking at bad sentences than good sentences and that sentences with higher than average fixation counts were more likely to be bad.

2 Comparative

Indices of cognitive effort in machine translation post-editing[7] had quite a broad range of objectives as stated above. It was shown that based on the variables outlined above in the summary, it was found that based on eye tracking data that PE cognitive difficulty may be more accurately assessed. Based on the number of references cited, and the length and detail of the background and procedure, it is clear that a lot of research went into this paper. It is very information heavy, however and the results and discussion is difficult to parse. a greater degree of **cognitive effort** was required to understand these sections, compared to Doherty, Brien, and Carl [2]'s paper. Eye tracking as an MT evaluation technique[2] is a shorter paper, and much more readable for it. It is quite easy to understand the results and discussion and how it relates to the conclusion and findings. This may be in part due to a lower complexity in the stated objective of the paper, which is not necessarily a bad thing.

References

- [1] Michael Denkowski and Alon Lavie. "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 85–91. URL: <https://www.aclweb.org/anthology/W11-2107>.

- [2] Stephen Doherty, Sharon Brien, and Michael Carl. “Eye tracking as an Automatic MT Evaluation Technique”. In: *Machine Translation* 24 (Mar. 2010), pp. 1–13. DOI: 10.1007/s10590-010-9070-9.
- [3] R. Kittredge. “Kring, Hans P. (2001): Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes (Geoffrey S. Koby, ed.), The Kent State University Press, Kent, Ohio & London, 558 p.” In: *Meta: Translators’ Journal* 47 (2002), pp. 435–436.
- [4] Sabine Lehmann et al. “Applying CNL Authoring Support to Improve Machine Translation of Forum Data”. In: *Controlled Natural Language*. Ed. by Tobias Kuhn and Norbert E. Fuchs. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–10. ISBN: 978-3-642-32612-7.
- [5] Deborah McCutchen. “A Capacity Theory of Writing: Working Memory in Composition”. In: *Educational Psychology Review* 8.3 (1996), pp. 299–325. ISSN: 1040726X, 1573336X. URL: <http://www.jstor.org/stable/23359419>.
- [6] L Mitchell, J Roturier, and S O’Brien. “Mitchell L, Roturier J, O’Brien S (2013) Community-based post-editing of machine-translated content: monolingual vs. bilingual. In: O’Brien S, Simard M, Specia L (eds) Proceedings of machine translationsummit XIV Workshop on Post-Editing Technology and Practice (WPTP2), 2 Sept 2013, Nice, pp35–43”. In.
- [7] Lucas Nunes Vieira. “Indices of cognitive effort in machine translation post-editing”. In: *Machine Translation* 28.3/4 (2014), pp. 187–216. ISSN: 09226567, 15730573. URL: <http://www.jstor.org/stable/44113775>.