MT-Pese: Machine Translation and Post-Editese

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

Natália Resende
ADAPT Centre
School of Computing
Dublin City University
natalia.resende@adaptcentre.ie

Abstract

This paper introduces the MT-Pese project which is an umbrella name for a series of experiment venues that started in 2019. The project aims at researching the post-editese phenomena in machine-translated texts. We describe a range of experiments performed in order to gauge the effect of post-editese in different domains, back-translation, and quality.

1 Translationese and Post-editese

A number of studies (Volansky et al., 2013) have shown evidence of the so-called translationese phenomena (Gellerstam, 1986), that is, statistical differences between translated texts and non-translated texts. Recently, post-editing (PE) of machine-translated (MT) texts has secured its space in the translation workflow for a variety of domains, and consequently, the research interest for the typical features of human-translated texts has shifted for the typical features of post-edited texts. However, results of studies searching for typical features of post-edited texts - what has been called "post-editese - have presented mixed results, that is, while some studies found evidence for the existence of post-editese (e.g. Toral, 2019; Castilho et al. 2019), other studies did not find evidence of the phenomena (e.g. Daems et al. 2017).

The aim of the MT-Pese project is to investigate the post-editese phenomena on MT PE texts, using the rationale behind the translationese features as proposed by Baker (1996): *simplification*, *explicitation, normalisation* (or *conservatism*) and *levelling out* (or *convergence*). We define post-editese as the difference between the characteristics of human-translated texts (HT) and the PE versions, in relation to the raw MT output. MT-Pese has researched what influences the features of post-editese in two different textual domains, namely, news and literature (Castilho et al., 2019). We found that the literature domain contained more post-editese features. In a further study, we looked into the post-editese features in two different genres within the literature domain (Castilho and Resende, 2022). Currently, the project is focused on investigating the features of Post-editese on backtranslations (BT), with the aim to identify, for instance, if BT of PE versions would still carry strong post-editese features. Finally, the project also aims at addressing the question of whether the features of post-editese could be related to MT quality (section 4).

2 What influences the features of post-editese? A preliminary study

This study (Castilho et al 2019) investigated the presence of post-editese in a corpus composed by HT, MT and PE texts post-edited by either professional translators or student translators in two domains: news and literature. We also tested whether the PE level (light PE vs. full PE). Results showed evidence of post-editese features manifested as PE texts closer to the source texts and raw MT output rather than HT texts, and that the translators' experience as well as the text domains influence the magnitude of the post-editese features

3 Post-Editese in Literary Translations

This study (Castilho and Resende 2022) investigated the existence of post-editese features in a literary corpus composed of two different genres:

Alice's Adventures in Wonderland (AW) and The Girl on the Train (TGOTT), which were post-edited by nine professional translators. Results show a clear difference between the literary genres: while literary texts whose author's style is full of figurative language pose a harder challenge to the MT system, texts that emphasise action over language style are less challenging. We validate this assumption based on our observations that AW involved more edits than the TGOTT test set, suggesting that the MT output is capable of expressing the meaning of the source text more efficiently than for the AW. Moreover, we find a more visible pattern in terms of features for the TGOTT test set when compared to the AW which, in turn, is unstable in terms of pattern manifestation. This allowed us to confirm our post-editese hypothesis for almost all features in the TGOTT but for none in the AW.

4 Post-Editese in Backtranslations

This ongoing study aims at researching whether the post-editese features remain on backtranslated texts. To this end, we backtranslated the previous PE versions of the TGOTT and AW texts using an MT system, and extracted the same features examined in the previous studies in order to address the following questions:

a) Are the post-editese features reported in Castilho & Resende (2022) preserved in the BT texts?

b) How are post-editese features manifested in BT? Are BT features closer to the PEs or to the source texts?

The results will shed a light on whether BT from post-edited versions show more features from human involvement, and if so, whether that means PE-BTs have a higher quality. This will help the MT field, especially in regards to data augmentation.

5 Post-editese and Translation Quality

Finally, MT-Pese will look into whether post-editese features can be correlated with translation quality and creativity. For that, a few main research questions have been designed:

a) Which post-editese features are correlated to high quality post-edited texts?

b) Are there any features that can be correlated with naturalness?

c) Are there any features that can be correlated with creativity?

The results of this study will shed light on whether post-editese features mean that the PE version are of higher quality when compared to the raw MT output. If so, these features could be used to develop new evaluation metrics.

Acknowledgement

Both authors contributed equally to this work. This research was conducted with the financial support of the innovation programme under the Marie Skłodowska-Curie grant agreement No 843455 and the Irish Research Council (GOIPD/2020/69). Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University.

References

Baker, Mona. 1996. Chapter corpus-based translation studies: The challenges that lie ahead. In Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager. Amsterdam: John Benjamins Publishing Company, page 175-186.

Castilho, Sheila, and Natália Resende. 2022. "Post-Editese in Literary Translations" Information 13, no. 2: 66. Online. https://doi.org/10.3390/info13020066

Castilho, Sheila, Natalia Resende, Ruslan Mitkov. 2019. What Influences Post-editese features? A preliminary study. Proceedings of the second workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019). 5-6 September, 2019, Varna, Bulgaria, pages 19-27.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia New Series - Themes in Translation Studies 16:89–103.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Wollin, L. and Lindquist, H. Transla-tion Studies in Scandinavia. CWK Gleerup, Lund, volume 4, pages 88–95.

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In Proceedings of Machine Translation Summit. Dublin, Ireland.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. Digital Scholarship in the Hu-manities 30 (1):98–118.https://doi.org/10.1093/llc/fqt031.