With or without post-editing processes? Evidence for a gap in machine translation evaluation

Caroline Rossi
Univ. Grenoble Alpes (ICEA4)
France
Caroline.Rossi@univ-grenoble-alpes.fr

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes (LIG)
France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Abstract

Machine translation evaluation (MTE) is performed differently and with different goals in academia and industry (Drugan 2013, in Castilho et al. 2018 : 11). However, with the current integration of neural machine translation into human translation workflows, reliable measures of the amount of effort needed to post-edit machine translation (PEMT) outputs have become a common goal for researchers, language service providers and machine translation vendors (ibid., p. 29). Translation process research has developed tools to gather and analyse empirical data, but while a variety of measures have proved useful and reliable to measure PEMT effort (see e.g. Vieira 2016 : 42), translation processes are seldom considered when assessing the relevance of a given MTPE scenario.

Against this background, our study seeks to determine the impact of including MTPE in the evaluation process. We selected two of the most commonly used scales for the "declarative evaluation" of MT (Humphreys et al. 1991, in Way 2018b : 164): adequacy and fluency ratings. Based on two distinct experimental conditions, we then compared the ratings produced without performing PE and those produced immediately after a light PE process.

Data was collected with a group of 14 trainee translators, using two different text types and two different tools. A first series of assessments was conducted with KantanMT's language quality review system (LQR), which allows for a simple comparative evaluation of two systems without post-editing the outputs. The second series was done a few weeks later, in Post-Editing Tool (PET, Aziz et al. 2012). Each experimental condition includes two source texts from two different domains (environmental discourse and patents). We generated usable SMT and NMT outputs using eTranslation with environmental texts and WIPO translate with patent extracts. In both conditions, the students were given a realistic scenario -- i.e. they performed the evaluation, with a view to determining whether the MT output was relevant to a particular order.

Interrater reliability was assessed for each segment in each text (N=55) using Fleiss' kappa for adequacy and fluency scores, and an intraclass correlation coefficient (Vieira 2016 : 52) for temporal measures. While the reliability of the measures collected without PE was low, the measures collected in PET were for the most part homogeneous. Thus, evaluation was more reliable when performed with PE than without. Similarly, and even though there was more variation in temporal measures, homogeneity was stronger in PET data, suggesting that the activity was performed in a similar way across trainee translators.

We finally sought to determine what went wrong by performing qualitative analyses of the problematic segments, as evidenced by both kappa and intraclass correlation coefficients. Overall, our results suggest that it is very difficult, at least for trainee translators, to assess MT without PE. Specific training combining MTPE and evaluation might be particularly helpful to prepare them for a changing industry.

References

Aziz W, Sousa SCM, Specia L (2012). PET: a tool for post-editing and assessing machine translation. In: Calzolari N, Choukri K, Declerck T, Dogan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) *Proceedings of the eighth international conference on language resources and evaluation*, Istanbul, pp 3982–3987.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In *Translation Quality Assessment* (pp. 9-38). Springer, Cham.

Vieira, L. N. (2016). How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation*, *30*(1-2), 41-62.

Way A (2018a) Machine translation: where are we at today? In: Angelone E, Massey G, Ehrensberger-Dow M (eds) *The Bloomsbury companion to language industry studies*. Bloomsbury, London.

Way, A. (2018b) Quality expectations of machine translation. In *Translation Quality Assessment* (pp. 159-178). Springer, Cham.