DA + Criteria: A New Quality Assessment Method for Bridging the Gap Between Human and Machine Translation

Bettina Hiebl
University of Vienna
bettina.hiebl@univie.ac.at

Abstract

Direct Assessment (DA) + Criteria is a translation quality assessment method proposed based on a comprehensive systematic literature review on the concepts of quality in machine translation and translation studies. In the presented project the method was tested alongside Multidimensional Quality Metrics (MQM) on the German translations by humans, DeepL and ChatGPT of English non-fiction texts, using the results of the study as well as the participants' answers to further refine the method.

1 Introduction

Quality assurance is a central component of human translation (HT) and machine translation (MT). However, quality is conceptualized differently in the humanities-based field of Translation Studies (TS) and the AI-based field of MT: while MT focuses mostly on developing metrics (increasingly including concepts put forward by TS scholars, e.g. evaluation at document level and limitations of relying on reference translations), TS still focuses on time-consuming and costly manual evaluation.

The idea of linking the theoretical foundations of translation studies with the practical quality frameworks used in MT is not new (Čulo, 2014). However, previous surveys tend to concentrate on only one aspect of the field, such as post-editing (Koponen, 2016) or the perspective of translation studies (Koby et al., 2014). Additionally, the Multidimensional Quality Metrics (Lommel et al., 2014) proposes a comprehensive catalog of quality issues that can be used to derive evaluation scores for translations.

Focusing on bridging the gap between TS and MT in translation quality assessment (TQA), this project proposes and tests a unified assessment method (DA + Criteria) based on the results of a comprehensive systematic literature review on the concepts of quality in the two fields.

2 Project Overview

The overall project is a dissertation project at the University of Vienna comprising a systematic literature review on human and machine TQA based on the PRISMA method (Page et al., 2021). This review consists of 250 publications assessed focusing on both the authors' background and the main topics (preliminary results published in Hiebl and Gromann (2023b; 2023a)) building the basis for creating a criteria catalog with numerous criteria for TS and MT with relevant scientific references. Based on this a unified assessment method is proposed, i.e. DA + Criteria, combining Direct Assessment (Graham et al., 2013) with criteria from the catalog, i.e. in addition to rating a translation with a slider from 0 to 100, participants can choose criteria as reason for the rating.

The one-year project of testing the unified assessment method has been sponsored by the EAMT Sponsorship of Activities, Students' Edition 2025 and concerns the validation of the proposed unified assessment method by asking translation professionals to perform TQA of human and machine-translated texts with MQM and DA + Criteria. 12 professional translators have been asked to assess translation quality with both methods (each of the methods was used by 6 participants on each of the texts, i.e. all 12 participants assessed all

texts). They evaluated 6 parts of English language texts out of non-fiction books on different topics (politics, history, and biology) and their officially published human translations as well as MT output by DeepL and ChatGPT into German and answered questions on both methods. For assessment with DA + Criteria they could choose up to four criteria (Accuracy, Fluency, Cultural Context, Terminology). While for MQM the assessment was only done sentence by sentence, for DA + Criteria the rating was done both sentence by sentence and overall for the texts.

3 Results

Table 1 shows the type of translation per text, the number of segments and the overall assessments per text and method. For MQM the best-rated text according to the total number of errors (30) and the total error points (37.1)[1] is biologyA translated by a human. Whereas this text also received good results by those participants who used DA + Criteria for its assessment, the best-rated text with this method was politicsB translated using DeepL. Both methods indicated that the worst translation was the one of biologyB by ChatGPT.

4 Conclusions

The main conclusions to be drawn by the participants' answers regarding both methods and the evaluation of the results are that the slider from 0 to 100 is considered too fine-grained, and having a criteria catalog to choose from is seen as useful, but not being able to indicate whether the ticked criteria had a positive or negative influence was considered not ideal. Therefore, the method DA + Criteria will be updated: instead of using just one set of criteria without specifying whether the influence is positive or negative, there will be two sets of the same criteria used for indicating the type of influence, and the slider will be changed to 0 to 10.

[1] 5 error points for each major error, 0.1 error points for each minor Fluency/Punctuation error, 1 error point each for all other minor errors

References

Čulo, Oliver. 2014. Approaching machine translation from translation studies–a perspective on commonalities, potentials, differences. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 199–206, Dubrovnik, Croatia.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Hiebl, Bettina and Dagmar Gromann. 2023a. Human & machine translation quality: comparing & contrasting concepts. In *Transl. Comput 45*, pages 108–128.

Hiebl, Bettina and Dagmar Gromann. 2023b. Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384.

Koby, Geoffrey S, Paul Fields, Daryl R Hague, Arle Lommel, and Alan Melby. 2014. Defining translation quality. *Tradumàtica*, 12:0413–420.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2).

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of Surgery*, 88:105906.

Info	**biologyA**	**biologyB**	**historyA**	**historyB**	**politicsA**	**politicsB**
translated by	HT	ChatGPT	DeepL pro	ChatGPT	HT	DeepL pro
segments	10	9	7	9	10	10
MQM total errors	**30**	**79**	62	51	68	38
MQM total error points	**37.1**	**234.1**	200	72.1	161.1	48
DA+Crit mean scale	76.72	**64.81**	68.14	70.87	76.98	**78.27**
DA-Crit Doc mean scale	70.50	**53.83**	57.50	71.17	75.83	**79.17**

Table 1: Main Results; mean values are per participant