Evaluating Machine Translation for Emotion-loaded User Generated Content (TransEval4Emo-UGC)

Shenbin Qian[1], **Constantin Orăsan**[1], **Félix do Carmo**[1], **Diptesh Kanojia**[2]
[1]Centre for Translation Studies
[2]Institute for People-Centred AI
University of Surrey, UK
{s.qian, c.orasan, f.docarmo, d.kanojia}@surrey.ac.uk

Abstract

This paper presents a dataset for evaluating the machine translation of emotion-loaded user generated content. It contains human-annotated quality evaluation data and post-edited reference translations. The dataset is available at our GitHub repository.[1]

1 Introduction

Machine translation (MT) technology has developed so rapidly in recent years that some claimed to have achieved human parity in Chinese–English news translation (Hassan et al., 2018). Different from news translation, automatically translating user generated content (UGC) has revealed additional challenges for MT systems including handling slang, emotion, literary devices such as irony and sarcasm (Saadany et al., 2023). This is particularly prominent in Chinese social media texts, as various homophones are used to replace offensive words to avoid censorship (Qian et al., 2023).

To evaluate how MT systems perform on emotion-loaded UGC, we collected Chinese microblog texts, and employed Google Translate[2] (GT) to translate them to English. Trained annotators were recruited to directly evaluate translation quality in terms of emotion preservation (through error annotation). Professional translators were hired to post-edit the GT outputs to produce reference translations. Post-edited translations can be used to compare with the MT outputs and to showcase the high-quality translations achievable by human translators. The human evaluation process

[1]https://github.com/surrey-nlp/HADQAET for data & licence
[2]https://translate.google.co.uk/ on the 30th of May, 2022

was funded by the University of Surrey. The post-editing activity was funded by the European Association for Machine Translation (EAMT) through its 2022 sponsorship of student activities.

2 Data Description

This project delivered a dataset comprising 5538 instances of Chinese microblog texts (source), their machine-translated English versions, information on human-annotated errors and quality evaluation scores (QEval information), and post-edited translations (reference).

2.1 Source

The source originated from the dataset released by the *Evaluation of Weibo Emotion Classification Technology on the Ninth China National Conference on Social Media Processing* (SMP2020-EWECT). The original dataset was sourced from *Weibo*,[3] the largest microblogging platform in China. It has a size of 34,768 instances. Each instance is a tweet-like text segment, which was manually annotated with one of the six emotion labels, i.e., *anger*, *joy*, *sadness*, *surprise*, *fear* and *neutral* (Guo et al., 2021). We selected a random sample of 5538 instances (20%) with non-neutral emotion labels as our primary resource to examine how MT renders emotion-loaded UGC.

2.2 QEval information

Two annotators with Chinese–English translation qualifications were recruited to evaluate the quality of these GT outputs in terms of emotion preservation. The evaluation employs the Multi-dimensional Quality Metrics (MQM) framework (Lommel et al., 2014), to assess the translation quality across various error dimensions like

[3]https://weibo.com/

accuracy, fluency, terminology, and more. We introduced a modified framework inspired by MQM to annotate error types and severity levels related to emotion preservation. The output of this process is human-annotated errors and their corresponding severity levels for each of these 5538 instances. Details for the new MQM-based framework, error annotations (including annotation guidelines and inter-annotator agreement), error analysis and data distribution can be seen in Qian et al (2023).

These errors can be used to calculate a quality evaluation (QEval) score based on the weight assigned to each severity level. This score represents the human assessment of the MT quality in terms of emotion preservation. The words that cause the errors were also annotated as wrong/bad translations, which can provide insights into translation quality at word level. QEval scores and error annotations can be utilized to train machine learning systems. These systems can then predict similar scores and word annotations, offering a way to approximate human evaluation in the absence of a gold standard.

2.3 Reference

We hired a translation company to post-edit 2778 instances of the GT output. These instances were identified as having errors related to emotion preservation during the quality evaluation process. Before the post-editing process starts, the translators received clear instructions that: 1) the task involves the post-editing of the provided GT translations, and 2) maintaining the source's emotion is just as crucial as conveying its meaning. They were given enough time (approximately one month) to complete the job to avoid fatigue and ensure quality. The post-edited translations were delivered in two batches for quality checks using random sampling.

While training quality estimation systems can serve as a proxy for quality evaluation (Specia et al., 2018), the system performance improves when human-translated references are accessible (Wan et al., 2022). These high-quality reference translations are valuable for comparing machine translation and training automatic QEval systems.

3 Conclusion

To our best knowledge, this dataset is the first open-sourced Chinese–English resource in the MT area that includes human-annotated translation errors, words that cause the errors, quality evaluation scores in terms of emotion preservation, post-edited reference translations, and emotion labels. We believe it is valuable for the evaluation of translation quality for emotion-loaded UGC, and for the training of new MT systems.

References

Guo, Xianwei, Hua Lai, Yan Xiang, Zhengtao Yu, and Yuxin Huang. 2021. Emotion Classification of COVID-19 Chinese Microblogs based on the Emotion Category Description. pages 916–927. Chinese Information Processing Society of China, August.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXive preprint*.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality. *Tradumàtica: tecnologies de la traducció*, 12:455–463, December.

Qian, Shenbin, Constantin Orasan, Felix Do Carmo, Qiuliang Li, and Diptesh Kanojia. 2023. Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland, June. European Association for Machine Translation.

Saadany, Hadeel, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. Analysing mistranslation of emotions in multilingual tweets by online MT tools. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland, June. European Association for Machine Translation.

Specia, Lucia, Caroline Scarton, and Gustavo Henrique Paetzold. 2018. *Quality Estimation for Machine Translation*. Morgan Claypool.

Wan, Yu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland, May. Association for Computational Linguistics.