Cultural Transcreation with LLMs as a new product

Beatriz Silva[1], Helena Wu[1 2 3], Yan Jingxuan[1], Vera Cabarrão[1], Helena Moniz[2 3], Sara Guerreiro de Sousa[1], João Almeida[1], Malene Sjørslev Søholm[1], Catarina Farinha[1], Paulo Dimas[1]

[1]Unbabel, Lisbon, Portugal
[2]University of Lisbon, Portugal
[3]INESC-ID, Lisbon, Portugal
{beatriz.silva, helena.wu.int, yan.jingxuan.int, vera.cabarrao, helena, sara.guerreiro, joao.tiago.almeida, malene.soeholm, catarina.farinha, pdimas}@unbabel.com

Abstract

We present how at Unbabel we have been using large language models (LLMs) to apply a cultural transcreation product on customer support emails and how we have been testing the quality and potential of this product. We discuss our preliminary evaluation of the performance of different MT models in the task of translating rephrased content and the quality of the translation outputs. Furthermore, we introduce the live pilot programme and the corresponding relevant findings, showing that transcreated content is not only culturally adequate but it is also of high rephrasing and translation quality.

1 Introduction

As defined by Díaz-Millón and Olvera-Lobo (2021:358), transcreation is "a type of translation characterized by the intra-interlingual adaptation or re-interpretation of a message intended to suit a target audience (...) paying special attention to the cultural characteristics of the target audience". While transcreation has multiple uses and areas to which it can be applied to, our focus is on the cultural dimension, particularly in the field of machine translation in the domain of customer support, exploring in a near future the extrapolation to other domains, such as marketing. Our goal is to enable companies to move to markets distant from their culture while feeling confident that they will be able to effectively communicate with their customers.

Our approach involves prompting an LLM for rephrasing the source text produced by the customer support (CS) agents in English, before the text is translated with an MT model into the target languages (TL) of Japanese (JA), Korean (KO) and Mandarin Chinese (ZH). Our product can adapt messages in the source to the culture of the target audience without depending on the knowledge of the agent, making the final message culturally appropriate and, thus, improving communication between both parties, as shown in **Figure 1**.

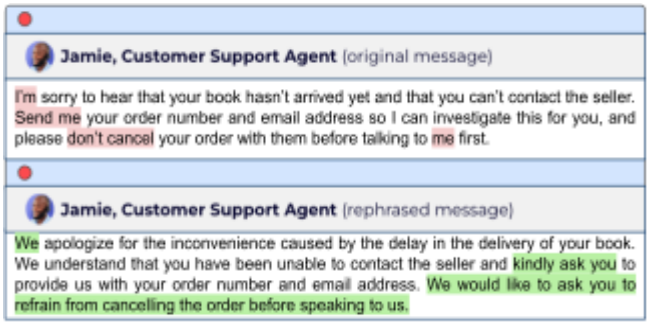

Figure 1: Rephrasing example for JA as TL

2 Prompts and MT Quality

The first step in our transcreation task was to construct the rephrasing prompts for each of the TL. This was achieved through compiling the findings of our research on cultural and linguistic aspects of the TG, specifically regarding CS communication etiquette, as well as the input of native speakers, who are linguistic experts, and non-native speakers, which have lived in these countries for a period of time and learned the language there, into language specific guidelines. As languages which live in the confucianist cultural sphere of influence, our object languages share characteristics such as conveying politeness by showing deference and honoring the interlocutor (Kádár and Mills, 2011). However, the degree and form through which these should be applied differ between them and, thus, different prompts were built depending on the TL.

The next step was ensuring that the content rephrased in the source using our prompts could produce high-quality translations. In order to test this, we translated a total of approximately 1000 rephrased segments distributed across three language pairs (LPs): ~400 segments for EN–JA and EN–KO

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2), pages 56–57
June 24-27, 2024 ©2024 European Association for Machine Translation

and ~200 for EN–ZH, with six different MT providers (Azure, AWS, DeepL, Google, ChatGPT and GPT-4) and annotated the outputs so the quality of the translations could be evaluated through Multidimensional Quality Metrics (MQM) framework scores (Lommel *et al.* 2014). The resulting average scores were of around 88 MQM for both en–ja and en–ko, with en–zh scoring higher with an average of 94.3 MQM, reflecting that different MT engines can be successful in the task of translating rephrased content with no adaptation or customization. In addition, by running an automatic quality estimation metric (QE) (Kepler *et al.* 2019) on the translation of the original message and their rephrased versions for the three LPs, we could see that the transcreated messages score higher on average, thus indicating that cultural transcreation (CT) has the potential to improve translation quality.

For the purpose of testing our CT product, we have integrated it into a CS platform in the form of a widget with a "Rephrase" button which calls an API endpoint attached to the CT service after a pre-processing step, including data anonymization. The service relies on an LLM (GPT-4), to rephrase the message according to the prompts, and returns the final rephrased message to the agent in seconds, ready to be translated and sent to the recipient.

3 Live Pilot Programme

Three customers were selected for our pilot evaluation and the data produced during this period, namely the original text, the rephrased text, and the target text (MT version sent to recipients), are being continuously assessed and analyzed quality-wise.

During the first three weeks of the pilot, around three hundred CS emails (based on traffic volume per language) were rephrased: 58.7% for JA, 30.3% for ZH and 11% for KO. In terms of CT quality, almost all the rephrased texts achieved the target-culturalization, and only about 8% were not culturally aware, but no critical level of errors were found. The more frequent rephrasing failures in all three languages were unnatural expressions, inappropriate word substitutions, and errors in the format of greetings and closings.

3.1 LLM Comparison

In order to optimize the prompt version and to choose whether the product needs to update the LLM used and which LLM to adopt specifically, we have been observing and recording the bugs and feedback based on the progress of the pilot evaluation. One month into the pilot, a third version of the prompts for each of the three languages was built on top of the second version. Then, three LLMs were chosen for comparison: GPT-3.5-turbo-16k-0613, GPT-4 and GPT-4-turbo-preview. In this phase the rephrasing outputs were assessed manually in order to evaluate their performance, namely cultural adequateness and adopted prompt rules, and select the most suitable LLM for this task. The final comparison showed that the best performing model was GPT-3.5-turbo-16k-0613 for JA, GPT-4 for KO, and GPT-4-turbo-preview for ZH.

With the continuous quality monitoring, the improvement of the prompts and the LLM comparison, many of the issues have been reduced. For example, the greetings and closings format errors in EN–ZH rephrasing no longer occur in the latest pilot data. These changes may not only be influenced by the optimization factors of prompts, but also due to the improved compliance and stability of the latest versions of the LLM.

3.2 Future Work

As ongoing work, we are including this product in the translation pipeline for other customers, languages and domains. This is supported in the high quality of the rephrasing, its impact on the MT quality without any customization or adaptation, and the flexibility of managing the best LLM per language with automatic metrics (e.g. QE).

Acknowledgements

This work was developed within the scope of the project nº 62 - "Center for Responsible AI", financed by European Funds, namely "Recovery and Resilience Plan - Component 5: Agendas Mobilizadoras para a Inovação Empresarial", included in the NextGenerationEU funding program and was partially founded by FCT, Fundação para a Ciência e a Tecnologia, under project UIDB /50021/2020(DOI:10.54499/UIDB/50021/2020).

References

Díaz-Millón, M., & Olvera-Lobo, M.D. (2021). Towards a definition of transcreation: A systematic literature review. *Perspectives, 31(2)*, 347–364

Kádár, D., & Mills, S. (2011). Politeness in east Asia: An introduction. *Cambridge University Press*, 1–17

Lommel, A., Burchardt, A., Popović, M., Harris, K., Avramidis, E. & Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. Mauro Cettolo *et al.* (eds) (2014) *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik: European Association for Machine Translation, 165–172

Kepler, F., Trénous, J., Terviso, M., Vera, M. & Martins, André F. T. (2019). OpenKiwi: An Open Source Framework for Quality Estimation. arXiv:1902.08646