University *of Ljubljana*
Faculty *of Computer and Information Science*

Human Evaluation Template: Machine Translation Comparison

Lea Vodopivec, Ines Karažija and Isabell Furkert

Abstract

Advisors: Aleš Žagar

Introduction

The differentiation between formal and informal language registers play an important role in communication. However, machine translation systems still struggle to distinguish between them accurately [1]. This project examines how different Large Language Models (LLMs) handle register-sensitive translation, analysing their ability to adapt to formal and informal contexts, while identifying common errors and biases. Since control over formality is essential for producing natural and culturally appropriate translations, failures in this area can lead to awkward or even disrespectful output. Therefore, improving the ability of machine translation to manage formality remains a key challenge in natural language processing. European languages, such as English, Slovene and German, often express formality through complex syntax (e.g. subordinates clauses) or sophisticated vocabulary and phrasing.

Related Work

A study conducted by Wang et al. [1] in 2023 explored the effectiveness of different prompting techniques to guide pretrained LLMs such as GPT-3 and ChatGPT by producing translations with controlled levels of formality. They introduce a robust Transformer-based classifier that outperforms previous evaluation metrics for formality-controlled translation. Ndejde et al. [2] also addressed the challenge of controlling formality in translations done by LLMs in their research. Their study highlights the issue of honorifics, using the example of translating „Are you sure?“into both formal and informal German. They introduce the CoCoA-MT dataset and an evaluation metric for training LLMs, demonstrating that fine-tuning on labeled contrastive data allows models to achieve high accuracy while maintaining translation quality. In the work of Li & Hu [3], the two researchers explore the shining-through effect in English translations from Chinese across four different registers, comparing human and machine translations. Their results show that the effect is present in both but more pronounced and persisted in machine-generated texts. The effect varies depending on the register, being most present in general and academic texts. Mekki et. al [4] also contributed to the field of register-aware NLP by introducing the TREMoLo-Tweets corus, a large-scale French dataset of over 200 000 tweets annotated with labels based on their register. The project itself both demonstrates a way to automatically identify and analyze register in informal genres, such as social media, and provides a replicable framework for training register classifiers, which is relevant to evaluating register preservation in translated outputs as well. Moreover, the authors explore how linguistic descriptors (i.e. contractions, emoji use, verbal tense diversity) could be adapted to evaluate whether LLMs shift or preserve register when translating text between languages, and as such highlight the importance of capturing linguistic cues when evaluating whether an LLM translation maintains the tone, intention, and formality of the original. Similarly, Vela and Lapshinova-Koltunski [5] propose a register-based evaluation framework that incorporates lexico-grammatical features into the evaluation of machine translation output. By training classifiers on various linguistic characteristics, associated with different registers, they show that machine translations tend to share more register features with human translations than with original source text. This suggests that register awareness is often already implicitly handled during translation, but still not always consis-

tently preserved across genres. Above all, their work emphasizes the need to evaluate machine-generated outputs in terms of how well they reflect the context and stylistic expectations of the source, not only based on content accuracy and similarity to the golden standard. Their findings also underscore the importance of assessing whether LLM-generated translations preserve or distort register, particularly when translating between formal and informal contexts.

Initial idea

The goal of our project is to analyse, assess and compare how large language models handle register variation when producing translations. Were specifically interested into the distinction between formal and informal language and identifying whether and how these models preserve or shift the intended register of the source text across languages, genres and prompts. Through analysis, we would like to discover if certain models show a tendency toward a neutral, overly formal, or casual style. Our research will involve evaluating translation outputs from different LLMs across a range of text types, including both formal text (i.e. government proceedings) and informal (i.e. social media posts) language. We aim to assess the extent to which register is maintained or altered and under which circumstances, and to explore whether these shifts reveal any underlying biases or limitations in model behaviour. Ultimately, we seek to offer insight into strengths and weaknesses of current LLMs in handling stylistic variation in register. To evaluate register accuracy and variation, we will employ a combination of methods, including automatic formality classifiers, BLEU scores for general translation quality, and human evaluations by native speakers who can assess the appropriateness and fluency of translated output in terms of register.

Dataset & Methods

For our research, we will use the Grammarly's Yahoo Answers Formality Corpus [6], which contains parallel English sentences in formal and informal registers that we will translate to German and Slovene.

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

Discussion

References

[1] P. C. Wang, E. Marrese-Taylor, Y. Matsuo, Prompting pre-trained Large Language Models for formality-controlled En-Ja Translation, in , vol. JSAI2023, pp. 2E5GS602-2E5GS602, 2023.

[2] M. Nadejde, A. Currey, B. Hsu, X. Niu, M. Federico and G. Dinu, "Cocoa-mt: A dataset and benchmark for contrastive controlled mt with application to formality," 2022.

[3] Y. Li, A better or worse communicator? Comparing human and machine translation in source language shining through across registers, in Lingua, vol. 312, p. 103834, 2024.

[4] J. Mekki, G. Lecorv, D. Battistelli, N. Bchet, TREMoLo-tweets: A multi-label corpus of French tweets for language register characterization, in RANLP 2021-Recent Advances in Natural Language Processing, Sept. 2021.

[5] M. Vela, E. Lapshinova-Koltunski, Register-based machine translation evaluation with text classification techniques, in Proceedings of Machine Translation Summit XV: Papers, 2015.

[6] S. Rao, J. Tetreault, Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer, in NAACL, 2018.