Avoiding ‘generatese’: the optimization of NLG systems through fit-for-purpose data collections

María do Campo Bayón
Pilar Sánchez-Gijón
GELEA2LT, Universitat Autònoma de Barcelona, Spain
pilar.sanchez.gijon@uab.cat
maría.docampo@uab.cat

Paper Abstract

Most linguistic research on the use and exploitation of Natural Language Generation (NLG) systems, whether through graphical interfaces (as in the case of ChatGPT or Gemini) or without them, has primarily focused on their ability to generate text on the basis of prompts. These systems have a wide range of applications, one of which is the interlingual translation of text. They are also able to generate text from a prompt, either in response to a question or a request to perform a linguistic task. Their apparent ability to generate coherent text from another text surpasses the functionalities of any previous linguistic resource.

A translated text often retains certain traces of the source text and language, a phenomenon known as "translationese" (Baker, 1993). With the widespread adoption of machine translation, especially in certain genres, there has been an observable intensification of this phenomenon, which has been termed "post-editese" (Toral, 2019). This can be detected through measurements of specific linguistic aspects and comparisons of human and machine translations using parallel and reference corpora.

Recently, AI systems known as Large Language Models (LLMs) have begun to be used in both professional translation and translator training. The potential footprint such systems leave on translated texts could be called "generatese" (Sánchez-Gijón, forthcoming). The principle of language agnosticism that underlies NLG systems can affect not only the form of discourse (the linguistic features of a text) but also its content (the concepts and ideas it contains and how they are developed) (Sánchez-Gijón, 2022; Imran et al., 2023). This paper aims to study the impact of using small, highly fit-for-purpose data collections to optimize NLG systems by reducing the randomness of their responses and mitigating "generatese". We will explore the creation and, in particular, the description of such data collections, along with their potential for enhancing the quality of translations produced by NLG systems.

REFERENCES

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (233 – 250). John Benjamins Publishing.

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology* 15.4. https://doi.org/10.30935/cedtech/13605

Sánchez-Gijón, P. (2022). Neural machine translation and the indivisibility of culture and language. *FORUM*. Vol. 20. No. 2. John Benjamins Publishing Company, 2022. https://doi.org/10.1075/forum.00025.san

Sánchez-Gijón, P. (2024). Towards characterizing "generatese". American Translation & Interpreting Studies Association (ATISA) Conference. Rutgers U., 5-7 April 2024.

Toral, A. (2019). Post-editese: An exacerbated translationese. *Proceedings of the Machine Translation Summit*, Dublin, Ireland, 19–23 August 2019. https://doi.org/10.48550/arXiv.1907.00900