Automatic detection of (potential) factors in the source text leading to gender bias in machine translation

Janiça Hackenbuchner **Arda Tezcan** **Joke Daems**
Language and Translation Technology Team
Department of Translation, Interpreting and Communication
Ghent University
Belgium
firstname.surname@ugent.be

Abstract

This research project aims to develop a comprehensive methodology to help make machine translation (MT) systems more gender-inclusive for society. The goal is the creation of a detection system, a machine learning (ML) model trained on manual annotations, that can automatically analyse source data and detect and highlight words and phrases that influence the gender bias inflection in target translations. The main research outputs will be (1) a manually annotated dataset, (2) a taxonomy, and (3) a fine-tuned model.

1 Credits

This project is a strategic basic PhD research fully funded by The Research Foundation – Flanders (FWO) for the timespan of four years from 01.11.2023 until 31.10.2027, and hosted within the Language and Translation Technology Team (LT3) at Ghent University.

2 Introduction

With an increasing use of and interest in the developments of machine translation (MT) and a growing demand for gender inclusiveness in society, research on gender bias in MT is increasing (Savoldi et al., 2021). An MT system is considered to be gender-biased if it "systematically and unfairly discriminates against certain individuals or groups in favor of others" (Savoldi et al., 2021, p. 846), perpetuating "inaccurate and potentially discriminatory stereotypes" in society (Vanmassenhove, 2024, p. 3).

Word-level studies show that word embeddings used in MT training are highly gender-inflected, where word embeddings of different parts-of-speech (POS) strongly cluster based on their gender in varying domains (e.g., in sports, kitchen, big tech, sexual profanities) (Caliskan et al., 2022). However, research on word embeddings is limited to word-level analysis and has not yet been extended to the full context of natural language source sentences and resulting systematic gender associations in machine translations.

3 Project Description

In this research project, we apply a novel approach to analyse aligned linguistic and morphological features with a focus on gender in source and target texts and apply machine learning methodologies.

English is chosen as the source language, where role names are generally not marked with a gender (e.g. *teacher*) and the target translations are analysed in German and Spanish, grammatical gender languages (Savoldi et al., 2021), where gender is clearly marked (e.g. *Lehrer/Lehrerin*). The aim is to automatically identify and classify "trigger words" in a source context that influence the grammatical gender inflection in the target translation (i.e. whether an MT translates a person as female or male, or instead opts to neutralise or rephrase the word). The project consists of the following main deliverables that differ from previous research: (1) a manually annotated dataset of human gender associations in sentence contexts, (2) a taxonomy based on these annotations, (3) a comparative analysis of human gender associations vs. the MT gender inflections, and (4) a fine-tuned large language model (LLM) that highlights trigger words for gender in a source text.

3.1 Data collection

The first step is the collection of a list of candidate words (role names, e.g. *friend*) including their gender inflection, as sampled from their word embeddings in previous studies. These words are used to filter monolingual English data from different domains, slightly resembling the methodology taken by Ondoño-Soler and Forcada (2022). Following the automatic filtering, the English sentences are filtered manually on a monolingual-level to select gender-ambiguous cases in terms of the singular candidate word. We aim for 2,000 to 5,000 sentences for model fine-tuning.

Next, the filtered data will be machine translated into German and Spanish with publicly available MT toolkits. We document and compare into which gender the MT systems translate each candidate word, first on a word level (e.g. the individual term *friend*) and then in a sentence context (e.g. *After a friend suggested she try it, Ann said, "Sure!"*). The two bilingual corpora (EN-DE, EN-ES) will be aligned and enriched at word level for morphosyntactic information.

3.2 Data Annotation and Analysis

In the next step, the ambiguous English data will be manually annotated to analyse how context influences gender associations. From these annotations, we can compare to what extent human gender associations overlap with an MT system's choice of grammatical gender in a target language.

For each sentence, annotators will be asked to annotate what context influences their gender association. Specifically, they are asked to annotate any trigger words or phrases (e.g., a reference, location or any POS) that they personally consider to influence the gender inflection of a candidate word in that sentence. The annotations will be verified and classified and from this, a taxonomy will be created. A case study with 22 annotators of different genders was already conducted to assess annotation guidelines, annotator agreement and gender influence (Hackenbuchner et al., under review).

Next, we will analyse morphosyntactic features of the data using automated tools. The combination of all morphosyntactic information will reveal patterns between trigger words and the candidate word in question. Based on this analysis, we want to exclude "irrelevant" trigger words, allowing us to focus on "relevant" words or phrases that have the greatest influence on the candidate's gender.

3.3 Model Fine-Tuning and Evaluation

In our final step we apply machine learning (ML) by fine-tuning an LLM based on our annotated and verified data and the information extracted from the morphosyntactic analysis. By seeking previously unstudied patterns that lead to gender bias in MT, with this fine-tuned model we aim to automatically detect gender-triggering words or phrases in a source text and highlight these.

4 Aligned Projects

In line with this research project, the PhD fellow, first author of this paper, is a co-founder and member of DeBiasByUs[1] and a co-organiser of the two International Workshops on Gender-Inclusive Translation Technologies.

5 Conclusion

Our benchmark could make MT users aware of gender inflections of source texts that are machine translated, technologically support translators in post-editing MT output, direct developers of MT systems to persisting issues of gender bias, and help content creators identify potential triggers in text that may lead to gender-biased translations.

References

Caliskan, Aylin, Pimparkar P. Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, page 156–170.

Hackenbuchner, Janiça, Arda Tezcan, Aaron Maladry, and Joke Daems. under review. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with mt.

Ondoño-Soler, Nerea and Mikel L. Forcada. 2022. The exacerbation of (grammatical) gender stereotypes in english–spanish machine translation. *Revista Tradumàtica*, 20:177–196.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. In *Transactions of the Association for Computational Linguistics*, volume 9, page 845–874, Cambridge, MA.

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. In *arXiv preprint*, page 1–24.

[1] https://debiasbyus.ugent.be/