Enhancing Gender Representation in Neural Machine Translation: A Comparative Analysis of Annotating Strategies for English-Spanish and English-Polish Language Pairs

Celia Soler Uguet csuguet@transperfect.com
Fred Bane fbane@transperfect.com
Mahmoud Aymo mahmoud.aymo@transperfect.com
João Pedro Fernandes Torres joao.torres@transperfect.com
Anna Zaretskaya azaretskaya@transperfect.com
Tània Blanch Miró tblanch@transperfect.com
TransPerfect

Machine translation systems have been shown to demonstrate gender bias (Savoldi et al., 2021; Stafanovičs et al., 2020; Stanovsky et al., 2019), and contribute to this bias with systematically unfair translations. In this presentation, we explore a method of enforcing gender in NMT. We generalize the method proposed by Vincent et al. (2022) to create training data not requiring a first-person speaker. Drawing from other works that use special tokens to pass additional information to NMT systems, e.g. by Ailem et al. (2021), we annotate the training data with special tokens to mark the gender of a given noun in the text, which enables the NMT system to produce the correct gender during translation. These tokens are also used to mark the gender in a source sentence at inference time. However, in production scenarios, gender is often unknown at inference time, so we propose two methods of leveraging language models to obtain these labels.

Our experiment is set up in a fine-tuning scenario, adapting an existing translation model with gender-annotated data. We focus on the English to Spanish and Polish language pairs. Without guidance, NMT systems often ignore signals that indicate the correct gender for translation. To this end, we consider two methods of annotating the source English sentence for gender, such as the noun *developer* in the following sentence:

The developer argued with the designer because she did not like the design.

1. We use a coreference resolution model based on SpanBERT (Joshi et al., 2020) to connect any gender-indicating pronouns to their head nouns.
2. We use the GPT-3.5 model prompted to identify the gender of each person in the sentence based on the context within the sentence.

For test data, we use a collection of sentences from Stanovsky et al. (2019) including two professions and one pronoun that can refer only to one of them. We use the above two methods to annotate the source sentence we want to translate, produce the translations with our fine-tuned model and compare the accuracy of the gender translation in both cases. The correctness of the gender was evaluated by professional linguists.

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 171–172
September 4–8, 2023, Macau SAR, China.

	Spanish	Polish
Total sentences	577	314
Baseline (no gender annotation in the source)	448	112
'Gold' gender marking	538	175
SpanBERT method	479	168
GPT method	478	158

Table 1: Results of the human evaluation of gender translation.

Overall, we observed a significant improvement in gender translations compared to the baseline (a 7% improvement for Spanish and a 50% improvement for Polish), with SpanBERT outperforming GPT on this task. The Polish MT model still struggles to produce the correct gender (even the translations produced with the 'gold truth' gender markings are only correct in 56% of the cases). We discuss limitations to this method. Our research is intended as a reference for fellow MT practitioners, as it offers a comparative analysis of two practical implementations that show the potential to enhance the accuracy of gender in translation, thereby elevating the overall quality of translation and mitigating gender bias.

References

Ailem, M., Liu, J., and Qader, R. (2021). Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation.

Stafanovičs, A., Bergmanis, T., and Pinnis, M. (2020). Mitigating gender bias in machine translation with target gender annotations.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation.

Vincent, S. T., Barrault, L., and Scarton, C. (2022). Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 121–130, Ghent, Belgium. European Association for Machine Translation.