Cole Bennett

4/7/2023

A Summary of Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, Marco Turchi

University of Trento

Fondazione Bruno Kessler

With the explosion in the popularity of the use of natural language processing, a key aspect that is not investigated with the rigor it is due is the encoding of societal biases into the patterns learned by constructed models – it has been demonstrated that automatic translation systems have a tendency to overrepresent masculine forms of words and amplify gender stereotypes when performing translation into grammatical-gender languages. According to the authors of this paper, modern research into the surfacing of gender bias in language technologies suffers in two major aspects: evaluation techniques are too coarse-grained, i.e. they do not allow for the exploration of potential correlations between word category and influence on gender bias, and they overlook the fact that when translating to grammatical-gender languages (Spanish, French, Italian, etc.), many words in a particular chain must share the same gender inflection to be grammatically correct. In order to investigate these identified deficiencies, the researchers modified the MuST-SHE training corpus to include additional depth to the linguistic information contained: part-of-speech labelling and agreement chains.

The MuST-SHE training corpus, developed by these same authors along with Mattia Antonio Di Gangi and Roldano Cattoni, addresses the limitations on previous gender-bias oriented training corpora which largely consisted of challenge datasets, which are “controlled artificial benchmarks that provide a limited perspective on the extent of the phenomenon and may force unreliable conclusions” (Bentivogli et al., 2020). MuST-SHE is constructed out of a subset of the MuST-C dataset, an English-French and English-Italian translation dataset of TED talks containing triplets of audio, transcript, and translation data that have been annotated to include balanced and differentiated gender-related phenomena. The addition of POS tagging for the current paper’s inquiry was constricted to only six parts of speech, as this was an exploratory study into the potential impacts of training with this additional encoding. Agreement chains are defined and encoded as “gender-marked words that constitute a phrase, such as a noun plus its modifiers, and verb phrases for compound tenses” (Savoldi et al. 2022). These features were added to the dataset via manual annotations by 6 annotators, with 2 per language pairing. For models on which performance was evaluated, various state-of-the-art transformer-based models were selected.

Evaluation of model performance was carried out at various levels of granularity. SacreBLEU, an evaluation metric allowing for generally comparable Bi-Lingual Evaluation Understudy scores (a modified precision metric accounting for word frequency), semantic coverage, and gender accuracy were the primary metrics reported for broad-level performance. For fine-grained performance, only accuracy in the two additional areas of inquiry was reported. Broadly speaking, they found that invariant to language and model used, all metrics showed favorable performance for male-associated data by roughly 30%. Looking at their first novel evaluation criteria, POS tagging, a wider range of discrepancies was observed (as low as 6% but as high as 77%) with the bias still in favor of males. More interestingly, a significant interaction between part of speech and gender accuracy was found, with nouns having significantly larger performance differences relative to verbs and adjectives, irrespective of language. Examining performance regarding agreement chains, the appearance of gender bias becomes most apparent. Accuracy in this task was categorized into three parts: agreement respected with correct gender, agreement respected with wrong gender, and agreement not respected. Feminine language was misclassified roughly 50% of the time as male language on average across all languages and models, while masculine language was only misclassified approximately 10% of the time across languages and models. The models were successful at maintaining agreement as a whole however, as non-agreement accuracy values ranged from 0%-2%.

This paper serves as an important milestone in the deeper exploration of gender bias in speech and machine translation, as only two papers prior to this research have explored the interaction between syntactic agreement and gender bias, and this paper is the very first study exploring the intersection of part-of-speech, word agreement, and gender bias. Not only do the avenues used to examine gender bias in the present study offer fascinating insight into deeper trends underlying the issue, but also open the door for more fine-grained exploration of the subject and allow for the construction of more targeted corrective measures. The authors on this study have the following number of citations on Google Scholar: Beatrice Savoldi: 125, Marco Gaido: 255, Luisa Bentivogli: 6039, Matteo Negri: 7200, and Marco Turchi: 7443.

Works Cited

Savoldi, Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2022). Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. <https://doi.org/10.48550/arxiv.2203.09866>

Bentivogli, Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., & Turchi, M. (2020). Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. <https://doi.org/10.48550/arxiv.2006.05754>