# Assessing gender bias at sentence versus paragraph levels

Audrin Yann, Morinais Suzie, Rabehanta Line
(line-niryantso.rabehanta@etu.u-paris.fr)

## Abstract

Here, we analyze mistakes that different automatic translators can make compared to a human translation regarding gender, especially regarding feminine entities.
When it comes to gender bias in machine translation, the challenge lies in determining when a bias effectively occurred, and finding ways to debias while maintaining the overall quality of the translation. Some of the previous works ([SB20]) focused on each aspect separately (i.e assessing bias and showing possible debiasing methods). We tried to determine what are some possible causes of gender bias when we consider a whole paragraph (e.g bias might still occur because of the first name, despite clear gender signs in the following sentences).

## Dataset

**Characteristics :**

- 6/138 biographies of Olympic athletes from the translated Wikipedia biographies dataset made by Google ([Ste]).
- 3 -> women, 3 -> men
- originally in English, translated into French manually
- total of 12 texts
- each text tokenized into sentences (sentence-level evaluation)

## Experiments

1. Demonstration of bias with commercial automatic translators such as Google Translate.
2. Use of the mBART model. Unfortunately, it has yet to perform on larger inputs (paragraphs) : we used it only for the sentence-level translations.
3. We realized that gender bias is strongly linked to context, and in particular to information such as first names.

**Evaluation method :**

1. manual comparison of the resulting translations with target sets (at both sentence and paragraph-levels)
2. percentage calculation of correct translation for each system (paragraph-levels)

## Results

**Experiment 3 :** we chose 3 first names which are considered both feminine and masculine (Sam, Alex and Morgan), to see how they would influence the translation of gender (paragraph-level).
**Source :** *Jorien ter Mors (born 21 December 1989) is a Dutch speed skater on both short track and long track.*
**Expected :** **First name* Mors (née le 21 décembre 1989) est une patineuse à la fois de vitesse et de longue piste.*
**Results :**

| | **Google Translate** | **DeepL** |
|---|---|---|
| | Alex Mors (né le 21 décembre 1989) est un patineur de vitesse néerlandais sur piste courte et longue piste. | Alex Mors (née le 21 décembre 1989) est une patineuse de vitesse néerlandaise sur courte piste et longue piste. |
| | Sam Mors (né le 21 décembre 1989) est un patineur de vitesse néerlandais sur piste courte et longue piste. | Sam Mors (née le 21 décembre 1989) est une patineuse de vitesse néerlandaise sur courte piste et longue piste. |
| | Morgan Mors (né le 21 décembre 1989) est un patineur de vitesse néerlandais sur piste courte et longue piste. | Morgan Mors (née le 21 décembre 1989) est une patineuse de vitesse néerlandaise sur courte piste et longue piste. |
| **Correct translation rate overall (%)** | 33.33 | 66.67 |

## Conclusion

To conclude, automatic translators tend to choose masculine over feminine when they confront a name that is not clearly labelled as feminine. Masculine seems to be used for neutral or uncertain translations. It proved neural and commercial translators are biased and designed to attribute some actions to a specific gender. When it comes to paragraph versus sentence-levels translations, DeepL showed how the first one mostly performs better thanks to the several gender indicators being used after the first "unlabelled" sentence.

[Epi07] Dominique Epiphane. Resumen. *Travail, genre et societes*, 18:65–85, 2007.

[SB20] Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*, 2020.

[Ste] Romina Stella. A Dataset for Studying Gender Bias in Translation.

[WZBY21] Guillaume Wisniewski, Lichao Zhou, Nicolas Ballier, and François Yvon. Biais de genre dans un système de traduction automatiqueneuronale : une étude préliminaire (Gender Bias in Neural Translation : a preliminary study ). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 11–25, Lille, France, jun 2021. ATALA.

[ZWY+18] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *arXiv:1804.06876 [cs]*, apr 2018.

This project was supported by IdEx Université de Paris ANR-18-IDEX-0001 (AAP Initiation à la Recherche en Licence, vague 2).