Gender Bias in Machine Translation: a statistical evaluation of Google Translate and DeepL for English, Italian and German

Argentina Anna Rescigno and Johanna Monti
UNIOR NLP Research Group University of Naples "L'Orientale"
a.rescigno1@studenti.unior.it
jmonti@unior.it

Abstract. Despite the significant advancements made in the field of Machine Translation (MT) technology, there are still some challenges that need to be addressed. One such challenge is represented by the issue of gender bias in machine translation systems. The main objective of this study is to examine and investigate the presence of gender bias in MT systems and identify any potential issues related to the use of sexist language. The research evaluates the performance of Google Translate and DeepL in terms of natural gender translation, particularly the frequency of male and female forms used in translating sentences that refer to professions without any other gender-specific words. The evaluation is carried out using the MT-GenEval corpus [2] contextual subset, for English-Italian and English-German language pairs. The paper presents the statistical findings obtained from the evaluation.

Keywords: Machine Translation · Gender Bias · Natural Language Processing.

1 Introduction

Bias in automated systems refers to the tendency of these systems to repeatedly make the same assumptions. In the scientific community, bias in artificial intelligence (AI) has become a significant concern due to the increasing use of AI applications such as machine learning architectures. These systems learn by maximizing prediction accuracy, which means that they optimise themselves based on patterns that appear more frequently in their training data. However, if a certain phenomenon is overrepresented in the training data, the program will optimize for it, as this increases its accuracy [14]. The algorithms are evaluated on sub-samples of original training sets making them more likely to exhibit the same biases observed during training. This recurring biased behaviour is responsible for the lack of diversity in machine translation outputs on multiple levels, which can be attributed to the observed algorithmic bias [12].

In 2005, Koehn [3], identified in a preliminary analysis of the Europarl corpus a male-to-female speakers ratio of 2:1. The corpus is made of about 30 million
words for each of the 11 official languages of the European Union (Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish) and it is one of the most used corpora to train MT systems. Consequently, masculine noun endings or pronouns are more frequently used in some languages where gender agreement with the speaker is required, leading to a bias in translation. As a matter of fact, this translation bias reinforces the existing gender disparity and further exacerbates it.

Nonetheless, not until recent times, limited attention has been given to gender bias in machine translation, specifically regarding the natural gender of languages [10]. One way to identify examples of biases in machine translation is to examine an ambiguous sentence in the source language that is unambiguous in the target language. Ambiguity can occur when a linguistic feature is explicit in one language and not in another one or vice-versa. For instance, in English, natural gender is implicitly understood, while in languages such as Italian and German, it must be grammatically expressed. For example, in a simple sentence such as “I am happy”, the subject “I” has an ambiguous gender in English. However, in Italian, the gender must be grammatically expressed as “Io sono contento” for a male subject and “Io sono contenta” for a female subject. Typical users of machine translation systems may not be aware of this bias in the machine translation outputs, especially if they are not proficient in the source language, and there is currently no tool to notify them about it. However, those who are familiar with the target language may recognise when a specific gender is being used inappropriately or offensively. To mitigate gender bias in machine translation, gender characteristics can be integrated into the training data for neural machine translation systems. However, these types of approaches are still limited and further investigation is needed.

The majority of modern machine translation systems translate at the sentence level, meaning that gender-related issues are usually addressed through statistical methods using training data. As a result, translation errors relating to gender are more common in translations between languages, such as English, which retains features relating to natural gender only with some nouns and pronouns and neuter pronouns for sexless objects, and morphologically rich languages like Italian, French, Spanish, and German. These latter languages require additional information in order to accurately translate gender. When this information is missing, the systems tend to generate the most frequent variants, often resulting in biased translations. A study conducted by Prates et al. [6] found that Google Translate defaults to producing more male-gendered outputs than expected, even when taking into account demographic data. This suggests the existence of a phenomenon known as machine bias [6] [13].

2 Related Works

Several studies have been carried out in the field of Machine Translation to address gender bias in translation outputs. These studies involve training trans-
loration systems, developing specific datasets to mitigate gender bias, and creating external tools to integrate into the MT systems themselves.

Monti [4] carried out a study to explore the issues related to the translation of gender evaluating four MT systems - i.e., Google Translate, Microsoft BING Translator, Lucy KWIK Translator and SYSTRANet. While the first two systems adopt a neural approach, Lucy is a rule-based system and SYSTRAN combines both statistical and linguistic approaches. Monti’s study explored various types of issues in gender translation, such as subject-nominate predicate agreement (e.g., Mary is a diligent employee), subject-object agreement (e.g., Jane seems very nervous), name-apposition agreement (e.g., Mary, our doctor, went away a few minutes ago), name-past participle agreement (e.g., Mary is happily married), name-anaphoric/cataphoric reference (e.g., The student studied really hard for her test). The study revealed that all approaches had translation errors regarding gender.

Despite advancements in the field of Natural Language Processing (NLP), the issue of gender translation continues to be a significant concern, as highlighted by Monti [5]. This issue was further exemplified by a recent study conducted by Rescigno et al. [8] which evaluated the frequency of feminine, masculine and neutral forms in the translation of nouns referring to professions when related to certain adjectives (i.e., beautiful, efficient, intelligent, sad, famous) using three state-of-the-art MT systems. The evaluation focused on English–Italian, English–Spanish and English–French language pairs.

Interlinguistic differences between languages require that implicit information from the source language be made explicit in the target language. Human translators can consider the context (both linguistic and extra-linguistic) to infer any necessary information and translate accordingly, while MT systems may encounter challenges in accomplishing this task. Vanmassenhove and Monti [11] developed gENder-IT, a collection of annotated sentences that focuses on gender-related phenomena, to address the lack of high-quality datasets tailored to investigate specific interlinguistic phenomena, including gender and/or number. The dataset was manually adapted from the MuST-SHE Corpus [1] but, unlike this one, gENder-IT tags also include - other than masculine/feminine referents – a neutral referent, whereas the sentence itself does not provide any explicit hint to detect the gender. The dataset indeed focuses on ambiguous English sentences providing the correct translations in Italian. For example, the English sentence "Do you remember that patient you sent home?" the other nurse asked matter-of-factly can, in fact, have four different translation alternatives, all correct:

1) "Si ricorda quel paziente (M) che ha mandato a casa?" mi ha chiesto l’altro infermiere (M);
2) "Si ricorda quel paziente (M) che ha mandato a casa? " mi ha chiesto l’altra infermiera (F);
3) "Si ricorda quella paziente (F) che ha mandato a casa? " mi ha chiesto l’altra infermiera (F);
4) "Si ricorda quella paziente (F) che ha mandato a casa? " mi ha chiesto l’altro infermiere (M).
Finally, more recent works focusing on the creation of specific datasets, tailored to evaluate the accuracy of MT outputs according to gender, include the one from Currey et al. [2], MT-GenEval. This set is based on real-world data from Wikipedia and covers the translations from English towards 8 target languages (Arabic, French, Hindi, Italian, Portuguese, Russian, Spanish, and German). Moreover, this set integrates existing benchmarks, such as WinoMT,\(^1\) MuST-SHE corpus [1], and GeBioCorpus,\(^2\) providing realistic and gender-balanced counterfactual data for all the languages considered. Further evaluations which include more linguistic combinations and even more centred and balanced regarding gender are presented by Rarrick et al. [7] with the GATE corpus (Gender-Ambiguous Translation Examples). GATE is a linguistically diverse corpus consisting of gender-ambiguous source sentences in English along with multiple alternative translations in the target languages considered, i.e., Spanish, French and Italian.\(^3\) The corpus is made up of thousands of segments which consist of a single English sentence paired with one or two alternative translations in Spanish, French or Italian; each sentence contains at least an Arbitrarily Gender-Marked Entity (AGME), which results in an unmarked name referring to an animate entity in the source language (English) but marked for gender in the target languages.

3 Materials and Methods

This study employs a statistical approach to analyse gender bias in machine translation. Its main objective is to detect instances of gender stereotypes in widely used MT systems, Google Translate and DeepL, for English–German and English–Italian language pairs. Furthermore, the present study aims to examine whether the utilization of an extended context, as proposed by Tiedemann [9], specifically through the incorporation of extended translation units encompassing the sentence under investigation and the preceding sentence, yields any positive or negative effects in the disambiguation and translation of gender, particularly when the gender of the referent or speaker is ambiguous. For example, in the sentence “Godoy is an accountant by profession and is currently studying to receive a degree in social work”, the gender of the entity described with the English substantive “accountant” can only be accurately identified by considering the context, namely the previous sentence, from which it is evident that it is referring to a female entity due to the presence of the pronoun “her”:

\(<\text{context}>\) At the age of 17, Godoy was assaulted by a police officer while protesting the Augusto Pinochet dictatorship and was put in prison for her social activism. \(<\text{sentence}>\) Godoy is an accountant by profession and is currently studying to receive a degree in social work.

---

1 https://github.com/gabrielStanovsky/mt_gender
2 https://github.com/PLXIV/Gebiotoolkit
3 All data, including the corpus and the specific evaluation tool, are publicly available at Github repository https://github.com/MicrosoftTranslator/GATE
3.1 Description of the Dataset

The study utilized the MT-GenEval dataset (Machine Translation Gender Evaluation benchmark), which is a reference dataset created specifically to assess the accuracy of machine translation outputs regarding gender. Unlike other datasets involved in bias identification research, the MT-GenEval dataset does not use synthetic data, but it is based on authentic data from Wikipedia. It contains realistic counterfactual data that is gender-balanced across all included languages. The dataset covers 8 language pairs, consisting of a total of 2,400 sentences each.\(^4\)

For this evaluation, the contextual dataset was considered, which includes sentences containing at least one profession noun without any other phrasal element that could indicate the gender of the referent. These selected sentences are accompanied by one or two antecedent contextual sentences that aim to clarify the gender of the selected referent.

Consequently, the dataset consists of sentences that are ambiguous in terms of the gender of the referent but have a previous contextual sentence that clarifies the gender information (i.e., contains gender information). The sentence set was manually edited, removing any unclear or excessively ambiguous examples, for example
\[
<\text{context}> \text{Cook and Dickerman made this their home and Eleanor had her own room, although she rarely spent the night.} <\text{sentence}> \text{Cook, an expert woodworker, made all furniture.}
\]

Also, some other sentences lacking in the contextual information have been removed as well, such as:
\[
<\text{context}> \ldots <\text{sentence}> \text{A surrogate’s life may be very similar to that of the author.}
\]
\[
<\text{context}> \ldots <\text{sentence}> \text{Cook was pressed into service as a player as a result of injuries.}
\]

The final dataset used for the evaluation includes 448 sentences in Italian and 557 sentences in German.

3.2 Experimental Setup

The methodology used for this research can be divided into two phases. In the first phase, translations of single sentences, i.e. without context, obtained from Google Translate and DeepL for both the Italian and German language sets were evaluated manually. The translations were tagged based on the gender of the referent, which could be either masculine (M), feminine (F), or "non-available" (N/A) when the gender could not be inferred. Subsequently, the outputs from the machine translation systems were compared to the benchmark translations, which consisted of the original sentences from the MT-GenEval corpus. The MT system outputs were manually tagged as either "positive correspondence" (Y), "negative correspondence" (N), or "ambiguous" (A) when the gender of

\(^4\) https://github.com/amazon-science/machine-translation-gender-eval
the referent was not clear. In this way, a statistic on the percentage of male and female outputs that matched the gender of the benchmark sentences has been obtained.

The second part of the experiment focused on examining the sentences again by reintroducing them into the machine translation (MT) systems along with the sentences containing the contextual information to ascertain any changes in the outputs. Similarly, also these latter results have been gender tagged and, after being compared to the benchmark sentences, their correspondence regarding the gender of the referent has been annotated as well.

All the machine translations and related manual evaluations were conducted between February and March of 2023 and are accessible in the publicly available GitHub repository.  

4 Results and Analysis

The gender-annotated sets for Italian and German languages present the following results (Table 1), which will be hereinafter indicated as Benchmark (BM).

<table>
<thead>
<tr>
<th>ITALIAN</th>
<th>GERMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>199</td>
<td>249</td>
</tr>
<tr>
<td>44.4%</td>
<td>55.6%</td>
</tr>
</tbody>
</table>

Table 1. “Clean” dataset statistics (Benchmark).

After the manual input of the sentences into the MT systems under consideration, the obtained results were assessed based on their level of correspondence, with the gender of the referent as presented in the benchmark translation.

A preliminary analysis reveals that the majority of translation outputs have a masculine referent in both the MT systems. Specifically, Google Translate produces about 90% male outputs for both languages; DeepL generates 85% male outputs for Italian and 88% for German.

It is worth noting that there were relatively fewer benchmark sentences with masculine gender compared to the MT outputs (Google Translate produces 406 male-gendered sentences for Italian and 506 male-gendered sentences for German in comparison with the benchmark which includes 249 male-gendered sentences for Italian and 312 male-gendered sentences for German; similarly, DeepL produces 385 male-gendered sentences for Italian and 492 male-gendered sentences for German) (Table 2), making the higher production of masculine translation outputs particularly noticeable. In contrast, the number of feminine-gendered outputs is considerably low for both systems: on average, Google Translate produces 5% female-gendered output results, while DeepL has an average of 8%

5 https://github.com/argentina-res/genderbias_dissertation
female-gendered outputs. From Table 2, it emerges very clearly that both MT systems are biased, as they show a higher percentage of masculine gender in comparison with the benchmark translations in Italian and German.

Besides distinguishing translation results between male (M) and female (F) genders - and, neutral (Neut) for what concerns German language - a further differentiation was made for some translated sentences where there was no other useful element for defining the gender of the referent; this category includes sentences tagged as N/A.

Table 2. Results in number (#) and % of masculine (M), feminine (F) and non-available gender (N/A) outputs (also, neutral (N) translation outputs for German) obtained from Google Translate (GT) and DeepL (DL) MT systems for the Italian (IT) and German (DE) languages in relation to the Benchmark (BM) translation statistics.

<table>
<thead>
<tr>
<th></th>
<th>BM</th>
<th>%BM</th>
<th>#GT</th>
<th>%GT</th>
<th>#DL</th>
<th>%DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>249</td>
<td>55.6</td>
<td>406</td>
<td>90.6</td>
<td>385</td>
<td>85.9</td>
</tr>
<tr>
<td>F</td>
<td>199</td>
<td>44.4</td>
<td>19</td>
<td>4.3</td>
<td>36</td>
<td>8.1</td>
</tr>
<tr>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>5.1</td>
<td>27</td>
<td>6.1</td>
</tr>
<tr>
<td>Total</td>
<td>448</td>
<td>100</td>
<td>448</td>
<td>100</td>
<td>448</td>
<td>100</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>312</td>
<td>56</td>
<td>506</td>
<td>90.8</td>
<td>492</td>
<td>88.3</td>
</tr>
<tr>
<td>F</td>
<td>245</td>
<td>44</td>
<td>34</td>
<td>6.1</td>
<td>55</td>
<td>9.9</td>
</tr>
<tr>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>14</td>
<td>2.5</td>
<td>7</td>
<td>1.3</td>
</tr>
<tr>
<td>Neut</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>0.5</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>557</td>
<td>100</td>
<td>557</td>
<td>100</td>
<td>557</td>
<td>100</td>
</tr>
</tbody>
</table>

The evaluation process entailed comparing the outputs of the machine translation (MT) systems with the benchmark translations. A systematic analysis was performed for each MT output to ascertain and classify the correspondence between the gender of the sentence’s referent and the gender of the referents in the benchmark sentences, thereby assigning either a positive (Y) or negative (N) tag. In cases where this kind of correspondence cannot be identified, the ambiguous (A) tag has been employed.

Table 3. Results in number (#) and % of cases of positive (Y), negative (N) and ambiguous (A) correspondence between benchmark translations and sentences without context translated by the MT systems Google Translate (GT) and DeepL (DL) for the Italian (IT) and German (DE) languages.

<table>
<thead>
<tr>
<th></th>
<th>#GT</th>
<th>%GT</th>
<th>#DL</th>
<th>%DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>254</td>
<td>56.7</td>
<td>275</td>
<td>61.4</td>
</tr>
<tr>
<td>N</td>
<td>173</td>
<td>38.6</td>
<td>146</td>
<td>32.6</td>
</tr>
<tr>
<td>A</td>
<td>21</td>
<td>4.7</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>333</td>
<td>59.8</td>
<td>363</td>
<td>65.2</td>
</tr>
<tr>
<td>N</td>
<td>214</td>
<td>38.4</td>
<td>187</td>
<td>33.6</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>1.8</td>
<td>7</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Table 3 presents the results of the evaluation of the correspondence between the MT outputs and the benchmark translations, indicating the number of positive, negative and ambiguous cases accompanied by their respective percentages. The findings demonstrate a relatively equal distribution between the two systems and the considered languages, regarding their negative and positive correspondence. Google Translate, in fact, shows always more than 50% positive correspondence (Y) with the benchmark sentences for both language pairs (56.7% regarding Italian, 59.8% for German), whereas DeepL presents a percentage above 60% (61.4% for Italian and 65% for German).

In regards to cases where the translated output does not correspond in terms of gender with the benchmark translation, the percentages of such cases are below 40% for both systems: Google Translate achieves approximately 38% negative results for both language pairs, while DeepL results range between 32% and 33%.

At this stage, the experiment includes adding antecedent contextual sentences to all the single sentences under examination in order to detect any positive or negative changes in the MT output for what concerns gender disambiguation. Subsequently, those results have been compared with the respective MT outputs without context.

The findings for Google Translate and DeepL for both the English-Italian and English-German language pairs (Fig. 1) show that the number of male outputs obtained after translating the sentences with the contextual sentences differs from the previous results concerning the translation of the sentences without context (Table 2): in particular, the number of feminine and N/A outputs increases. What differs between the two systems is the different percentages of masculine, feminine, and non-available gender outputs, resulting in DeepL outperforming Google Translate, as it produces less masculine defaults.

Subsequently, the aforementioned procedure was repeated, where the translated sentences, along with their corresponding context, were compared to the benchmark sentences.

As Fig. 1 shows, it is evident that the benchmark sentences exhibit higher percentages of positive correspondence with respect to the gender of the referents in the benchmark sentences. Once again, Google Translate demonstrates a marginal improvement compared to DeepL, with the disparity between the "before" and "after" stages of incorporating contextual information becoming more pronounced.

However, the findings also reveal some instances in which the inclusion of contextual information not only fails to enhance the performance of the machine translation (MT) systems but instead has the opposite effect. This occurs due to an erroneous disambiguation of the gender of the referent, a situation that had not previously arisen despite context was not provided. However, such occurrences are relatively infrequent, as evidenced by the data presented in Fig. 2. In the case of Google Translate, there were only a total of 6 instances where the correct gender disambiguation deteriorated for the English-Italian language pair and 16 instances for English-German translations. Conversely, DeepL presents a
Fig. 1. Graphs showing a comparison of the number of gendered outputs resulting for both Google Translate and DeepL before (no context) and after (w/context) adding the contextual information for English-German translations (upper graphs) and English-Italian translations (lower graphs).

Fig. 2. Graphs showing a comparison of the number of positive (Y), negative (N), and ambiguous (A) concordance results for both Google Translate and DeepL before (no context) and after (w/context) adding the contextual information for English-Italian translations (upper graphs) and English-German translations (lower graphs).
lower number of cases for both language pairs - respectively, 3 for Italian and 4 for German.

5 Conclusions and Future Work

The present study involved comparing two machine translation (MT) systems for the English-Italian and English-German language pairs by examining the frequency of male and female gender usage.

The findings of this study indicate, that the MT systems still have a tendency to default to the masculine gender, with some slight differences between Google Translate and DeepL. Moreover, the inclusion of contextual sentences demonstrates overall improvements in the results concerning the positive correspondence of gender agreement with the benchmark translations. Finally, comparing the two system outputs, it is clear that DeepL takes more advantage of the context sentences than Google Translate as it outperforms the latter in increasing its percentage of positive correspondences with the benchmark translations both for the Italian and German languages. However, it is also interesting to highlight that the integration of contextual information also produces some cases of incorrect gender disambiguation, whereas the gender was correctly identified in the first phase of the experiment. Also, in this case, DeepL presents a lower number of degraded occurrences.

Nevertheless, this investigation presents some limitations, including its exclusive focus on binary gender categories - masculine and feminine. The examination of binary gender within current society might be considered restrictive considering the prevailing trend towards increasing inclusivity, encompassing various aspects such as linguistic inclusivity. As a matter of fact, this and other related studies could lay the groundwork for future research endeavours that encompass a broader range of gender identities.

As a potential future direction for this research methodology, there is also the possibility of conducting more comprehensive evaluations that encompass a wider range of language combinations and more targeted and balanced datasets, especially regarding gender - such as the very recent GATE (Gender-Ambiguous Translation Examples) [7].

References


