OccGen: Selection of Real-world Multilingual Parallel Data Balanced in Gender within Occupations

Marta R. Costa-jussà
Meta AI
costajussa@meta.com

Christine Basta
Universitat Politècnica de Catalunya
IGSR, Alexandria University, Egypt
christine.raouf.saad.basta@upc.edu

Oriol Domingo
Batou XYZ
oriol@batou.xyz

Andre Niyongabo Rubungo*
Princeton University
rn3004@princeton.edu

Abstract

This paper describes the OccGen toolkit, which allows extracting multilingual parallel data balanced in gender within occupations. OccGen can extract datasets that reflect gender diversity (beyond binary) more fairly in society to be further used to explicitly mitigate occupational gender stereotypes. We propose two use cases that extract evaluation datasets for machine translation in four high-resource languages from different linguistic families and in a low-resource African language. Our analysis of these use cases shows that translation outputs in high-resource languages tend to worsen in feminine subsets (compared to masculine), specially in the directions containing English. This is confirmed by the human evaluation. We hypothesize that a sound language generation may contribute to pay less attention to the source sentence and to overgeneralize to the most frequent gender forms.

1 Introduction

Biased NLP systems can mainly cause harm in allocations (e.g., giving job opportunities to particular social groups) and in representation (e.g., by propagating and amplifying stereotypes) Blodgett et al. [2020]. Typical examples are associating neutral words with one gender Bolukbasi et al. [2016] or wrongly translating the gender of an entity because of the influenced of a social stereotype Prates et al. [2020], e.g., *The doctor decided to bring her phone.* to *El* doctor* decidió llevar su teléfono.* instead of translating to the correct form *La doctora decidió llevar su teléfono.*.

NLP biases have several dimensions that should be tackled, including detection, evaluation, and mitigation. This paper focuses on generating balanced datasets to address evaluation and mitigation issues. Our motivation to create balanced datasets for training purposes comes from the fact that previous works have shown that fine-tuning with balanced data Saunders and Byrne [2020], Costa-jussà and de Jorge [2020] mitigates gender bias, while creating balanced datasets for evaluation aligns to further progress in responsible artificial intelligence evaluation[2]. Therefore, we propose a methodology to collect monolingual, bilingual, and multilingual datasets balanced in gender and occupations to train or evaluate Machine Translation (MT) models. The outcomes of this paper are two-fold. On the one hand, the OccGen toolkit can be customized according to research and development needs towards languages and gender definition (beyond binary). On the other hand, two use-cases of this toolkit are presented in this paper, which provide two evaluation benchmarks for

*Work done while at Universitat Politècnica de Catalunya

[2] https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

https://doi.org/10.52202/068431-0106

the particular case of binary gender (masculine and feminine). One use case includes a multiparallel dataset in four high-resource languages of different linguistic families (Arabic, English, Russian, and Spanish). The other use case includes parallel dataset in a low-resource African language (Swahili) with English. We provide MT evaluation of different models with these two benchmark datasets. We analyze the impact on gender performance without requiring an additional specific measure for gender and relying on standard MT automatic evaluation methods. We additionally perform a human evaluation that evaluates the gender accuracy of marked words from our datasets. Our data[3] and toolkit[4] are freely available in Github.

2 Related Work and Definitions

Related work As follows, we summarize studies most similar to ours in terms of datasets created to evaluate gender bias. While there are proposals in a wide variety of natural language processing applications (e.g., language modeling Nadeem et al. [2021], Nangia et al. [2020]), we focus on summarizing the ones in MT. The evaluation of gender bias generally combines the proposal of new datasets and a proposed evaluation methodology. Compilation of datasets for evaluation has been approached by proposing synthetic patterns Stanovsky et al. [2019], Nangia et al. [2020], Nadeem et al. [2021] or real-world selection Costa-jussà et al. [2020], Levy et al. [2021]. Some of these datasets have been analyzed and proven to contain several critical pitfalls Blodgett et al. [2021], including unstated assumptions, ambiguities, and inconsistencies. More recently, Stella [2021] extracted English-Spanish and English-German datasets also from Wikipedia biographies that are balanced in gender, occupations, and diversity in nationalities (up to 90). This dataset has been designed to analyze common gender errors such as gender choices in pro-drop, possessives, and gender agreement. The released dataset provides information about the document ID, the source text, the translated text, the perceived gender, the entity name, and the source URL. One limitation of this dataset is that it is only released in two language pairs (English-Spanish and English-German) and contains nine occupations. A structural limitation that is difficult to overcome is that gender is addressed in a binary way since there is little representation of feminine or masculine occupations in Wikipedia. In Renduchintala et al. [2021], authors build SimpleGEN, which is a gender bias test set based on gendered noun phrases. These phrases follow diverse patterns that cover a single, unambiguous, correct answer. Authors specifically calculate the percent of correctly gendered nouns, incorrectly gendered nouns and inconclusive results. Our approach belongs to the category of real-world datasets. The study most similar to ours Costa-jussà et al. [2020] proposes a toolkit to extract multilingual balanced datasets in binary gender (men and women) from Wikipedia biographies. Differently, OCCGEN is customizable in gender categories considering the broad gender spectrum and balanced within occupations. Limitations of these categories are discussed in Appendix H.

Definitions In this paper, we define bias as one relevant factor that prevents our systems from being equitable. In this sense, an example of occupation bias when translating from English to French in our translation systems will generate more translations to masculine doctors than to feminine doctors because historically, there are more references to masculine doctors in our data. Even if gender is a spectrum more than a categorical variable D'Ignazio and Klein [2018], in our use cases, we are limiting gender to binary (masculine and feminine), and we are relying on the tagged category of the perceived gender from our sources. We are only considering binary gender in our use cases because of the limited representation of other genders in our data (see Figure 2). The gender categories from this paper are exclusively based on our sources, and we are exempt from any responsibility in this matter since it is out-of-the-scope of this paper to perform any categorization of gender.Interested readers for this issue can refer to wonderful online resources[5]. Moreover, within articles, we also balance in number of sentences. We categorize languages as high-resource or low-resource based on the categorization that has been described by Goyal et al. [2022] where languages with available parallel data less than one million samples are considered low-resource, between one to one hundred millions are considered medium-resource, and greater than one hundred millions are categorized as

[3] https://github.com/mt-upc/OccGen_dataset

[4] https://github.com/mt-upc/OccGen_toolkit

[5] `https://www.morgan-klaus.com/gender-guidelines.html`. Our extracted datasets are balanced in gender within occupations in the sense that they have the same number of Wikipedia entities (one entity corresponds to one Wikipedia article) in all genders under consideration for each particular occupation entity. For example, for the case of the *politician* occupation, if limiting to binary categorization of gender (masculine and feminine), we would have *N* a number of articles for feminine politicians and *N* for masculine.

https://doi.org/10.52202/068431-0106

high-resource languages. Finally, we provide impact and bias statements and limitations in Appendix H and a data card in Appendix I.

3 Proposed Methodology: OccGen toolkit

We describe the proposed methodology of the OccGen toolkit (summarized in Figure 1).

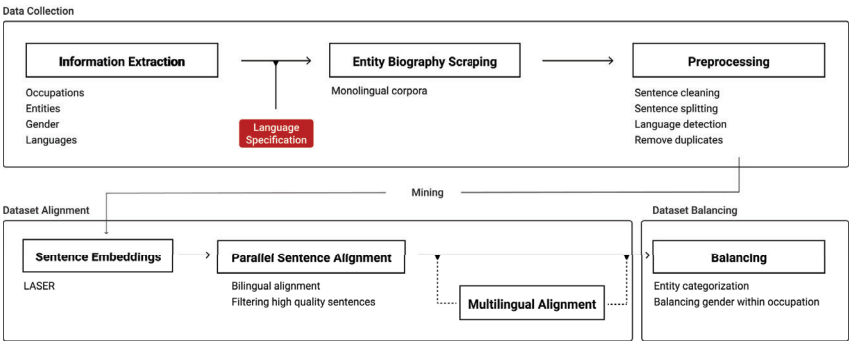

Figure 1: OccGen pipeline overview.

3.1 Data Collection

On the one hand, our metadata are collected from the Wikidata[6] knowledge base. Wikidata is a project that maintains its data quality by monitoring methods and evaluations to guarantee that it suits users' needs. Briefly, our metadata contains a set of people (*from now on* entities) with their occupation(s), gender, and Wikipedia links in all available languages. On the other hand, the textual data that compose our dataset are extracted from Wikipedia[7], similar to Costa-jussà et al. [2020], Stella [2021]. Textual data consists of the text from the entity's biography for one language from Wikipedia.

Information Extraction In this first step, we extract metadata from Wikidata, which relate a set of entities with their working occupations, gender, and biographies from Wikipedia in all available languages. The information extraction procedure is described as follows (see also Appendix A):

1. We extract all the occupations present in Wikidata.
2. For each occupation, we gather the data of every entity that works in the related occupation.
3. For each entity from the previous step, we determine the gender information and related Wikipedia links (corresponding to biographies) in all available languages.

Note that we remove the occupations that do not have related entities and entities that lack gender information. Furthermore, we remove language tags in each entity that do not have a valid ISO language code[8] nor special language code from Wikimedia[9].

Entity Biography Scraping At this step, we specify the languages that will be included in our final dataset. The size of the dataset at the end of the pipeline will be heavily influenced by the type and number of selected languages. For instance, high-resource languages are more likely to have more biographies; nevertheless, a multilingual dataset with high-resource languages may significantly reduce the number of sentences compared to a bilingual dataset. As a result, there are implicit trade-offs between high-resource and low-resource languages and between bilingual and multilingual datasets. By specifying a set of ISO language codes to the system, we scrape all the monolingual data from the corresponding Wikipedia biography for entities with a link for all of the given languages.

Preprocessing As follows, we describe the steps used to preprocess monolingual data.

[6] https://www.wikidata.org
[7] https://www.wikipedia.org
[8] http://www.lingoes.net/en/translator/langcode.htm
[9] https://meta.wikimedia.org/wiki/Special_language_codes

https://doi.org/10.52202/068431-0106

- **Sentence cleaning** Regex expressions are applied to remove the information between brackets and parenthesis, which is mainly related to phonetics, dates, and references.
- **Sentence splitting** Monolingual data are split into sentences; consequently, the sentences are prepared for alignment individually.
- **Language detection** Sentences are fed into a language detection module to exclude those that are not labeled correctly, as Wikipedia pages can mix sentences from several languages, to ensure that all sentences are from the intended language.
- **Remove duplicates** Duplicated sentences are removed to ensure unique sentences for each entity.

3.2 Dataset Alignment

Our mining strategy prepares the data so that each entity's data are represented individually. The next steps perform the sentences embeddings of each language independently and compute the candidates sentences between a source and target language on each entity individually. Then, the final data set is obtained from a multilingual alignment.

Sentence Embeddings We obtain sentence embeddings of each language through a multilingual sentence encoder based on the architecture Schwenk [2018] in which semantically similar sentences are closer to each other, independent of their language Schwenk et al. [2021]. This allows for a common ground for sentences from different languages. It facilitates the use of the multilingual encoder to extract parallel sentences relying on distance-based metrics to perform the next step parallel sentence alignment.

Parallel Sentence Alignment Parallel sentence alignment follows the margin-based criterion introduced in Artetxe and Schwenk [2019] as a metric to execute the nearest neighbor. The margin criterion between two candidate sentences x and y is defined as the ratio between the cosine distance between the two embedded sentences and the average cosine similarity of its nearest neighbors in both directions (equation 1).

$$margin(x, y) = \frac{\cos(x, y)}{\sum\limits_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum\limits_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}}, \tag{1}$$

where $NN_k(x)$ denotes the k unique nearest neighbors of x in the other language and $NN_k(y)$ denotes the same for y. This alignment step allows for getting parallel bilingual candidates which are sorted according to their margin scores, and a threshold is applied to get the desired quality of parallel sentences. This step is performed on each pair of languages independently.

Multilingual Alignment In case of multilingual dataset, we consider one language as the target (pivot) to all other languages, and perform a parallel sentence alignment for each pair of languages. For example, for multilingual alignment of English, Arabic, Spanish and Russian, we can consider English as the target language and perform parallel alignment for English-Arabic, English-Spanish and English-Russian (illustrated on Appendix B). Then we obtain the intersection of common sentences between the language pairs depending on the target language (same pivot sentences). The desired quality of these sentences depend on the chosen threshold of the scores of the alignments (illustrated on Appendix B).

3.3 Dataset Balancing

We aim to obtain a balanced dataset that will contain the same number of sentences per gender within an occupation.

Entity Categorization Entities could have more than one occupation. We categorize entities by the number of occupations they include. Such categorization informs us about the multiplicity of occupations and their corresponding entities in our data. This information enables the choice of categories intended for balancing later. For example, category one represents the entities that have one occupation, category two represents the entities that have two occupations and so on.

Balancing Gender within Occupation The output of this step will be a balanced set with respect to numerous occupations. Each occupation will be represented by a similar number of gender entities, and the total numbers of sentences per gender will be the same. There might be an occupation's

https://doi.org/10.52202/068431-0106

name that refers to a single gender, but the data within this occupation will be balanced regarding all the genders (e.g. *actor/actress*). During balancing, we balance each category (i.e., number of occupations) separately and incrementally, for example, balancing category one (i.e., one occupation) followed by category two (i.e., two occupations). Balancing higher categories (i.e., with multiple occupations) means excluding occupations that already exist in previous lower categories. For example, in the case of extracting category two, if we have an entity with occupations of *doctor* and *politician*, this entity is excluded if either *doctor* or *politician* or both were included occupations in category one. This guarantees that the balancing of the new occupations is not conditioned by balancing the occupations of the last category. We continue by computing the number of gender entities for each occupation and the sum of each gender sentence from all the corresponding entities. For each occupation, balancing will be carried out according to the minimum number of entities and sentences. For example, for binary gender (masculine and feminine), if an occupation has four women and five total related sentences and seven men and ten total sentences, then four entities and five sentences are the maximum intended values for each gender in this occupation. Consequently, this step excludes occupations that have one gender representation (feminine or masculine). We prioritize the masculine and feminine entities[10] that have a similar number of sentences with a higher degree of similarity among languages (i.e., this similarity is based on the margin criterion defined in section 3.2). Details are illustrated in Algorithm 1 for the specific case of using binary gender as we do later in our use cases.

Algorithm 1: *Balancing gender within occupations.*

Input : U_{dic} // An unbalanced dictionary containing the information about occupations, entities in each gender, and aligned sentences with their alignment score.
Output : B_{dic} // A balanced dictionary where each occupation has the same number of sentences in balanced masculine and feminine entities.

$Occs$; // A list of occupations in U_{dic}.
$Em_i; Ef_i$; // A list of masculine and feminine entities with ith occupation, respectively.
$Sm_i; Sf_i$; // Number of sentences in Em_i and Ef_i, respectively.
$B_{dic} = \{\}$; // Initialize the empty dictionary to store the balanced information.
for $i \leftarrow 0$ **to** $len(Occs)$ **do**
    if $len(Em_i)$==$len(Ef_i)$ **then**
        Balance the entities from Em_i and Ef_i such that Sm_i is equal to Sf_i and update B_{dic};
    else
        $Emin$ = min($len(Em_i), len(Ef_i)$);
        if $Emin$==$len(Em_i)$ **then**
            Select only $Emin$ feminine entities with high-quality sentences from Ef_i;
            Balance the entities from Em_i and Ef_i such that Sm_i is equal to Sf_i and update B_{dic};
        else
            Select only $Emin$ masculine entities with high-quality sentences from Em_i;
            Balance the entities from Em_i and Ef_i such that Sm_i is equal to Sf_i and update B_{dic};
        end if
    end if
end for
return B_{dic}

4 Use-case Study

In this section, we report the experimental details of our methodology by including details on two use cases (high- and low-resource languages) limited to balancing in binary gender (masculine and feminine).

4.1 High-resource languages

The top-7 languages with the largest number of entities at Wikipedia are English, German, French, Spanish, Russian, Italian, and Arabic. Among these top languages, there are four linguistic families, Germanic, Latin, Slavic, and Semitic, and we choose one language representing each family. We extract multiparallel data among the high-resource languages that cover different linguistic families, including Semitic (Arabic, ar), Germanic (English, en), Slavic (Russian, ru), and Latin (Spanish, es). The motivation of this use case is to have a balanced dataset in languages that are well-studied in the community. Nonetheless, this dataset can also be used in conjunction with other existing benchmarks

[10]Note that we use the terms feminine/masculine and women/man coherently with linguistic praxis instead of the Wikipedia category female/male

https://doi.org/10.52202/068431-0106

that may contain occupational stereotypes or unbalances in gender. Hereinafter, we alternatively refer to this use case either as the high-resource or en-es-ru-ar use case. The latter mentions specifically the covered languages.

4.2 Low-resource languages

We extract parallel data for the bilingual case of Swahili (sw) and English (en). We choose Swahili to represent low-resource African languages because it has moderate coverage in Wikipedia and is supported by the multilingual sentence encoder we use in the alignment process. The motivation of this use case is to have a balanced benchmark in both occupations and gender for low-resource languages and increase the representation of African languages in the NLP community. Hereinafter, we alternatively refer to this use case as either the low-resource or en-sw use case. The latter explicitly mentions the covered languages.

4.3 Implementation details

We extract data from Wikidata using a Python SPARQL wrapper. For entity biography scraping, we implement an algorithm that works with Beautiful Soup[11], whose purpose is pulling data from HTML content. After that, the following preprocessing techniques are implemented to improve the outcome of our collection process:

- We use regex expressions to clean the collected textual data.
- We use the nltk[12] sentence tokenization package[13] to split sentences across all languages except Arabic, which uses a sentence splitter wrapper[14] for CoreNLP[15].
- We apply language detection to sentences using Compact Language Detector 3[16], which can identify up to 108 languages, to remove sentences that are not labeled with the appropriate language.
- We also remove sentences repeated within an article.

We prepare the text for each entity individually for mining. Then, to execute parallel sentence alignment, we utilize LASER Schwenk and Douze [2017], which provides multilingual sentence embeddings. After embedding the sentences, the aligned parallel sentences in a language pair are computed using the distance in the embedding space. Candidates are sorted according to the order of the similarity between sentences. When aligning multiple languages, English is the pivot language. Finally, note that both datasets, after being automatically extracted, are postedited to be used as evaluation benchmarks. Details on the human postedition are on the Appendix C.

5 Experimental Results

In this section we report the experimental results of the use cases that we proposed.

5.1 Data Statistics

Entities per language. Figure 2 shows the number of entities for each language and gender. We see the difference between high-resource languages and low-resource languages. English is the language with the highest number of entities, and Swahili is the language with the lowest. There is a large difference between gendered representations. We observe very few entities that are not man or woman. The figure shows that all languages have three times more masculine representations than feminine.

Number of entities, occupations and sentences through the pipeline. Table 1 shows the number of entities, occupations, and sentences at the different stages of our pipeline (Figure 1): entity biography scraping, preprocessing, alignment (multilingual or bilingual), and balancing. These statistics show how the number of entities and occupations is reduced at each step for our user cases. They show that alignment and balancing steps have a great impact in reducing the number of occupations and entities,

[11] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[12] https://www.nltk.org
[13] https://github.com/Mottl/ru_punkt
[14] https://github.com/chaojiang06/CoreNLP_sentence_splitter
[15] https://stanfordnlp.github.io/CoreNLP/index.html
[16] https://github.com/google/cld3

https://doi.org/10.52202/068431-0106

Figure 2: Distribution of entities' gender across languages.

illustrating the reason for losing entities and corresponding sentences when choosing more languages to align and balance. The numbers of sentences can only be provided from the alignment step onward, since sentences per language can only be noted individually before this step. As predicted in section 3.3, among the balanced occupations in the high-resource use case, we found occupations' names characterizing only the masculine gender, such as *pornographic actor* or *monarch*. Appendix E reports details on the statistics of entity categorization explained in section 3.3.

		Entity biography scraping	Preprocessing	Alignment$^{(*)}$	Balancing
en-es-ru-ar	entities	15421	14635	2436	286
	occupations	644	256	203	59
	sentences	-	-	6732	524
en-sw	entities	1647	1371	1053	277
	occupations	281	281	73	38
	sentences	-	-	5426	730

Table 1: Evolution of the number of entities and occupations through the pipeline. $^{(*)}$Multilingual alignment, en-es-ru-ar and bilingual alignment, en-sw.

5.2 Machine Translation

System description and implementation To evaluate our dataset, we used the downstream MT task. We used three multilingual models that include the languages from our use cases: M2M_100 Fan et al. [2020], mBART50_m2m Tang et al. [2020] and Opus-MT Tiedemann and Thottingal [2020]. These systems are transformer-based models Vaswani et al. [2017], and they use SentencePiece-based segmentation Kudo and Richardson [2018]. M2M_100, supports translation between any direction for 100+ languages, includes many-to-many supervised training covering thousands of language directions. mBART50_m2m supports translation between any direction for 50+ languages and it has been trained with supervised translation from and to English. Opus-MT supports 1200+ translation directions for 150+ languages.We used the default implementation from EasyNMT[17].

Results Figure 3 reports the results in terms of BLEU for all (top) and the two feminine/masculine subcorpora from our high-resource use case benchmark. See Appendix F for all results in all translation directions. For English and Arabic (both directions) and translating to Russian, the performance is better (or similar) in the masculine set for all models. In the case of translating from Russian, the performance of the feminine subdataset is better than that of the masculine subdataset in all models except for the mBART50_m2m model. When translating to Spanish, this improvement only holds for the M2M_100 model.

For the low-resource languages use case, the results are reported in Figure 4. The M2M_100 model performs the best, and similar to high-resource languages, mBART50_m2m performs the least. We could not obtain the results for Swahili-to-English with Opus-MT models since it does not support this direction.

[17] https://github.com/UKPLab/EasyNMT

https://doi.org/10.52202/068431-0106

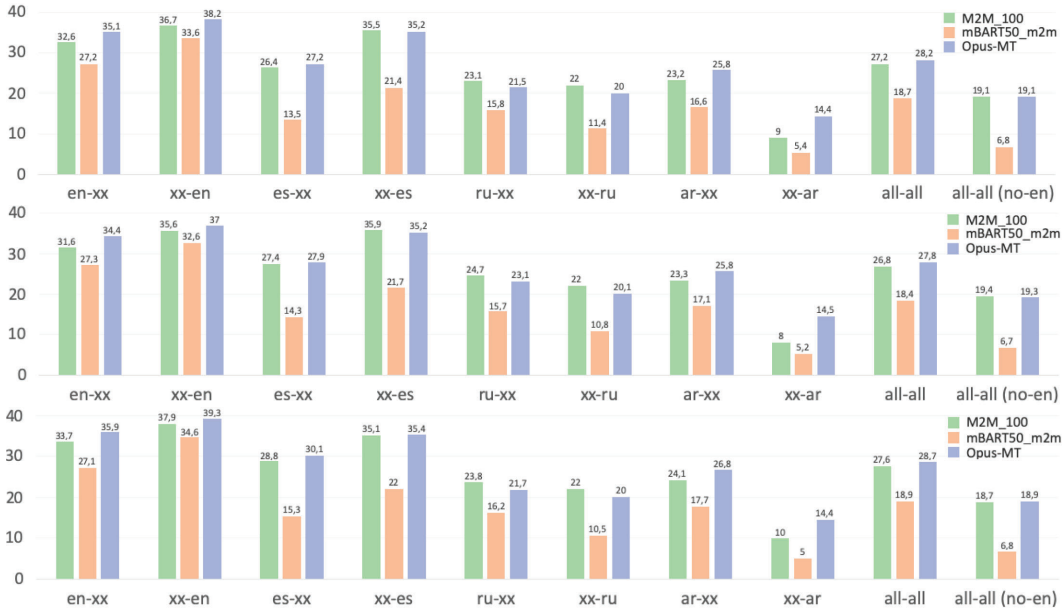

Figure 3: High-resource language results. Average BLEU for M2M_100, mBART50_m2m and Opus-MT models. (Top) All, (Mid) Woman (Bottom) Man.

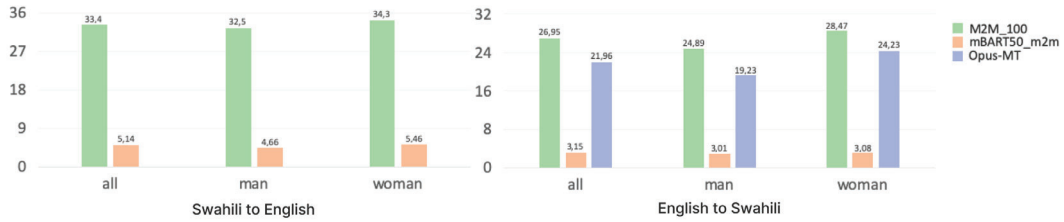

Figure 4: Low-resource language results. BLEU scores for the Swahili to English and English to Swahili direction with M2M_100, mBART50_m2m and Opus-MT models. Note that for Swahili to English, results for Opus-MT model were not obtained because it does not support this direction.

Discussion There is an intriguing pattern that shows that masculine English translations improve feminine set in the all-all performance but feminine translation improve masculine set in the case of all-all (no-en). Furthermore, if we analyze performance with English either on the source or the target side (en-xx or xx-en) on both genders (Figure 3 bottom mid), we found that the difference between the performances of masculine and feminine translations using English on either side is the greatest for gender comparisons in any other high-resource language pair, which explains the change in translation performance with and without English. This reveals that the performance in English may be skewed towards masculine translations in any direction.

When looking at the translation direction (e.g., language A to all or all to language A), we observe English and Spanish to exhibit differed behaviors than Russian and Arabic. The former languages exhibit a higher BLEU performance when the languages are on the target side rather than on the source side. For the same languages, on the target side, the performance on the feminine set tends to be lower than the masculine set. We hypothesize that English and Spanish have a solid language generation. This strong language generation might be, in part, due to the fact that they are written in Latin script, which is shared among many high-resource languages (i.e., Italian, French, and Portuguese). Having a sound language generation may help to achieve higher performance when on the target side. However, it may also contribute to paying less attention to the source sentence and more attention to the previous words in the target sentence, which may explain why the performance in the feminine set did not improve. Less attention given to the source sentence and more attention given to the previously generated words in the target can overgeneralize to the most frequent gender, which tends to be masculine. This suggests that when translating to English, even if source languages (Spanish,

https://doi.org/10.52202/068431-0106

Russian, and Arabic) have high morphological information, the performance on the masculine set is higher than that on the feminine set. This could be explored and confirmed with interpretability measures Ferrando and Costa-jussà [2021], Ferrando et al.. For the latter language set, Russian and Arabic, the performance of the translation direction does not vary as it does in the previous case. This may be because even if they are high-resource languages, they have Cyrillic and Arabic scripts, respectively, which are not shared among other high-resource languages. This may explain the poorer language generation.

Regarding the low-resource use case in Figure 4, the performance in the feminine set is better than that in the masculine set for all models and directions. Both Swahili and English are low-inflected languages in gender, which means we are analysing a language pair which is less prone to having gender bias. The results obtained for Swahili-to-English are better than those obtained for English-to-Swahili, probably due to the language generation for English being better than that for Swahili. The large difference between the performances of the high-resource and low-resource use cases might be caused by the few Swahili data included in the training corpora used to pretrain the used models. Furthermore, the fact that the Swahili-to-English results obtained with the Opus-MT model were not obtained using EasyNMT even though Swahili was claimed to be supported also reveals the challenges of working on low-resource languages, thus the NLP community should increase their representation.

Human Evaluation For human evaluation of gender accuracy, we marked the critical gender word of the sentence in the English part of the high-resource dataset. For example, in the sentence *Sopita Tanasan is a Thai weighlifter.*, the word *weighlifter* was marked. For all language pairs in this set, native annotators marked if the gender of the marked words was correctly translated. In the previous example, when translating to Spanish, the output was *Sopita Tanasan es un levantador* de pesas tailandés.* Since the gender of the marked word is wrong (masculine, *levantador*, instead of feminine, *levantadora*), the sentence is annotated as incorrectly translated. See more details on the annotation in Appendix section D.

We computed gender accuracy as the number of correctly translated gender of marked words (each representing its sentence) divided by the total number of marked words. We conducted the human evaluation for the translation directions of English-to-Arabic, English-to-Russian and English-to-Spanish. Overall, the results agree with the automatic evaluation, where the feminine accuracy is lower than the masculine accuracy for these specific translation directions. The highest feminine accuracy is found in the Russian translations, where the feminine accuracy reaches 0.68 while the masculine accuracy is 0.81. On the other hand, the lowest feminine accuracy occurs in Arabic translations, where the feminine accuracy is 0.41. However, the masculine accuracy is much higher, reaching 0.78. The accuracy of Spanish translations difference between man and woman is the least, where the feminine accuracy is 0.62, and the masculine accuracy is 0.71.

6 Conclusions

This paper proposes the OccGen toolkit to generate monolingual, bilingual, and multilingual balanced datasets in gender within occupations. This freely available toolkit is customizable in languages and gender. Our toolkit simplifies the extraction of parallel data and it is already been used by the community Zhou [2022], Liu [2022]. We present a high and low-resource benchmark. The former includes a multilingual English, Spanish, Russian, and Arabic dataset. The latter includes a bilingual dataset on English-Swahili translations. Both of them are balanced for the particular case of binary gender (masculine/feminine). Note that with our benchmarks, we have postedited the output of the OccGen toolkit to provide an accurate evaluation set. However, we can also use our toolkit to extract training data which is not necessary to postedit. Appendix G provides this kind of data for the English-Arabic, English-Spanish and English-Russian pairs. Differently from the provided benchmark, we have not balanced this data in order to have as many data as possible, covering all genders available in the Wikipedia for the extracted entities. We suggest that this data can be balanced by means of artificial techniques such as oversampling, counterfactual techniques or other synthetic techniques in future. Once balanced, this data can further be used to fine-tune models in order to mitigate gender biases as conducted in previous studies Saunders and Byrne [2020], Costa-jussà and de Jorge [2020]. Our toolkit, benchmarks and training data are released in Github as referenced in section 1.

https://doi.org/10.52202/068431-0106

We report experiments using our benchmark datasets to evaluate MT models. We provide an accurate analysis of performance behavior for the particular case of binary gender. We conclude that feminine translations tend to be worse for high-resource languages with a high-quality language generation model. We hypothesize that, in these cases, the model gives less attention to the source words than the target prefix, and using the target context may overgeneralize to most frequent patterns (which tend to be masculine patterns) rather than producing an accurate translation. This is confirmed by the human evaluation that computes gender accuracy on marked words (for English to Arabic, Spanish, Russian), which shows a higher accuracy for the masculine cases than feminine cases.

Our balanced datasets are not strictly comparable across genders since each gendered subset has different vocabularies. Therefore, a reasonable further step is to automatically generate counterfactual data Qian et al. [2022]. This would modify the masculine subdataset into a feminine subdataset and vice versa. Ultimately, with this data augmentation, our balanced sets would allow for the analysis of gender performance with standard evaluation methods and without requiring new ones.

Acknowledgments and Disclosure of Funding

Authors would like to specially thank annotators for tasks described in Appendix C and D: Carlos Escolano, Gerard Gallego, Ksenia Kharitonova, Gerard Sant, Elena Zotova, Marianne Raouf, Merna Onsy, Engy Basta, Delphina Ndekupatia Severin, Perez Ogayo, and Lawrence Ilagoye Gwajekale. Authors appreciate the insightful comments received by the anonymous reviewers. Authors are extremely thankful to Pascale Fung and Kendra Albert for their valuable ethics discussion; to Christophe Ropers and to Eric Smith for their feedback; and to Delice Musabyimana Agahozo for helping in contacting some Swahili annotators. The work is partially supported by Research Grants (AGAUR) through the FI PhD Scholarship, Universitat Politècnica de Catalunya with the collaboration of Banco de Santander.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. Assessing gender bias in machine translation – a case study with google translate. *Neural Computing and Applications, 32*, page 6363–6381, 2020.

Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.690. URL https://aclanthology.org/2020.acl-main.690.

Marta R. Costa-jussà and Adrià de Jorge. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.3.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://www.aclweb.org/anthology/P19-1164.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.502.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. *arXiv:2109.03858 [cs]*, September 2021. URL http://arxiv.org/abs/2109.03858. arXiv: 2109.03858.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

Romina Stella. A Dataset for Studying Gender Bias in Translation, June 2021. URL https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.15. URL https://aclanthology.org/2021.acl-short.15.

Catherine D'Ignazio and Lauren Klein. Data feminism. MIT Press, 2018.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Holger Schwenk. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037. URL https://aclanthology.org/P18-2037.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.115. URL https://aclanthology.org/2021.eacl-main.115.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

https://doi.org/10.52202/068431-0106

Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay B. Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 157–167. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-2619. URL https://doi.org/10.18653/v1/w17-2619.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125, 2020. URL https://arxiv.org/abs/2010.11125.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401, 2020. URL https://arxiv.org/abs/2008.00401.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018. URL http://arxiv.org/abs/1808.06226.

Javier Ferrando and Marta R. Costa-jussà. Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.39. URL https://aclanthology.org/2021.findings-emnlp.39.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. URL https://arxiv.org/abs/2205.11631.pdf.

Chenuye Zhou. Building a catalan-chinese parallel corpus for use in mt. Master's thesis, Universitat Pompeu Fabra, Barcelona, July 2022.

Zixuan Liu. Improving chinese-catalan machine translation with wikipedia parallel. Master's thesis, Universitat Pompeu Fabra, Barcelona, July 2022.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp, 2022. URL https://arxiv.org/abs/2205.12586.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534627. URL https://doi.org/10.1145/3531146.3534627.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. How to write a bias statement: Recommendations for submissions to the workshop on gender bias in nlp, 2021. URL https://arxiv.org/abs/2104.03026.

Christine Basta. Gender bias in natural language processing, October 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Please see Section 5 and 6 for more details.
   (b) Did you describe the limitations of your work? [Yes] See Section H
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See bias and impact statements in Section H
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.2
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We have generated the translations using EasyNMT, which is a publicly available MT framework that can accessible at `https://github.com/UKPLab/EasyNMT`.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See the Appendix I
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The released toolkit can publicly be accessible at `https://github.com/mt-upc/OccGen_toolkit`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Please check in the Appendix C.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] The human annotation and evaluation were done voluntarily and all the participants are acknowledged.

https://doi.org/10.52202/068431-0106