Lost in Translation: The Case of Gender

Apurva Bhargava **Bashar Alhafni** **Jiaqi Dong** **Bella Lyu**
New York University
{ab8687,ba63,jd3036,hl4229}@nyu.edu

Abstract

In this work, we present an approach to extend single-output gender-unaware NLP systems with user-specific gender reinflections. We focus on two gender marking languages, French and Spanish. Our contributions are the development of a user-aware gender reinflection model and the building of two gender parallel corpora for training and evaluating gender reinflection in French and Spanish.

1 Introduction

The recent progress in many Natural Language Processing (NLP) applications has raised expectations about the quality of results and especially their impact in a social context, including not only race (Merullo et al., 2019) and politics (Fan et al., 2019), but also gender identities (Font and Costa-jussà, 2019; Dinan et al., 2019, 2020). Human-generated data, reflective of the gender discrimination and sexist stereotypes perpetrated through language and speaker's lexical choices, is considered the primary source of these biases (Maass and Arcuri, 1996; Menegatti and Rubini, 2017). However, as Habash et al. (2019) pointed out, NLP gender biases do not just exist in human-generated training data, and models built from it; but also stem from gender-unaware systems designed to generate a single text output without considering any target gender information. Beyond being simply incorrect in many cases, such output patterns create representational harm by propagating social biases and inequalities of the world we live in. One example is the *I-am-a-doctor/I-am-a-nurse* problem in machine translation (MT) systems targeting many morphologically rich languages. While English uses gender-neutral terms that hide the ambiguity of the first-person gender reference, morphologically rich languages need to use grammatically different gender-specific terms for these two expressions. In Spanish, as in other languages with grammatical gender, gender-unaware single-output MT from English often results in *soy-un-doctor 'I'm a [male] doctor'/soy-una-enfermera 'I'm a [female] nurse'*, which is inappropriate for female doctors and male nurses, respectively.

In contrast, gender-aware systems should be designed to produce outputs that are as gender-specific as the input information they have access to. For example, gender information may be contextualized (e.g., the input 'she is a doctor' or 'he is a nurse'). But, there may be contexts where the gender information is unavailable to the system (e.g., 'the student is a nurse'). In such cases, generating both gender-specific forms is more appropriate.

In this work, we propose an approach for gender reinflection using sequence-to-sequence models. We focus on two gender marking languages French and Spanish and formulate the problem as a user-aware grammatical error correction task at the character level. As such, we use as our primary metric the MaxMatch (M^2) scorer (Dahlmeier and Ng, 2012). Our system takes a French or a Spanish sentence and a target gender as input and generates a gender-reinflected sentence based on the target gender.

This report is organized as follows. In Section 2, we discuss some related work. Section 3 introduces the approach we took to construct the French and Spanish Gender Parallel Corpora. In Section 4, we discuss the architecture we used to build our gender reinflection model. Then in Section 5 and 6, we discuss the experiments we did and present some results and we conclude in Section 7.

2 Related Work

Many NLP systems have the ability to embed and amplify societal (gender, racial, religious, etc.) biases across a variety of core tasks such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), machine translation (Rabinovich et al., 2017; Vanmassenhove et al., 2018; Font and Costa-jussà, 2019; Moryossef et al., 2019; Stanovsky et al., 2019; Bergmanis et al., 2020), named entity recognition (Mehrabi et al., 2019), dialogue systems (Dinan et al., 2019), and language modeling (Lu et al., 2018; Bordia and Bowman, 2019).

For the case of gender bias, various research efforts have shown that this could be caused by either human-generated training datasets (Font and Costa-jussà, 2019; Habash et al., 2019), pre-trained word embeddings (Bolukbasi et al., 2016; Zhao et al., 2017; Caliskan et al., 2017; Manzini et al., 2019), or language models (Kurita et al., 2019; Zhao et al., 2019). To mitigate this problem, several researchers proposed approaches in which they focus mainly on debiasing word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b; Gonen and Goldberg, 2019) or using counterfactual data augmentation techniques (Lu et al., 2018; Zhao et al., 2018a; Zmigrod et al., 2019; Hall Maudslay et al., 2019). Most of the solutions were mainly proposed to reduce gender bias in English and may not work as well when it comes to morphologically rich languages.

In this work, we attempt to reduce gender bias that is caused by single-output gender-unaware NLP systems for French and Spanish. To do so, we built two parallel gender corpora and developed a sequence-to-sequence gender reinflection model to extend the output of gender-unaware NLP systems.

3 The French and Spanish Gender Parallel Corpora

To train and evaluate our gender-reinflection models, we need a corpus of French and Spanish sentences that are gender-annotated and gender-reinflected. That is, for every sentence in such corpus, we would like the gender of the sentence's speaker to be identified as F (feminine) or M (masculine) and we would like the equivalent opposite gender form. To the best of our knowledge, no such corpus exists for French nor Spanish. We describe next the approach we followed to build these corpora.

Given the limited resources available to us, we built a synthetic corpus consisting of gender-inflected sentences in Spanish and French. To do so, we created 21 templates and covered 238 entities to produce parallel examples in masculine and feminine forms. For the entities, we used the English entity list provided by Bolukbasi et al. (2016). To produce the masculine form of each entity in French and Spanish, we leveraged Google Translate. Each entity was as added to the third person template "*He is a/an [entity]*" which is then fed to Google Translate to obtain the masculine form of the entities. To get the feminine form of the entities, we used French and Spanish pretrained FastText embeddings (Grave et al., 2018) instead of Google Translate. We decided to use FastText embeddings because the output of Google Translate was not plausible when we tried to obtain the feminine form of the entities. We used the Spanish and French versions of following analogy to obtain the feminine from of the entities:

$$\mathrm{entity_M} - \mathrm{man} + \mathrm{woman} = \mathrm{entity_F}$$

where *man* translates to *hombre* and *homme* in Spanish and French, respectively, whereas woman translates to *mujer* and *femme* in Spanish and French, respectively. By following the above analogy, we were able to obtain the most similar feminine forms of the masculine entities. However, this approach was not perfect and we had to correct some of entities returned by FastText manually to ensure high data quality.

To construct perfectly aligned gender-reinflected parallel sentences, we manually created a set of 21 templates in English. The templates have varying degrees of complexities and most of them contain at least one subjective pronoun (he/she), one objective pronoun (him/her), and an a single entity. All the templates were translated to Spanish by one native speaker, whereas for French, we relied on Google Translate. A complex sentence example is:

	Template Example
EN	I would like to work with [objective pronoun] as a [entity] because [subjective pronoun] cares about all people close to [objective pronoun]
FR	J'aimerais travailler avec [objective pronoun] en tant que [entity] car [subjective pronoun] se soucie de tous ses proches
ES	Me gustaria trabajar con [objective pronoun] como [entity] porque [subjective pronoun] se preocupa por la gente cercana a [subjective pronoun]

Source	Target	Source Gender	Target Gender
el es un doctor y pronto trabajaré con el	ella es una doctora y pronto trabajaré con ella	M	F
ella es una doctora y pronto trabajaré con ella	el es un doctor y pronto trabajaré con el	F	M
ella es una doctora y pronto trabajaré con ella	ella es una doctora y pronto trabajaré con ella	F	F
el es un doctor y pronto trabajaré con el	el es un doctor y pronto trabajaré con el	M	M

Table 1: Example covering all possible combinations of input and output grammatical genders for the Spanish translation of *"[subjective pronoun] is [indefinite article] [entity] and I will be working with [objective pronoun] soon"*. Changed words are underlined.

For each template, we generate its masculine and feminine forms and pair all the possible combinations together. That is, masculine with masculine, feminine with feminine, feminine with masculine, and masculine with feminine. In total, we ended up with $19,992$ $(4 * 238 * 21)$ gender-reinflected parallel sentences for Spanish and French. Additionally, we also used some of the data which was created by Stanovsky et al. (2019) and annotated it manually to increase the size of our parallel corpora. At the end, we ended up with $20,184$ gender-annotated and reinflected parallel sentences for Spanish and with $20,378$ gender-annotated and reinflected parallel sentences for French. We divided the datasets randomly into 80% for train, 10% for development, and 10% for test. Table 1 shows an example of the parallel sentences we generate in Spanish by using *"[subjective pronoun] is [indefinite article] [entity] and I will be working with [objective pronoun] soon"* as a template.

4 Gender Reinflection Model

In this section, we discuss our model architecture as well as the training settings and the model's hyperparameters.

4.1 Model Architecture:

Sequence-to-sequence models have achieved significant results in morphological reinflection tasks (Faruqui et al., 2016; Kann and Schütze, 2016; Aharoni and Goldberg, 2017). Given an input sequence $x_{1:n} \in V_x$ containing k words $w_{1:k} \in V_w$, a gender-reinflected output sequence $y_{1:m} \in V_y$, and a target gender $g \in \{F, M\}$, our goal is to model an auto-regressive distribution which is defined over the target vocabulary:[1]

$$P_{V_y}(y_{1:m}|x_{1:n}, g) = \prod_{t=1}^{m} P(y_t|y_{1:t-1}, x_{1:n}, g; \theta);$$

where θ represents the model's parameters. We implement this model using a character-level encoder-decoder neural network with an attention mechanism. On the encoder side, we use a two-layer bidirectional GRU (Cho et al., 2014). Each character in the input sequence will be mapped to an embedding that is learned during training. For the decoder, we use a two-layer GRU with additive attention (Bahdanau et al., 2015) over the last layer encoder hidden states. At each time step, the decoder receives two inputs: the embedding of the predicted decoder output character and the attentional context vector from the previous time step, to obtain a new a decoder hidden state. The target gender g is mapped to an embedding which is learned during training and is concatenated with the decoder hidden state, the attentional context vector, and the embedding of the predicted character from the previous time step to a create a single vector $\mathbf{z_t}$. We then project $\mathbf{z_t}$ to model the distribution over the target vocabulary using a linear layer followed by a softmax function.

4.2 Inference:

At inference time, we use greedy decoding to find the most likely sequence:

$$\begin{aligned}\hat{y}_{1:m} &= \operatorname*{argmax}_{\hat{y} \in V_y} P(\hat{y}|x_{1:n}, g) \\ &= \operatorname*{argmax}_{\hat{y} \in V_y} \prod_{\hat{y}_t \in \hat{y}} P(\hat{y}_t|\hat{y}_{1:t-1}, x_{1:n}, g)\end{aligned}$$

[1] F stands for Feminine and M stands for Masculine.

	French			Spanish		
	Precision	**Recall**	**F$_{0.5}$**	**Precision**	**Recall**	**F$_{0.5}$**
MLE (bigram)	**96.1**	69.4	89.2	88.0	82.7	86.9
seq2seq	90.4	**94.4**	**91.2**	**90.2**	**92.7**	**90.6**

Table 2: Results on the dev set.

	French			Spanish		
	Precision	**Recall**	**F$_{0.5}$**	**Precision**	**Recall**	**F$_{0.5}$**
MLE (bigram)	**95.0**	72.2	89.4	88.4	83.2	87.3
seq2seq	90.9	**95.3**	**91.8**	**89.2**	**92.6**	**89.9**

Table 3: Results on the test set.

4.3 Hyperparameters:

We use a batch size of 32, a character embedding size of 128, a gender embedding size of 10, a hidden size of 256, a scheduled sampling probability of 0.3, a dropout probability of 0.2, and gradient clipping with a maximum norm of 1. We train the model for 50 epochs by minimizing the average cross-entropy loss. We use the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0005, decaying by a factor of 0.5 if the loss on the development set does not decrease after 2 epochs.

5 Experiments and Evaluation

5.1 Metrics

We use the MaxMatch (M^2) scorer (Dahlmeier and Ng, 2012), which is a widely used metric in grammatical error correction tasks. The (M^2) scorer computes the word-level edits between the input and reinflected output. We report the precision, recall, and F$_{0.5}$ scores calculated against the gold edits, which will also be created by the M^2 scorer.

5.2 Baseline

For our baseline, we use a bigram maximum likelihood estimation (MLE) model. Given an input sequence of words $x_{w_{1:n}} \in V_{x_w}$, a target sequence of words $y_{w_{1:n}} \in V_{y_w}$, and a target gender $g \in \{F, M\}$, the MLE model is built as follows:

$$P(y_{w_i}|x_{w_i}, x_{w_{i-1}}, g) = \frac{count(y_{w_i}, x_{w_i}, x_{w_{i-1}}, g)}{count(x_{w_i}, x_{w_{i-1}}, g)}$$

The MLE baseline is suitable for our case because the input and output sentences are perfectly aligned on the word-level.

6 Results

We trained two separate systems: one for French and one for Spanish. The results of our evaluation on the dev set are presented in Table 2. For French, the MLE results are surprisingly competitive in terms of precision, scoring higher than the neural sequence-to-sequence model; while being worse in terms of recall and F$_{0.5}$. Whereas for Spanish, the neural reinflection model was superior to the MLE model across all metrics.

The results on the test set using the baseline and the neural system are given in Table 3. These results show consistent conclusions with the dev set results. We realize that the evaluation results for the dev and test sets are somewhat high. One possible explanation for this is that the majority of the parallel sentences in the corpora we built are easier than others.

7 Conclusion and Future Work

In this work, we proposed a solution to single-output NLP systems that allows users to specify their grammatical gender preference in both French and Spanish. Our intention is to enable users to reduce the harm that may be produced by NLP systems propagation of biased representations. We also introduced two new gender parallel corpora for French and Spanish. In future work, we would like to explore different architectures such as Transformer-based models (Vaswani et al., 2017). Furthermore, we are interested in exploring the added value of linguistic features into our neural models. We would also like to apply our approach to different languages and dialectal varieties.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Toms Bergmanis, Artūrs Stafanovičs, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *ArXiv*, abs/1911.03842.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing.

Anne Maass and Luciano Arcuri. 1996. Language and stereotyping. *Stereotypes and stereotyping*, pages 193–226.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to person as woman is to location: Measuring gender bias in named entity recognition.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of american football broadcasts. *arXiv preprint arXiv:1909.03343*.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.