The User-Aware Arabic Gender Rewriter

Bashar Alhafni, Ossama Obeid, Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
{alhafni,oobeid,nizar.habash}@nyu.edu

Abstract

We introduce the User-Aware Arabic Gender Rewriter, a user-centric web-based system for Arabic gender rewriting in contexts involving two users. The system takes either Arabic or English sentences as input, and provides users with the ability to specify their desired first and/or second person target genders. The system outputs gender rewritten alternatives of the Arabic sentences (provided directly or as translation outputs) to match the target users' gender preferences.

Bias Statement

Most NLP systems generate a single output for a specific input without taking their end users' grammatical gender preferences into consideration. Such systems typically result in output patterns that create representational harms by propagating biased stereotypes, such as associating certain professional activities or occupations with a particular gender. The system we present in this paper, allows the users to provide their desired gender preferences to provide them with user-aware unbiased outputs. We acknowledge that by limiting the choice of gender expressions to the grammatical gender choices in Arabic, we exclude other alternatives such as non-binary, gender-inclusive or no-gender expressions. We are aware of growing discussions around developing such alternatives in Arabic (UN, 2018; Ala'uldeen, 2022).

Figure 1: Google Translate's output for *"I am a doctor and you are a nurse"* in Arabic. Doctor is translated to the masculine form ('طبيب' *Tbyb*), whereas nurse is translated to the feminine form ('ممرضة' *mmrDħ*).

1 Introduction

Gender stereotypes, both negative and positive, are manifest in most of the world's languages (Maass and Arcuri, 1996; Menegatti and Rubini, 2017) and are further propagated and amplified by NLP systems (Sun et al., 2019; Blodgett et al., 2020) (see Figure 1). This is because NLP systems rely on human-created language corpora that mirror the societal biases and inequalities of the world we live in (Boyd and Crawford, 2012; Olteanu et al., 2019). For instance, Figure 2(a) presents part of a cooking recipe published on an Arabic popular cooking website targeting female readers,[1] whereas Figure 2(b) shows part of an article on career advice that is published on Harvard Business Review in Arabic targeting male readers.[2] However, even if overt gender biases are removed from datasets before using them to build NLP models,

[1] https://www.atyabtabkha.com/
[2] https://hbrarabic.com/

(a)

(b)

Figure 2: Examples of gender-specific text in the wild. Figure (a) is an example of text targeting female readers from a website about cooking recipes. The example is an introduction to a recipe for Baba Ghannouj. Figure (b) is an example of text targeting male readers from a website about career advice. The example is about an advice on how to find a new job. The underlined words are morphologically marked for the second person feminine in (a), and the second person masculine in (b).

this will not ultimately reduce the biases produced by systems that are designed to generate a single text output without taking their target users' gender preferences into consideration.

Some commercial NLP systems have solved this problem by generating more than one gender-specific output when the system encounters ambiguous scenarios. For instance, Google Translate generates both feminine and masculine translations when translating gender-neutral English sentences (e.g., *I am a doctor*) to a limited number of languages, such as Spanish (Kuczmarski, 2018; Johnson, 2020). However, this approach does not work well in multi-user contexts (first and second persons, with independent grammatical gender preferences), particularly when dealing with gender-marking morphologically rich languages. One example of this phenomenon is the Arabic machine translation of the sentence *I am a doctor and you are a nurse*. Figure 1 shows that Google Translate outputs the Arabic translation أنا طبيب وأنت ممرضة *ÂnA Tbyb wÂnt mmrDħ*[3] 'I am a [male] doctor and you are a [female] nurse', whereas a more suitable output would include all four possible Arabic translations of the input sentence.

One approach to mitigate the ambiguity is to provide the users with the ability to specify their desired target gender preferences so that NLP systems would generate personalized unbiased outputs. To this end, we build on the work of Alhafni et al. (2022b) where they formally introduced the task of gender rewriting and developed a user-centric gender rewriting model for Arabic.[4] We introduce the User-Aware Arabic Gender Rewriter, a user-centric web-based system for Arabic gender rewriting in contexts involving two users.[5] Our system takes either Arabic or English sentences as input, and provides users with the ability to specify their desired first and/or second person grammatical target genders. The system outputs gender rewritten alternatives of the Arabic input sentences (or their Arabic translations in case of English input) to match the target users' gender preferences. To the best of our knowledge, this is the first open-access web-based Arabic gender rewriting system.

Our goal behind creating an easy-to-use web-based multi-user Arabic gender rewriting tool is to enable users to rewrite any Arabic text based on their grammatical gender preferences that are consistent with their social identities. This reduces the gender bias that is caused by user-unaware NLP systems and increases the inclusiveness of Arabic NLP applications, leading to a better user experience. We envision a future in which websites such as those in Figure 2 could use automatic gender rewriting that fits the private preferences of their readers, or that is adjusted with simple website controls comparable to selecting different languages.

The rest of this paper is organized as follows. We discuss related work and Arabic linguistic facts in §2 and §3, respectively. We describe the design and implementation of the web-based Arabic gender rewriter in §4 and conclude in §5.

[3] Arabic HSB transliteration (Habash et al., 2007).

[4] https://github.com/CAMeL-Lab/gender-rewriting/

[5] http://gen-rewrite.camel-lab.com/

2 Related Work

Research has shown that NLP systems embed and amplify gender bias in a variety of core tasks such as machine translation (MT) (Rabinovich et al., 2017; Elaraby et al., 2018; Vanmassenhove et al., 2018; Escudé Font and Costa-jussà, 2019; Stanovsky et al., 2019; Costa-jussà and de Jorge, 2020; Gonen and Webster, 2020; Saunders and Byrne, 2020; Saunders et al., 2020; Stafanovičs et al., 2020; Savoldi et al., 2021; Ciora et al., 2021; Savoldi et al., 2022b; Savoldi et al., 2022a) and dialogue systems (Cercas Curry et al., 2020; Dinan et al., 2020; Liu et al., 2020a; Liu et al., 2020b; Sheng et al., 2021). Most existing solutions to mitigate gender bias in NLP systems either focus on debiasing pretrained representations used in downstream tasks (Bolukbasi et al., 2016; Zhao et al., 2018b; Manzini et al., 2019; Zhao et al., 2020) or on training systems on gender-balanced corpora (Lu et al., 2018; Rudinger et al., 2018; Zhao et al., 2018a; Hall Maudslay et al., 2019; Zmigrod et al., 2019).

More recently, text rewriting models were introduced to mitigate gender bias by either neutralizing the outputs of NLP systems or changing their grammatical genders to match provided users' gender preferences. Vanmassenhove et al. (2021) and Sun et al. (2021) presented rule-based and neural rewriting models to generate gender-neutral sentences in English. For morphologically rich languages and specifically Arabic, Habash et al. (2019) and Alhafni et al. (2020), introduced gender identification and rewriting models to rewrite first-person-singular Arabic sentences based on the target user gender requirements. The task of gender rewriting was formally introduced by Alhafni et al. (2022b) where they developed a new approach for Arabic gender rewriting in contexts involving two users (I and/or You) – first and second grammatical persons with independent grammatical gender preferences, and showed improvements over both Habash et al. (2019) and Alhafni et al. (2020) systems. The tool we introduce in this work uses the best gender rewriting model developed by Alhafni et al. (2022b).[4]

It is worth noting that our tool is similar to the recently introduced `Fairslator` (Měchura, 2022), a human-in-the-loop web-based tool for detecting and correcting gender bias in the output of MT systems translating from English to French, German, Czech, or Irish.[6] However, our work is different from theirs in the following ways:

- **Input**: our system takes either Arabic or English sentences as an input, whereas `Fairslator` only handles English sentences.
- **Models**: the Arabic gender rewriter relies internally on both rule-based and neural models as opposed to `Fairslator`'s rule-based gender reinflection system.
- **Evaluation**: the underlying gender rewriting model we use has been evaluated on Arabic gender rewriting and post-editing MT output, and it achieves state-of-the-art results, whereas `Fairslator` was not evaluated on any of the four languages it targets.
- **Visualization**: we focus on visualization by highlighting Arabic gender-marking words in both the input and the output to provide a better user-experience.

3 Arabic Linguistic Background

Arabic has a rich morphological system that inflects for gender, number, person, case, state, aspect, mood and voice, in addition to numerous attachable clitics (prepositions, particles, pronouns) (Habash, 2010). Arabic nouns, adjectives, and verbs inflect for gender: masculine (*M*) and feminine (*F*), and for number: singular (*S*), dual (*D*) and plural (*P*). Grammatical gender and number are commonly expressed using inflectional suffixes that represent some number and gender combination. Pronominal clitics also express gender and number combinations, e.g., طبيبتكم *Tbyb+km* 'your [masculine plural] doctor [feminine singular]. Gender and number participate in the morpho-syntactic agreement within specific constructions such as nouns and their adjectives and verbs and their subjects.

In practice, gender-specific words that are candidates for gender rewriting account for 10% of all words in all sentences and 17% of all words in gender-specific sentences. These statistics are calculated from the Arabic Parallel Gender Corpus (APGC) v2.1 (Alhafni et al., 2022a), which we use to train our models.

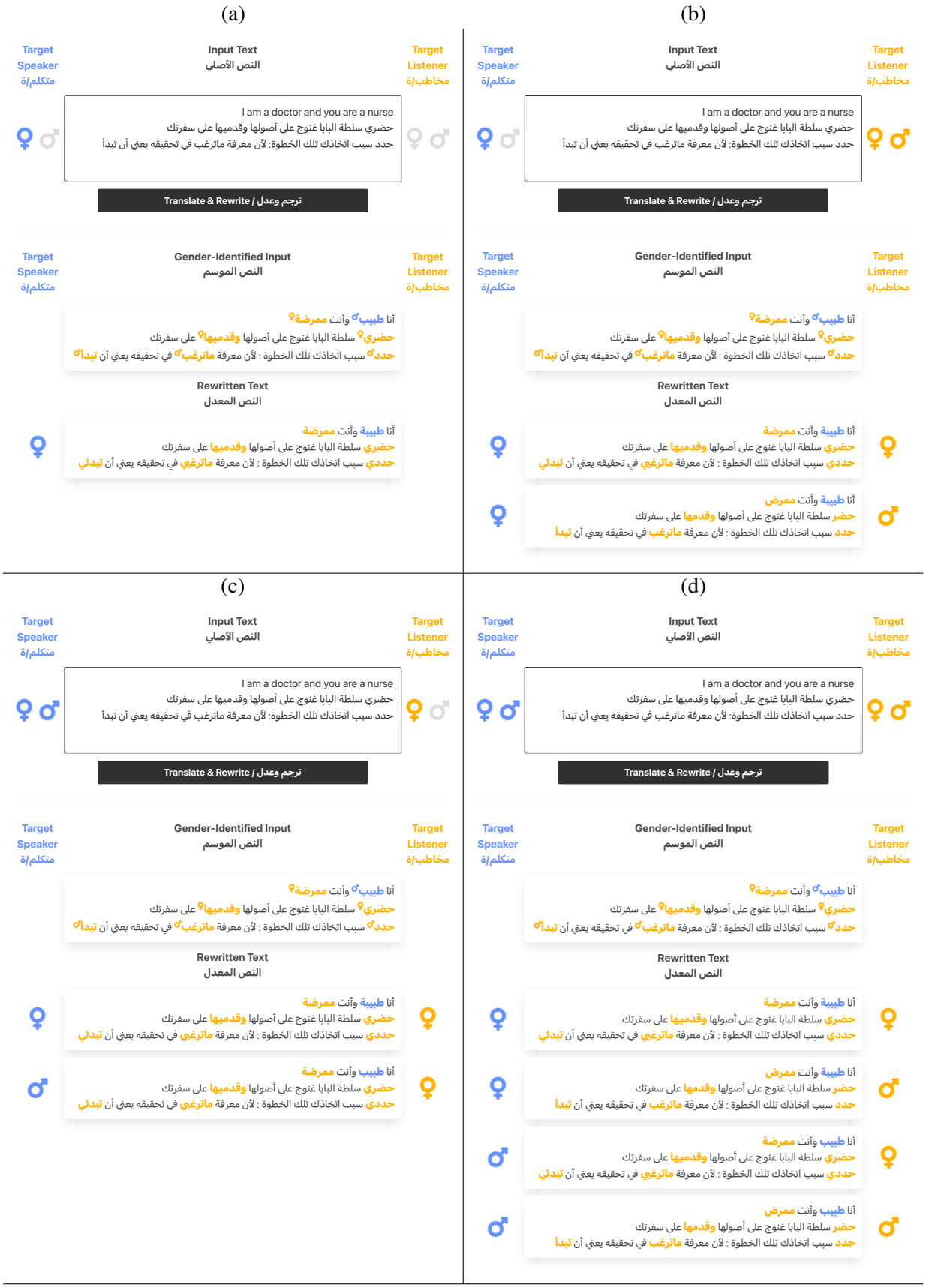

Figure 3: The Arabic Gender Rewriter interface showing gender rewritten alternatives of three input sentences in four modes: (a) Target speaker ♀ gender rewrites, (b) Target speaker ♀ and target listener ♀ and ♂ gender rewrites, (c) Target speaker ♀ and ♂ and target listener ♀ gender rewrites, and (d) Target speaker ♀ and ♂ and target listener ♀ and ♂ gender rewrites. Speaker gendered words are in **blue** and listener gendered words are in **orange**.

4 Design and Implementation

4.1 User Interface

Our gender rewriting interface is publicly available.[5] Figure 3(a) shows the basic structure of the interface. At the top, there is a text box to input either English or Arabic text. At each side of the text box, there are two selection buttons to indicate the desired target gender preferences for the speaker and the listener (♂ is for masculine and ♀ is for feminine). The user is able to select any possible combination of the desired target genders, including no target gender selection (i.e., requesting no rewriting).

Once the user clicks on the *Translate & Rewrite* button, all input English sentences will be passed to Google Translate's API to translate them into Arabic before generating their gender alternatives. When the gender rewriting process is done, additional text boxes will appear: the first text box will always contain the gender-identified Arabic inputs and the rest of the text boxes will contain the gender rewritten alternatives. Each gender marking word in the gender-identified input text box will be labeled as either masculine (♂) or feminine (♀). First-person (i.e., speaker) gendered words are colored in **blue** and second-person (i.e., listener) gendered words are colored in **orange**.

The number of the text boxes containing the gender rewritten alternatives is based on the selected target gender preferences. Each one of those boxes will have a label at its sides indicating a particular target gender combination based on the users' selections. For instance, Figure 3(a) has one text box containing first-person feminine gendered alternatives of the input sentences. We discuss the screenshots in Figure 3 in more details in §4.2.

Front-end The front-end was implemented using `Preact`[7] for view control and `Bulma`[8] for styling.

Back-end The back-end was implemented in Python using `Flask` to create a web API wrapper for the gender rewriting model.[9] We use the best performing gender rewriting model described in Alhafni et al. (2022b). The model was trained on the APGC v2.1 in addition to augmented data from the OpenSubtitles 2018 dataset (Lison and Tiedemann, 2016) and it consists of three components: gender identification, out-of-context word gender rewriting, and in-context ranking and selection.

The gender identification component identifies the word-level gender label for each word in the input sentence. It leverages a word-level BERT-based (Devlin et al., 2019) classifier that was built by fine-tuning CAMeLBERT MSA (Inoue et al., 2021). Once the gender labels have been identified for each word in the input and given the desired users target genders, out-of-context word gender rewriting is triggered based on the compatibility between the provided users' target genders and the predicted word-level gender labels. The gender rewriting component employs three word-level gender alternative generation models in a backoff cascade setup: 1) Corpus-based Rewriter: a bigram maximum likelihood estimation lookup model; 2) Morphological Rewriter: a morphological analyzer and generator provided by CAMeL Tools (Obeid et al., 2020); and 3) Neural Rewriter: a character-level sequence-to-sequence model with side constraints (Sennrich et al., 2016). Since the three implemented word-level gender rewriting models are out of context and given Arabic's morphological richness, this leads to producing multiple candidate gender alternative sentences. To select the best candidate output sentence, we rank all candidates in full sentential context based on their pseudo-log-likelihood scores (Salazar et al., 2020).

Results As we previously reported in Alhafni et al. (2022b), the results on the test set of APGC v2.1 show that the best gender rewriting model achieves an $\mathbf{M^2}$ $\mathbf{F_{0.5}}$ (Dahlmeier and Ng, 2012) score of **88.4** and an average of **1.2 BLEU** (Papineni et al., 2002) increase when automatically post-editing Google Translate's output.

4.2 Examples and Use Cases

Figure 3 presents the different outputs of the gender rewriting tool for three input sentences, one in English and two in Arabic. The three sentences come from the examples presented in Figure 1, Figure 2(a), and Figure 2(b), respectively.

In Figure 3(a), only the feminine target gender for the speaker is selected by the user. In this case, the system performs gender identification and then generates the first-person feminine gender alternative of the input sentences where all first-person masculine words are rewritten to feminine. Fig-

[6] `https://www.fairslator.com/`
[7] `https://preactjs.com/`
[8] `https://bulma.io/`
[9] `http://flask.pocoo.org/`

ure 3(b) shows an example where the feminine target gender for the speaker, and both the feminine and the masculine target genders for the listener are selected. In this case, the system outputs two gender rewritten alternatives for each input sentence, one for each selected target gender combination (i.e., speaker feminine – listener feminine, speaker feminine – listener masculine). Similarly, Figure 3(c) shows an example where both the feminine and the masculine target genders for the speaker, and the feminine target gender for the listener are selected. Lastly, Figure 3(d) is where all the target gender preferences are selected for both the speaker and the listener. In this case, the system generates all four possible gender rewritten alternatives for each input sentence.

5 Conclusion and Future Work

We introduced the User-Aware Arabic Gender Rewriter, a user-centric web-based system for Arabic gender rewriting in contexts involving two users. Our system takes either Arabic or English sentences as input, and provides users with the ability to specify their desired first and/or second persons target genders. The system outputs gender rewritten alternatives of the Arabic input sentences (or their Arabic translations in case of English input) to match the target users' gender preferences. Moreover, the system highlights Arabic gender-marking words in both the input and the output to provide a better user-experience.

In future work, we plan to continue improving our gender rewriting back-end by adding better gender rewriting models and enhancing inference efficiency, as well as expanding gender identification and rewriting to third person entities. We also plan to improve the interface by enabling users to provide feedback that can be collected and used to enhance the performance of gender rewriting. We will also improve the visualization we use to highlight Arabic gender marking words by examining the added value it provides to different end users, from language learners to native text editors.

References

Ala'uldeen, Rola. 2022. Gender and Arabic (in Arabic). Kohl Journal Vol 8. `https://kohljournal.press/node/346`.

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2022a. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France, June. European Language Resources Association.

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2022b. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States, July. Association for Computational Linguistics.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. 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. Association for Computational Linguistics.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Boyd, Danah and Kate Crawford. 2012. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679.

Cercas Curry, Amanda, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Ciora, Chloe, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK, August. Association for Computational Linguistics.

Costa-jussà, Marta R. and Adrià de Jorge. 2020. 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. Association for Computational Linguistics.

Dahlmeier, Daniel 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, June.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dinan, Emily, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, November. Association for Computational Linguistics.

Elaraby, Mostafa, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Escudé Font, Joel and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August. Association for Computational Linguistics.

Gonen, Hila and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online, November. Association for Computational Linguistics.

Habash, Nizar, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In van den Bosch, A. and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Habash, Nizar, 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, August.

Habash, Nizar Y. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Hall Maudslay, Rowan, 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, November.

Inoue, Go, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Johnson, Melvin. 2020. A scalable approach to reducing gender bias in google translate. Google AI Blog.

Kuczmarski, James. 2018. Reducing gender bias in google translate. Google Blog.

Lison, Pierre and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Liu, Haochen, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Liu, Haochen, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online, November. Association for Computational Linguistics.

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

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

Manzini, Thomas, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Měchura, Michal. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington, July. Association for Computational Linguistics.

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

Obeid, Ossama, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emr Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Rabinovich, Ella, 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, April.

Rudinger, Rachel, 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, June.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July. Association for Computational Linguistics.

Saunders, Danielle and Bill Byrne. 2020. 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. Association for Computational Linguistics.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 08.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022a. On the dynamics of gender learning in speech translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111, Seattle, Washington, July. Association for Computational Linguistics.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022b. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland, May. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online, June. Association for Computational Linguistics.

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

Stanovsky, Gabriel, 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, July.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Sun, Tony, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english.

UN. 2018. Guidelines for gender-inclusive language in Arabic. `https://www.un.org/ar/gender-inclusive-language/guidelines.shtml`.

Vanmassenhove, Eva, 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, October-November.

Vanmassenhove, Eva, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zhao, Jieyu, 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, June.

Zhao, Jieyu, 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, October-November.

Zhao, Jieyu, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July. Association for Computational Linguistics.

Zmigrod, Ran, Sabrina 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, July.