NeuTral Rewriter

Vanmassenhove, Eva; Emmery, Chris; Shterionov, Dimitar

DOI:
10.18653/v1/2021.emnlp-main.704

Publication date:
2021

NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives

Eva Vanmassenhove and **Chris Emmery** and **Dimitar Shterionov**
Department of Cognitive Science and Artificial Intelligence
Tilburg University
The Netherlands
e.o.j.vanmassenhove@tilburguniversity.edu
c.d.emmery@tilburguniversity.edu
d.shterionov@tilburguniversity.edu

Abstract

Recent years have seen an increasing need for gender-neutral and inclusive language. Within the field of NLP, there are various mono- and bilingual use cases where gender inclusive language is appropriate, if not preferred due to ambiguity or uncertainty in terms of the gender of referents. In this work, we present a rule-based and a neural approach to gender-neutral rewriting for English along with manually curated synthetic data (*WinoBias+*) and natural data (OpenSubtitles and Reddit) benchmarks. A detailed manual and automatic evaluation highlights how our NeuTral Rewriter, trained on data generated by the rule-based approach, obtains word error rates (WER) below 0.18% on synthetic, in-domain and out-domain test sets.

1 Introduction

Recent years have seen an increasing need for gender-neutral and inclusive language. This need is reflected, among others, by a surge in the use of *singular they*, [1] currently endorsed as part of APA style as the generic and gender-neutral pronoun. [2] Within the field of Natural Language Processing (NLP), there are various monolingual and bilingual use cases where gender neutral and inclusive language is appropriate, if not preferred due to e.g. ambiguity in terms of the gender of referents. Section 3 provides a short outline of potential NLP use cases.

To support these use cases, we present a rule-based and a neural approach to gender-neutral rewriting along with manually curated benchmarks, both of which we provide open-access/source. [3] First, a rule-based rewriter is implemented leveraging hand-written rules and an automatic error correction tool. Next, a neural rewriter is trained on output generated by the rule-based rewriter to remove the need for extensive pre-processing and the reliance on computationally expensive tools such as dependency parsers. Our manual and automatic evaluation show how the neural rewriter clearly improves over the rule-based approach with word error rates (WER) below 0.18% on synthetic, in-domain and out-domain test sets.

The main contributions of our work can be summarized as follows : (i) WinoBias+, an open-source manually curated extension of WinoBias (Zhao et al., 2018a) providing neutral alternatives for 3,167 sentences as well as a manually curated set of 1,000 natural sentences (domain : Reddit, OpenSubtitles), (ii) open-source code for rule-based and neural neutral rewriters which can convert (binary) gendered English sentences into their gender neutral counterparts, (iii) a detailed manual and automatic evaluation of errors made by the rule-based and neutral rewriter on synthetic and natural data.

2 Related Work

Recent years have seen an increase in research on gender and gender bias mitigation in NLP. While a relatively large body of research has focused on debiasing word embeddings (e.g., Bolukbasi et al., 2016; Font and Costa-jussà, 2019; Zhao et al., 2018c), our work is related to the generation of gender variants. We broadly distinguish between : (i) approaches that incorporate additional (meta-) information during training/testing allowing for a controlled generation of gender alternatives, and (ii) approaches that focus on gender rewriting. The synopsis will focus specifically on research related to the gender of human referents.

Within the field of Machine Translation (MT), Vanmassenhove and Hardmeier (2018); Vanmassenhove et al. (2019), and Basta et al. (2020) in-

1. The pronoun 'they' was announced word of the year in 2019 according to Merriam Webster https://www.nytimes.com/2019/12/10/us/merriam-webster-they-word-year.html
2. https://apastyle.apa.org/
3. https://github.com/anonymous-until-publication/NeuTralRewriter

corporate meta-information in the form of gender tags on the source side to enable gender alternative target translations for ambiguous source sentences. Moryossef et al. (2019) propose a black-box approach by appending gender information to the target sentences using parataxis constructions at translation time. Bau et al. (2019) describe work on controlling linguistic features (a.o. gender) in Neural MT by identifying and (de)activating the relevant neurons. They show that gender is the most difficult feature to control with a success rate of 21% using the top five identified neurons.

Lu et al. (2020) uses a Counterfactual Data Augmentation (CDA) technique to augment data sets by creating gender alternative sentences to decrease gender bias. Their approach consists of swapping gendered words with their male/female counterparts (e.g. he:she, father:mother...). Their results indicate that a CDA approach outperforms a simple word embedding debiasing technique (Bolukbasi et al., 2016). Habash et al. (2019) and Alhafni et al. (2020) present gender-aware reinflection models for Arabic. Using an Arabic sentence and a target gender, the desired gender alternative is generated by re-inflecting the input.

It is worth noting that all the previously described approaches focus on generating binary (female/male) gendered alternatives or translations, while our work focuses on generating gender-neutral alternatives. As such, the work that is most closely related to ours is Sun et al. (2021). Their work is contemporaneous to our submission.[4] Sun et al. (2021) present a rule-based and neural rewriter for the generation of gender-neutral *singular they* sentences as well as an evaluation benchmark[5] of 500 parallel sentences (gendered and gender-neutral) from five domains (Twitter, Reddit, movie quotes, jokes). Their rule-based and neural rewriters are able to generate gender-neutral sentences with an error-rate below 1% (0.63% and 0.99% respectively). In terms of resources, compared to Sun et al. (2021), we provide larger synthetic and natural benchmarks. In terms of performance, although complicated due to the lack of a publicly available benchmark, our models are seemingly better with error-rates of 0.52 (rule-based) and 0.02 (neural) on the most comparable benchmark, i.e. Reddit data.

4. Currently in $arxiv$ pre-print.

5. We contacted the authors to obtain their benchmark for comparison as it is currently not open-source, but have not been able to obtain it yet. We will nevertheless attempt to compare our result to theirs to the best of our ability.

3 Use Cases

Generating neutral alternatives for gendered sentences has applications for various monolingual language generation tasks (e.g. automatic responses), where (i) one does not want to assume the gender of the referents, or (ii) one wants to present the user with various options. Similarly, in a bilingual setting, more specifically for MT, a neutral rewriter allows for the generation of gender neutral alternatives for genderless and gender-neutral source languages (Hungarian, Turkish, Persian, Swahili...) or null-subject source languages (Spanish, Chinese, Arabic, Bulgarian...). For illustration, Example (1) and (2) demonstrate how gender-neutral alternatives can be useful in bilingual settings. Example (1) features a sentence in Armenian using the epicene (gender-neutral) pronoun '*Նա*' which can be either translated into 'he', 'she' or singular 'they'.

(1) HY : Նա բացեց դուռը
EN : **He**/She/They **opened the door.**[6]

Similarly, Example (2) illustrates the possible translations of a null-subject source in Spanish which can be translated as "works in a company".

(2) ES : Trabaja en una empresa.
EN : **He**/She **works in a company.**[7]
EN : They work in a company.

As a pre-processing step, rewriting into neutral alternatives could be useful to debias training data and thereby its embeddings (see a.o., Bolukbasi et al., 2016; Li et al., 2018; Gonen and Goldberg, 2019) and/or to obfuscate sensitive 'gender' features from real user data facing automatic profiling systems (Reddy and Knight, 2016; Shetty et al., 2018; Emmery et al., 2021).

4 Methodology & Experimental Setup

4.1 Datasets

All data is preprocessed using the Moses (de)tokenizer (Koehn et al., 2007). Training (Reddit) and test sets (WinoBias+, OpenSubtitles, Reddit) contain a balanced amount of the eight (binary) target pronouns/determiners : *he*, *she*, *her(s)*, *his*, *him*, *him/herself*.[8]

6. The translation in bold is the only one provided by Bing and Google Translate consulted on May 4, 2021.

7. The translation in bold is the only one provided by Bing, Google Translate and DeepL consulted on May 4, 2021.

8. For a set containing X sentences, we extracted at least $X/8$ sentences containing each form - a completely uniform

Reddit A set of 2,259,386 sentences (containing a total of 3M pronouns/determiners) was randomly sampled from Pushshift's Reddit snapshots (Baumgartner et al., 2020, including all subreddits) for the period of July–December 2019. This set we would later use for training our neural rewriter. Another set of 1,693 sentences (containing a total of 2K pronouns/determiners) extracted from Reddit in the same way would later be used as a development set. There are no overlaps between the two sets.

WinoBias+ an extension of the WinoBias benchmark, providing (manual) neutral alternatives for its 3,167 synthetic sentences, and corrections (e.g. for ungrammatical sentences [9]) of the original dataset.

OpenSubtitles, Reddit test additional sets of 1,000 (manually corrected) parallel sentences (500 for each set). The entire cleaned and extended version of the corpus—*WinoBias+*— the OpenSubtitles (Lison and Tiedemann, 2016), and Reddit benchmark is made publicly available under a CC BY-SA 4.0 [10] license. [11]

4.2 Rule-Based Rewriter

The rule-based rewriter (RBR), consists of two main components : (i) a rule-based pronoun rewriter, and (ii) an error-correction language model.

4.2.1 Rule-Based Pronoun Rewriter

Table 1 gives an overview of the binary forms and their gender-neutral alternatives. While most mappings are one-to-one, '*her*' can be either a pronoun (e.g. 'I gave it to her.' → 'I gave it to them.') or a possessive determiner (e.g. 'It is her book.' → 'It is their book') and '*his*' can be either a possessive determiner ('It is his book.' → 'It is their book') or an independent possessive pronoun ('The book is his.' → 'The book is theirs'). To disambiguate these forms, the POS tagger and dependency parser from Stanza (Qi et al., 2020) were used. [12]

binary	→	gender-neutral
he, she	→	they
him	→	them
her	→	them, their
his	→	their, theirs
hers	→	theirs
him/herself	→	themselves [13]

TABLE 1 – Mapping binary pronouns/determiners to their gender-neutral alternatives.

3^{rd} person	→	plural
works	→	work
has	→	have
is	→	are

TABLE 2 – Subject-verb agreement correction examples.

Following the guidelines from the European Parliament for gender neutral language [14], we provide an option to change gendered English animate nouns ('chair(wo)man' → 'chairperson', 'bar(wo)man' → 'bartender'...), unnecessary feminine forms of animate nouns (e.g. 'actress' → 'actor', 'heroine' → 'hero'...), and generic uses of 'man' (e.g. 'freshman'→'first-year student', 'man-made' → 'human-made'...). [15]

4.2.2 Subject-Verb Agreement Correction

The nominative pronouns (*he* and *she*) can be replaced by *they*. However, if they are in agreement with a simple present tense verb (or the verb 'to be') the 3^{rd} person form/ending should be replaced by a plural one (see Table 2). To address this, we used a Python wrapper for LanguageTool, an open-source grammar, style and spell corrector. [16] We limited the correction to grammar mistakes to avoid additional changes (e.g. insertion of commas, different word choices, removal of whitespaces...).

4.3 Neural Rewriter

We trained a Transformer model (Vaswani et al., 2017) using FAIRSEQ (Ott et al., 2019)—following the setup of (Sun et al., 2021) for comparison. For training we used the 2,259,386 Reddit sentences as source and their gender-neutral alternatives as

distribution was not achievable due to the fact that multiple pronouns/determiners can be present in a single sentence.

9. For example, the original WinoBias sentence "The laborer handed the application to the editor because she *want* the job." is corrected into "The laborer handed the application to the editor because she *wanted* the job."

10. `https://creativecommons.org/licenses/by-sa/4.0/`

11. `https://github.com/vnmssnhv/NeuTralRewritter`

12. The 'his' ambiguity can only be resolved using a dependency parser since the xpos and upos tags do not differ when 'his' is used as a independent or dependent possessive.

13. 'Themselves' is preferred over 'themself' according the APA guidelines : `https://apastyle.apa.org/style-grammar-guidelines/grammar/singular-they`

14. `https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf`

15. The complete list of nouns included can be found in the appendix.

16. `https://pypi.org/project/language-tool-python/`

Error Classes		Rule-Based			Neural		
		WB+	OpenS	Red	WB+	OpenS	Red
LM	SVA	9	16	11	0	5	0
	corr.	0	0	11	0	0	0
	‘s (has)	0	1	7	0	1	0
	space	0	0	3	0	0	4
	other	0	0	3	0	0	0
POS	error	12	0	3	0	0	0
	source	0	0	2	0	0	0
OTH.	cap.	0	4	2	0	1	0
	ungram.	0	2	0	0	0	0
	rule	0	1	1	0	1	0
	UNK	0	0	0	0	0	2
# of errors		21	24	43	0	8	6

TABLE 3 – Error classification and counts on the WinoBias+, OpenSubtitles and Reddit test set for the Rule-Based and Neural approach.

target; for validation we used the 1,693 Reddit sentences and their neutral alternatives (see Section 4.1). The gender-neutral alternatives, i.e. the target sides, are generated by applying the RBR on the original dataset. All hyperparameters and their values are listed in the Appendix along with the preprocessing and training commands and options.

5 Results & Discussion

Both rewriters were (manually) evaluated on synthetic (WinoBias+) and natural (Reddit, OpenSubs) evaluation benchmarks.

5.1 Manual Evaluation

Table 3 presents a detailed overview of the errors per test set for the Rule-Based and Neural approach. An overview and explanation of all error labels can be found in the Appendix.

Rule-Based Approach The errors can be divided broadly into “language model” (LM), “postag” (POS) and “other” errors. WinoBias+ consists of 3167 sentences. Only 21 of the synthetic sentences were rewritten incorrectly. Issues arose either due to incorrect disambiguation (‘her’ → ‘them’ (pronoun) instead of ‘their’ (determiner)) or due to incorrect subject-verb agreement (SVA).

The RBR struggled more with the noisy, often ungrammatical, natural data from OpenSubtitles and Reddit. The main issues observed are incorrect SVA, additional corrections by the language tool (unrelated to gender neutrality, e.g. *cause* → *because*) and incorrect disambiguation of "‘s". [17]

17. e.g. *He’s worked.* → *They are worked.* instead of *They have worked.*

WER (%)	WB+	OpenS	Reddit	Sun et al. (2021)
BASE	8.76	14.09	11.02	12.40
RBR	0.06	0.45	0.52	0.63
NR	**0.00**	**0.18**	**0.02**	0.99

TABLE 4 – WER on the synthetic WinoBias+ (WB+) test set and natural Reddit and OpenSubtitles benchmark vs WER obtained by Sun et al. (2021).

Neural Approach Interestingly, and in contrast with the findings described in Sun et al. (2021), our neural model trained on the rule-based generated training data, outperforms the rule-based approach. The error analysis reveals that the neural model resolves many of the longer distance SVA issues, the disambiguation of “’s” and errors that occurred due to incorrect postags.

No errors were made on the synthetic WinoBias+ data. Errors on the in-domain Reddit data were due to the removal of additional spaces (4 errors) or because of an unknown character/emoji (2 errors). On the out-of-domain OpenSubtitles set, we noted 8 errors the majority of which due to incorrect SVA (5 errors).

5.2 Automatic Evaluation

For comparison, we employed the same metric as Sun et al. (2021) : WER. A combination of the baseline WER (indicating the amount of changes needed in order to change to gender-neutral alternatives), and the WER computed between the correct neutral forms and the automatically generated forms provides insights into the performance of both approaches.

Given that Sun et al. (2021) use an evaluation benchmark of 500 sentences consisting of Twitter, Reddit, jokes and movie quotes data, its performance is probably most comparable to the scores we obtained on the Reddit set. Like the manual evaluation, and in contrast with Sun et al. (2021), the automatic evaluation (Table 4) confirms that our neural approach is able to generalize over the rule-based generated data, outperforming it with error rates below 0.18% (0.0% (WB+), 0.18% (OpenSubtitles) and 0.02% (Reddit)). Furthermore, these error rates are all substantially lower than those reported by Sun et al. (2021). We hypothesize this is due to the better performance of the RBR (confirmed as well by the automatic/manual evaluation) leading to better source (gendered)–target (neutral) training data for the NMT model.

We ought to note that WER does not take into account the removal of superfluous spaces (e.g. before the first character of a sentence, double spaces instead of a single one). We only observed the removal of such spaces by the neural rewriter on the Reddit data (see detailed manual analysis presented in Table 3).

6 Conclusion

This paper presents a rule-based and a neural gender-neutral rewriter for English. First, the rule-based approach was implemented, leveraging hand-written rules and an automatic error correction tool. Using the RBR, we generated a parallel gendered-to-neutral corpus on which an NMT system was trained. The NMT model removes the need for computationally expensive pre-processing steps and, according to the manual and automatic evaluation, outperforms the RBR on synthetic, in-domain and out-domain benchmarks. Along with our open-access/source code, we also provide three manually curated benchmarks for neutral rewriting.

For now, the neutral rewriter is limited to English using 'singular they' and recommendations for gender neutral writing specific to the English language. It is, in theory, possible to extend this approach (or a similar one) to other languages. However, so far, few languages have a crystallized approach when it comes to gender-neutral pronouns and gender-neutral word endings.

In future work, we intend to explore potential applications of the neutral rewriters (e.g. gender debiasing of corpora). We furthermore plan to extend our work to gender-neutral rewriting targeting specific referents within a sentence to accommodate the gender preferences of individual referents.

Ethics statement

Neutral Rewriter Application The Neutral Rewriter is intended to provide gender-neutral alternatives and increase the inclusiveness of NLP/MT applications. The rewriter can furthermore be used as a preprocessing step to obfuscate a potentially sensitive gender attribute from training data.

At this stage, the rewriter works on a sentence-level and does not allow for rewriting pronouns or determiners of specific referents. We followed the guidelines of the European Parliament for gender neutral language and provide an option to change gendered animate nouns, unnecessary feminine forms of animate nouns and generic uses of the word 'man' based on non-exhaustive word lists.

Datasets We present three openly available English benchmarks : (i) WinoBias+, (ii) OpenSubtitles and (iii) Reddit. (i) WinoBias+ consists of a curated and extended version of the synthetic WinoBias (Zhao et al., 2018b) dataset, distributed under the MIT License. [18] (ii) The open-source OpenSubtitles (Lison and Tiedemann, 2016) [19] data was used to randomly sample a subset for the OpenSubtitles benchmark. OpenSubtitles is distributed under a Creative Commons license. [20] (iii) The Reddit dataset was collected through the third-party snapshots of Reddit's publicly available API at https://pushshift.io. It is subject to Reddit's own User Agreement and Privacy Policy and covers the *free and public sharing of user data*. [21]

The neutral alternatives for the three benchmarks were manually created by a linguist. The curation rationale behind the selected datasets is summarized as follows. WinoBias was selected as it is one of the few benchmarks for gender bias in NLP. We extended it with gender-neutral alternatives. The natural Reddit and OpenSubtitles dataset allowed us to verify the robustness of the rewriters on more noisy and diverse data sets. The OpenSubtitles and Reddit datasets contain variety in terms of language and English social dialects. Training and test sets contain a balanced amount of the eight (binary) target pronouns/determiners. For a set containing X sentences, we extracted at least $X/8$ sentences containing each form - a completely uniform distribution was not achievable due to the fact that multiple pronouns/determiners can be present in a single sentence.

Carbon statement The neural model presented in this work has an ecological footprint equivalent to 1.68kg of CO2 emissions. [22] The training time, consumption and carbon emission can be found in Table 5.

Acknowledgements

We would like to thank the reviewers for their insightful feedback and comments.

18. https://opensource.org/licenses/MIT
19. http://www.opensubtitles.org/
20. Attribution-Non Commercial 4.0 International
21. See https://www.redditinc.com/policies/user-agreement and https://www.redditinc.com/policies/privacy-policy respectively.
22. Contribution based on GPU power consumption at training the NeuTral rewriter model.

Elapsed time (h)	Avg. power draw	kWh	CO2 (kg)
6.2	147.37	2.64	1.68 ±0.13

TABLE 5 – Train time, consumption and carbon emissions related to the training of the NeuTral rewriter.

References

Bashar Alhafni, 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.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2020. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, pages 1–14.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and Controlling Important Neurons in Neural Machine Translation. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*, New Orleans, USA.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Tolga Bolukbasi, 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 *Proceedings of Thirtieth Conference on Neural Information Processing Systems (NIPS)*, pages 4349–4357, Barcelona, Spain.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild : Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 2388–2402.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. In To appear in *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*, Florence, Italy.

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. 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 609–614.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 25–30.

Pierre Lison 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.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In TO APPEAR IN *1st ACL Workshop on Gender Bias for Natural Language Processing*, Florence, Italy.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza : A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 101–108.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1 : Long Papers*. The Association for Computer Linguistics.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4nt : author attribute anonymity by adversarial training of neural machine translation. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1633–1650.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs : Rewriting with gender-neutral english. *arXiv preprint arXiv :2102.06788*.

Eva Vanmassenhove and Christian Hardmeier. 2018. Europarl datasets with demographic speaker information. In *Proceedings of the 21st Annual Conference of the European Associations for Machine Translation (EAMT)*, Alicante, Spain.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018b. 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. 2018c. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4847–4853, Brussels, Belgium.

A Appendix

The appendix provides additional information on generation of gender-neutral alternatives (Section A.1), the error labels and analysis (Section A.2) and the training hyperparameters of the Neural Machine Translation model (Section A.3.1).

A.1 Advanced Rewriter

The advanced rewriter includes rewriting of gender-marked job titles (chairman, anchorman...), rewriting of unnecessary feminine forms (actress, comedienne, waitress...), avoidance of construction using a generic form of 'man' ('average man', 'man and wife'...), and rewriting of titles ('Mrs' and 'Miss').

A.1.1 Gender-neutral alternatives for gender-marked job titles

chairman/woman		businessman/woman	
chairman	→ chairperson	businessman	→ business person
chairmen	→ chairpeople	businessmen	→ business people
chairwoman	→ chairperson	businesswoman	→ business person
chairwomen	→ chairpeople	businesswomen	→ business people
anchorman/woman		**postman/postwoman**	
anchorman	→ anchor	postman	→ mail carrier
anchormen	→ anchors	postmen	→ mail carriers
anchorwoman	→ anchor	postwoman	→ mail carrier
anchorwomen	→ anchors	postwomen	→ mail carriers
congresswoman/congressman		**mailman/mailwoman**	
congressman	→ member of congress	mailman	→ mail carrier
congressmen	→ members of congress	mailmen	→ mail carriers
congresswoman	→ member of congress	mailwoman	→ mail carrier
congresswomen	→ members of congress	mailwomen	→ mail carriers
policeman/policewoman		**salesman/saleswoman**	
policeman	→ police officer	salesman	→ salesperson
policemen	→ police officers	salesmen	→ salespersons
policewoman	→ police officer	saleswoman	→ salesperson
policewomen	→ police officers	saleswomen	→ salespersons
spokesman/woman		**fireman/firewoman**	
spokesman	→ spokesperson	fireman	→ firefighter
spokesmen	→ spokespersons	firemen	→ firefighters
spokeswoman	→ spokesperson	firewoman	→ firefighter
spokeswomen	→ spokespersons	firewomen	→ firefighters
steward/stewardess		**barman/barwoman**	
steward	→ flight attendant	barman	→ bartender
stewards	→ flight attendants	barmen	→ bartenders
stewardess	→ flight attendant	barwoman	→ bartender
stewardesses	→ flight attendants	barwomen	→ bartenders
headmaster/mistress		**cleaning man/lady**	
headmaster	→ principal	cleaning man	→ cleaner
headmasters	→ principals	cleaning lady	→ cleaners
headmistress	→ principal	cleaning men	→ cleaner
headmistresses	→ principals	cleaning ladies	→ cleaners
foreman/forewoman			
foreman	→ supervisor		
foremen	→ supervisors		
forewoman	→ supervisor		
forewomen	→ supervisors		

TABLE 6 – Gender-neutral alternatives for gender-marked job titles

A.1.2 Gender-neutral alternatives for unnecessary feminine forms

actress		usherette	
actress	→ actor	usherette	→ usher
actresses	→ actors	usherettes	→ usher
heroine		**authoress**	
heroine	→ hero	authoress	→ author
heroine	→ heroes	authoresses	→ authors
comedienne		**mailman/mailwoman**	
comedienne	→ comedian	mailman	→ mail carrier
comediennes	→ comedians	mailwomen	→ mail carriers
executrix		**boss lady**	
executrix	→ executor	boss lady	→ boss
executrices	→ executors	boss ladies	→ boss
executrixes	→ executor		
poetess		**waitress**	
poetess	→ poet	waitress	→ waiter
poetesses	→ poets	waitresses	→ waiters

TABLE 7 – Gender-neutral alternatives for unnecessary feminine forms

A.1.3 Gender-neutral alternatives for generic 'man'

average man		layman	
average man	→ average person	layman	→ layperson
average men	→ average people	laymen	→ laypeople
best man for the job		**freshman**	
best man for the job	→ best person for the job	freshman	→ first-year student
best men for the job	→ best people for the job	freshmen	→ first-year students
mankind		**man-made**	
mankind	→ humankind	man-made	→ human-made
workmanlike		**man and wife**	
workmanline	→ skillful	man and wife	→ husband and wife

TABLE 8 – Gender-neutral alternatives for generic 'man'

A.2 Overview Error Analysis

A.2.1 Error Classification Rewriter

As explained in the paper, errors are divided into Language Model (LM) errors, postag error (POS) and other errors (OTHER). Within these three error classes, we identified multiple subclasses of LM, POS and OTHER errors. An explanation of the labels used in our error analysis and paper can be found in Table 9. Table 10 provides example input and output sentences.

Error Label	Explanation
LM 's	Wrongly disambiguated the contracted form 's as a verb form of 'to be' instead of 'to have'
LM space	Space added or removed by rewriter
LM correction (corr.)	Error correction done by rewriter (language tool) that is not related to gender-neutral rewriting
LM subject-verb agreement (SVA)	Failure to make correct subject-verb agreement, usually due to long distance dependencies.
POS	Wrong form of 'they' produced by rewriter due to incorrect postag
POS (source)	Wrong form of 'they' produced by rewriter due to incorrect postag which is related to an ungrammatical/incorrect soure sentences
OTHER rule	Some forms such as 'hisn's' are not standard language and does not covered by our rules. Similarly written forms such as 'hes' for 'he's' are not corrected by the rewriter
Other ungram.	Ungrammatical input sentence leading to an ungrammatical output
Other UNK	The Neural Rewriter outputs <UNK> for unknown characters (in our case "?", "!", "...", and emojis/special characters that did not appear in our Reddit training data)

TABLE 9 – Error label explanation

A.3 Neural Rewriter

Our neural model is trained with the following options : `transformer-iwslt-en-de` architecture with 4 attention heads and encoder and decoder embedding dimensions equal to 512, encoder and decoder embedding dimensions for the FFN equal to 1024, Adam learning optimizer (Kingma and Ba, 2015) with a learning rate of 0.005 and inverse square-root schedule with 4 000 warmup steps, an early stopping based on the improvement on the validation set with patience 5, dropout of 0.3, joint byte-pair encoding (Sennrich et al., 2016) with 32 000 operations, token-based batches with maximum size of 4096. For ease of replicability we provide our complete preprocessing and training scripts in Appendix.

A.3.1 Training Hyperparameters

```
fairseq-preprocess --source-lang $SRC \
  --target-lang $TRG \
  --trainpref $ENGDIR/data/train.tc.bpe \
  --validpref $ENGDIR/data/dev.tc.bpe \
  --testpref $ENGDIR/data/test.tc.bpe \
  --destdir $ENGDIR/data/ready_to_train

fairseq-train $ENGDIR/data/train_data \
  --arch transformer_iwslt_de_en \
  --lr 0.0005 --optimizer adam \
  --adam-betas '(0.9, 0.98)' \
  --max-tokens 4096 \
  --dropout 0.3 \
  --update-freq=1 \
  --lr-scheduler inverse_sqrt \
  --warmup-init-lr 1e-07 --min-lr 1e-09 \
  --warmup-updates 4000 \
  --save-dir $ENGDIR/model \
  --skip-invalid-size-inputs-valid-test \
  --patience 5
```

Error Label	Example	→	Output RBR
LM ('s)	He**'s** worked hard	→	They **are** worked hard.
LM (space)	... aren 't...	→	...aren't...
LM (corr.)	Bit pricey...	→	**A** bit pricey...
LM (SVA)	He works and works...	→	They work and **works**...
POS	He saw her run fast...	→	They saw **their** run fast...
POS (source)	...looked at **her weird** (she 's close..	→	... looked at **their weird** (they are close...
Basic rule	She's **hisn's**..	→	.They are **hisn's**
Other	Where's herself.	→	Where's themselves.

TABLE 10 – Examples Error Labels

With `$ENGDIR` we indicate the path where the data folder and the model folder are located.