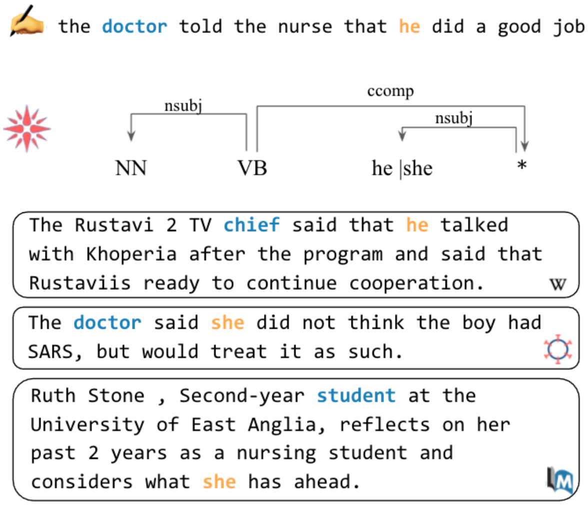
Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation

Shahar Levy **Koren Lazar** **Gabriel Stanovsky**
School of Computer Science and Engineering
The Hebrew University of Jerusalem, Jerusalem, Israel
shaharl6000@gmail.com {koren.lazar, gabriel.stanovsky}@mail.huji.ac.il

Abstract

Recent works have found evidence of gender bias in models of machine translation and coreference resolution using mostly synthetic diagnostic datasets. While these quantify bias in a controlled experiment, they often do so on a small scale and consist mostly of artificial, out-of-distribution sentences. In this work, we find grammatical patterns indicating stereotypical and non-stereotypical gender-role assignments (e.g., female nurses versus male dancers) in corpora from three domains, resulting in a first large-scale gender bias dataset of 108K diverse real-world English sentences. We manually verify the quality of our corpus and use it to evaluate gender bias in various coreference resolution and machine translation models. We find that all tested models tend to over-rely on gender stereotypes when presented with natural inputs, which may be especially harmful when deployed in commercial systems. Finally, we show that our dataset lends itself to finetuning a coreference resolution model, finding it mitigates bias on a held out set. Our dataset and models are publicly available at github.com/SLAB-NLP/BUG. We hope they will spur future research into gender bias evaluation mitigation techniques in realistic settings.


the doctor told the nurse that he did a good job

The Rustavi 2 TV chief said that he talked with Khoperia after the program and said that Rustaviis ready to continue cooperation.

The doctor said she did not think the boy had SARS, but would treat it as such.

Ruth Stone , Second-year student at the University of East Anglia, reflects on her past 2 years as a nursing student and considers what she has ahead.

Figure 1: We propose a semi-automatic method to vastly extend synthetic, small diagnostic datasets. We start with the texts of Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018), specifically designed to to be challenging for coreference and machine translation (top), extract syntactic patterns focusing on the salient entities in the artificial sentences (middle), and query real-world datasets for matching texts, using SPIKE (Shlain et al., 2020). The result is a large collection of diverse real-world texts exhibiting similar challenging properties which lends itself to both finetuning and testing (bottom).

1 Introduction

Gender bias in machine learning occurs when supervised models predict based on spurious societal correlations in their training data. This may result in harmful behaviour when it occurs in models deployed in real-world applications (Caliskan et al., 2017; Buolamwini and Gebru, 2018; Bender et al., 2021).[1]

Recent work has quantified bias mostly using carefully designed templates, following the Winograd schema (Levesque et al., 2012). Zhao et al. (2018) and Rudinger et al. (2018) probed for gender bias in coreference resolution with templates portraying two human entities and a single pronoun. For example, given the sentence *"the doctor asked the nurse to help her because she was busy"*, models often erroneously cluster "her" with "nurse", rather than with "doctor". Stanovsky et al. (2019) used the same data to evaluate gender bias in machine translation. When translating this sentence to a language with grammatical gender, models tend to inflect nouns based on stereotypes, e.g., in Spanish, preferring the masculine inflection over the correct feminine inflection (*"doctor-a"*).

While these experiments are useful for quanti-

[1] We acknowledge that gender identity is non-binary. Throughout this work we refer to *grammatical* gender, which has categorical inflections in the discussed languages (e.g., masculine and feminine pronouns in English).

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2470–2480
November 7–11, 2021. ©2021 Association for Computational Linguistics

fying gender bias in a controlled environment, we identify two shortcomings with this approach. First, the artificially-constructed texts diverge from natural language training distribution, which may inadvertently cause models to use prior distributions on such unseen constructions. Second, the small-scale templated data does not lend itself to training or finetuning to mitigate gender bias, limiting these datasets to diagnostic purposes.

In this work, outlined in Figure 1, we address both of these limitations by creating BUG, a large-scale dataset of 108K sentences, sampled semi-automatically from large corpora using lexical-syntactic pattern matching (see Figure 2 for examples). To construct BUG, we devise 14 diverse syntactic patterns, matching a wide range of sentences, ensuring that each mentions a human entity and a pronoun referring to it. Following, we use the SPIKE engine (Shlain et al., 2020)[2] to retrieve matching sentences over three diverse domains, including Wikipedia, Covid19 research, and PubMed abstracts. Finally, we filter the resulting sentences and mark each as either stereotypical or anti-stereotypical with respect to gender role assignments. The result is large corpus which is diverse, challenging, and accurate.

We use BUG to conduct a first large-scale evaluation of gender bias on real-world texts. We find that popular machine translation and coreference models struggle with feminine entities and anti-stereotypical assignments. Furthermore, BUG enables us to identify novel insights. For example, that machine translation models tend to be more biased when there are many pronouns in the input sentence.

Finally, we show that BUG can also help in mitigating gender bias. We finetune a state-of-the-art coreference resolution model on the anti-stereotypical portion of BUG and achieve a 50% error reduction on a held out test set, at the cost of only a modest drop in overall accuracy.

To conclude, our main contributions are:

- We present BUG, a first publicly available large-scale corpus for gender bias evaluation which consists of diverse, real-world sentences.
- We evaluate gender bias at large scale on natural sentences, leading to novel insights in machine translation and coreference resolution.
- We use BUG to finetune a coreference resolution model, showing that the resulting model is less prone to make gender biased predictions.

2 Data Collection

In this section, present BUG, a semi-automatic collection of natural, “in the wild” English sentences which are challenging with respect to societal gender-role assignments. Similarly to some of the synthetic gender bias datasets (Zhao et al., 2018; Rudinger et al., 2018), we are looking for sentences with a human entity, identified by their profession (e.g., “cop”, “dancer”) and a gendered pronoun (e.g., “he”, “she”). For example, see the first sentence in Figure 2, where the cop co-refers with a feminine pronoun (“she”), while the judge in the last sentence in Figure 2 co-refers with a masculine pronoun (“his”).

As opposed to previous work, we are interested in naturally occurring sentences, rather than generating artificial sentences from fixed lexical templates. The process for achieving this is outlined in Figure 1 and elaborated below. First, we perform syntactic search for sentences with challenging syntactic properties over corpora from three domains (Section 2.1). We then filter the sentences to verify they contain at least one entity, and a corresponding pronoun (Section 2.2). Finally, we manually assess BUG, finding it to be 85% accurate (Section 2.3).

2.1 Syntactic Querying with SPIKE

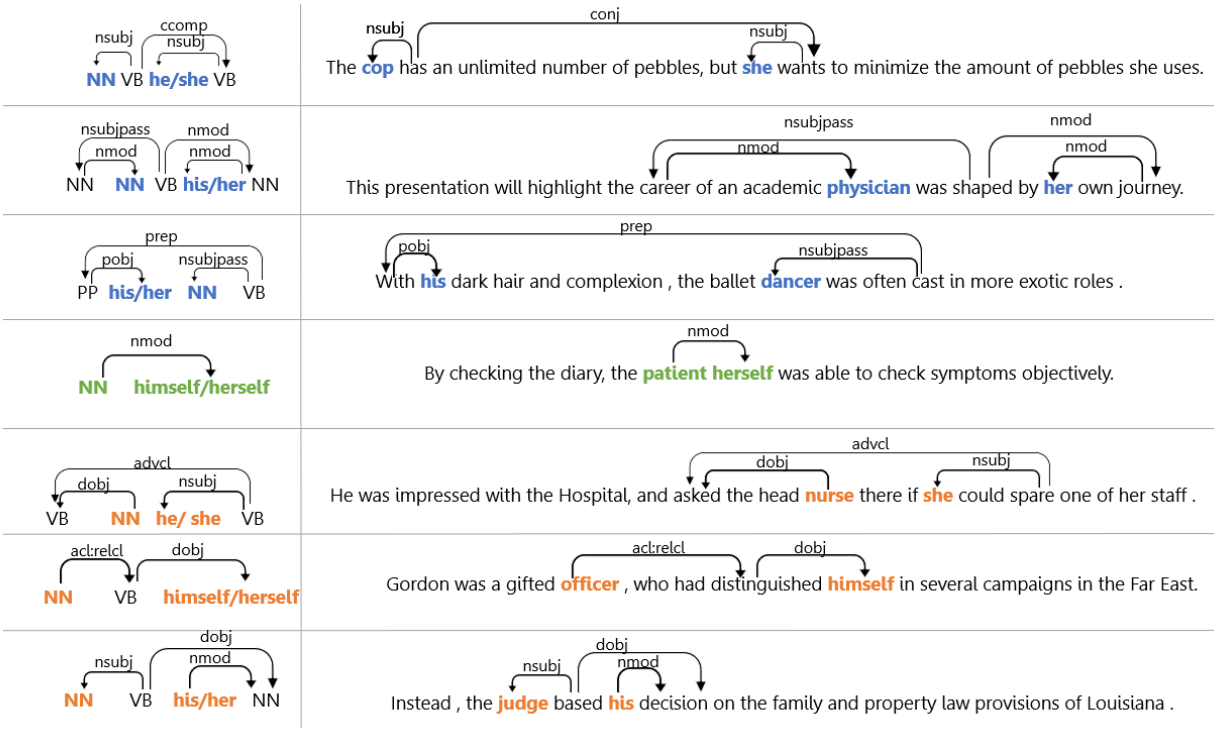
We devised 14 lexical-syntactic patterns, exemplified in Figure 2 to construct BUG. All our patterns have two anchors — a pronoun and a profession — which the pattern indicates are coreferring.[3]

For example, the last pattern in the figure links a noun (e.g., “officer”) with a relative clause relation (“acl:relcl”) to a verb (e.g., “distinguished”) modified by a direct object (“dobj”) gendered reflexive pronoun (“himself” or “herself”). These patterns were constructed by examining and expanding the sentences in the synthetic coreference corpora (Rudinger et al., 2018; Zhao et al., 2018).

To match these 14 patterns against real-world texts, we used SPIKE (Shlain et al., 2020), which

[2] spike.apps.allenai.org

[3] See Appendix for a full list of patterns: github.com/SLAB-NLP/BUG/blob/main/docs/appendix.pdf



The **cop** has an unlimited number of pebbles, but **she** wants to minimize the amount of pebbles she uses.

This presentation will highlight the career of an academic **physician** was shaped by **her** own journey.

With **his** dark hair and complexion , the ballet **dancer** was often cast in more exotic roles .

By checking the diary, the **patient herself** was able to check symptoms objectively.

He was impressed with the Hospital, and asked the head **nurse** there if **she** could spare one of her staff .

Gordon was a gifted **officer** , who had distinguished **himself** in several campaigns in the Far East.

Instead , the **judge** based **his** decision on the family and property law provisions of Louisiana .

Figure 2: Grammatical patterns (left) and corresponding examples sentences from BUG (right). Each instance in our dataset ensures at least a single human entity (marked by their profession) and a gendered pronoun, marked in bold. The sentences marked in blue are classified as anti-stereotypical while the sentences marked in orange are classified as stereotypical, and the sentence marked in green classified is neutral. The figure depicts 7 templates out of the 14 we designed. See the Appendix for a complete list.

indexes large-scale corpora and retrieves matching instances given a lexical-syntactic pattern. We queried corpora from three domains: Wikipedia, PubMed abstracts, and Covid19 research papers (Wang et al., 2020). The examples in Figure 2 highlight the diversity of the approach, while they all adhere to one of the predefined patterns, they vary widely in vocabulary and in syntactic construction, often introducing complex phenomena, such as coordination or adverbial phrases.

2.2 Marking Entities and Gender Roles

Following the lexical-syntactic querying, we filter BUG to make sure it contains human entities, and mark each instance as either stereotypical (bottom three examples in Figure 2), neutral (middle example) or anti-stereotypical (top three examples). This enables us to use BUG to measure gender bias in machine translation and coreference resolution models (Section 4).

We filter out two types of nouns: (1) nouns which do not refer to a person (e.g., “COVID-19”); (2) gendered English nouns (e.g., “princess”, “father”, or “sister”). To address both of these issues, we filtered the results with a predefined list of 183 professions, taken from the U.S. census.

Following, to mark each instance as either stereotypical or anti-stereotypical, we we follow Zhao et al. (2018) and Rudinger et al. (2018) and use the United States 2015 census’ gender distribution per occupation.[4] For instance, the first example Figure 2 is marked anti-stereotypical since “cop” is a predominantly male profession (76% in the census) and the referring pronoun is feminine.

2.3 Human Validation and Gold Standard

We estimate the accuracy of BUG by randomly sampling 1700 sentences from BUG, sampling uniformly across the data as well as from every pattern and domain. 17 human annotators proficient in English were asked to decide whether the gender BUG assigned to the entity matches their understanding of the sentence. The complete annotation guideline is presented in the Appendix. Overall we found that 85% of the instances were marked correct. We publish these annotation as a separate resource of diverse sentences with gold annotations (dubbed *Gold BUG*).

[4] https://www.kaggle.com/jonavery/incomes-by-career-and-gender

Category	Example	Comments
Disambiguated by noun (67%)	A **physician** who respects **her** autonomy should respect Ann’s right to make this decision.	Noun selection affects coreference decision. E.g., replacing “autonomy” with “job” would lead to a correct annotation.
Ambiguous (23%)	Hiei’s **captain** ordered **her** crew to abandon ship after further damage.	The antecedent is ambigous (either captain or Hiei).
Non-gendered pronoun (7%)	The IPP is a portfolio in which the **student** reflects on **his/her** learning and development during the production.	Reference to masculine and feminine pronouns.
Reported speech (3%)	We remove the comments , but this person keeps putting them back up - things like “**he** says he never met that woman”.	Quoted pronoun which does not refer to an entity in the sentence.

Table 1: Error analysis of 30 errors found in a sample of 200 randomly sampled sentences from BUG.

Corpus	Stereotypical	Anti-stereotypical	Neutral	Male	Female	Total
WinoGender + WinoBias	1,584	1,584	720	1,826	2,062	3,888
GAP*	-	-	-	2,227	2,227	4,454
Wikipedia	48,909	25,529	5,607	63,677	16,368	80,045
Pubmed abstracts	4,099	3,665	16,543	16,021	8,286	24,307
Covid19 research	1,001	683	2,383	2,572	1,495	4,067
Balanced BUG	12,922	12,922	-	12,922	12,922	25,844
Gold BUG	865	420	435	1,337	383	1,720
BUG Total	**54,009**	**29,877**	**24,533**	**82,270**	**26,149**	**108,419**

Table 2: Statistics for existing gender bias datasets (top) versus different BUG subsets (bottom). Stereotypical, anti-stereotypical and neutral refer to societal gender role assignments. E.g., a sentence with male doctor is stereotypical, while a sentence with a female doctor is anti-stereotypical; male, female refer to the number of sentences with masculine and feminine pronouns. BUG contains sentences from the three corpora listed above it. WinoMT contains sentences from WinoGender and Winobias. *Sentences in GAP do not have stereotypical classification.

3 BUG Analysis

The collection described in the previous section resulted in 108k sentences and 1700 human annotations. Following, we analyze key characteristics of BUG, finding it to be lexically diverse, and an order of magnitude larger than previous gender bias corpora.

3.1 Error Analysis and Inter-Annotator Agreement

The error analysis in Table 1 reveals that the most common errors are due to constructions where syntactic patterns are ambiguous with respect to coreference.

For instance, in the first example in Table 1, replacing “autonomy" with “job” changes the antecedent from the physician to the patient. Future work may address this by trying to refine our lexical-syntactic patterns to also include verb selection information.

Other types of errors were less frequent and included cases where two pronouns were used as a single gender-neutral word (“he/she”), and where the pronoun was part of a named entity or reported speech.

In addition, we test agreement between two annotators on a subset of 200 randomly selected sentences. We found a high level of agreement (95.5%; 0.73κ). Disagreements mostly occur on ambiguous sentences, such as “On the night of 17 August , Charlotte reported that the *child* had been taken from *her* tent by a dingo .”, where one annotator read “her” as referring to the child, while the other thought that the pronoun refers to Charlotte.

3.2 Data Characteristics

BUG statistics are presented in Table 2 in comparison with other datasets for gender bias. BUG is more than 24 times larger than the GAP coreference challenge set (Webster et al., 2018) and more than 30 times larger than WinoMT (Wino-

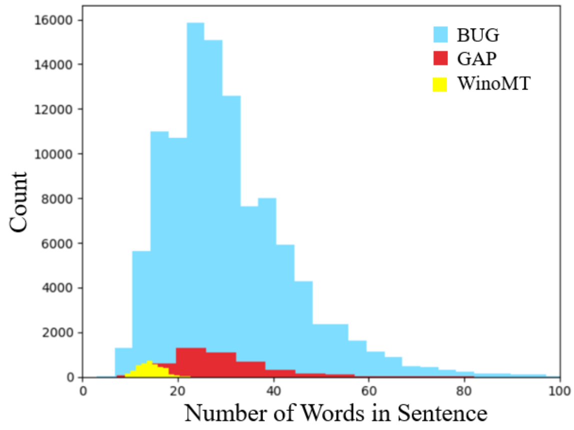


Figure 3: The distribution of the number of words in a sentence in BUG (in blue, average - 30.6 words per sentence) versus WinoMT used in Stanovsky et al. (2019) (in yellow, average - 14.3 words per sentence) and GAP used in Webster et al. (2018) (in red, average - 29.8 words per sentence). Word splitting was done with spaCy (Honnibal et al., 2020).

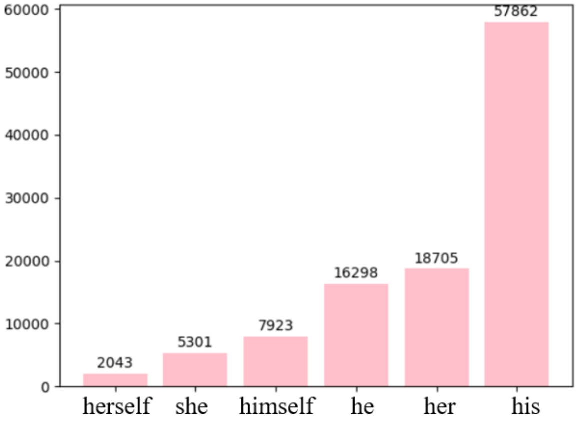


Figure 4: BUG pronoun histogram. In total, there are 82K (76%) masculine pronouns and 26K (24%) feminine pronouns.

Gender and Winobias combined) (Stanovsky et al., 2019). BUG consists of 110,544 unique words, while in the WinoMT corpus the vocabulary size is 1,868 and GAP's vocabulary size is 31,834. BUG is more diverse and naturally distributed, as can be seen in the histogram of sentence lengths depicted in Figure 3. Furthermore, the mean distance (in words) between entity and pronoun does not significantly differs between stereotypical (6.4[±4.5]) and anti-stereotypical (6.3[±4.6]) partitions, thus alleviating recent concerns about such artifacts in diagnostic datasets (Kocijan et al., 2021).

Our sentences were sampled from three corpora indexed in SPIKE. The majority were drawn from Wikipedia. Relative to the size of the original corpora, the yield from Wikipedia is 6 times more productive than PubMed and 4 times more than the Covid19 research domain. This is possibly since Wikipedia lends itself more to discussion of different entities in different settings.

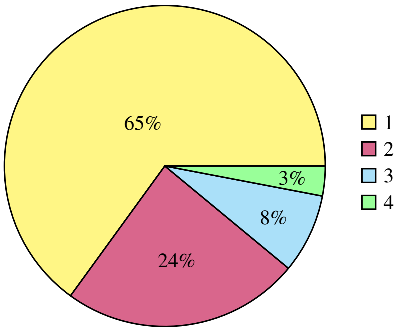


Figure 5: The distribution of the number of pronouns in our corpus. 35% (41K) of the sentences have more than one pronoun, further complicating the coreference resolution task.

As expected, since BUG is sampled from real texts, most of the data is stereotypical and most entities are male. There are three times more sentences with masculine pronouns compared to feminine pronouns, as shown in Figure 4; there are twice as many sentences with typically-male professions compared to typically-female professions; and twice as many sentences classified as stereotypical than anti-stereotypical. The natural texts also present a more challenging coreference setting. As evident in Figure 5 by large number of instances (35% of the corpus) with more than one pronoun.

To allow for more controlled evaluations, we publish two subsets of BUG. *Gold BUG* consists of the gold-quality human-validated samples, while *Balanced BUG* is randomly sampled from BUG to ensure balance between male and female entities and between stereotypical and non-stereotypical gender role assignments. We report statistics for both of these subsets in Table 2.

4 Evaluating Gender Bias in The Wild

We evaluate the performance of machine translation and coreference resolution models on BUG, using the metrics and tools established in previous work (Rudinger et al., 2018; Zhao et al., 2018; Stanovsky et al., 2019). To the best of our knowledge, this is the first quantitative evaluation of gender bias in such systems on a large scale using naturally occurring sentences. Such inputs better resemble real-world use where biases can affect

Target Language	Opus-MT			mBART50_m2m			m2m_100_418M		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
Arabic	75.4	19.1	12.4	**79.5**	26.0	15.5	73.8	52.3	16.6
Czech	83.2	26.3	23.6	**85.0**	20.7	21.7	76.1	48.6	20.4
German	75.7	24.9	15.2	**77.2**	25.0	17.2	70.0	44.1	16.8
Spanish	**63.4**	20.5	15.5	62.8	20.1	15.8	57.1	43.9	15.5
Hebrew	**75.8**	28.4	24.3	57.7	14.8	21.2	73.3	45.9	20.3
Italian	58.8	32.8	19.8	**61.1**	27.2	20.9	55.8	48.6	20.8
Russian	68.7	47.1	17.3	**73.5**	33.4	12.6	68.6	55.2	13.9
Ukrainian	67.1	35.4	17.3	**71.5**	26.1	16.2	67.8	48.4	15.8

Table 3: Results for machine translation gender bias evaluation evaluation across 8 diverse target languages on the BUG dataset. *Acc* represents the overall accuracy (F1) of gender translation. Δ_G is the difference in accuracy between masculine and feminine entities. Δ_S is the difference in performance between stereotypical and anti-stereotypical gender role assignments. Positive Δ_G and Δ_S values indicate that the translations are gender biased.

many users.

4.1 Experimental Setup

Machine translation. We used EasyNMT[5] to evaluate three machine translation models: mBART50_m2m (Tang et al., 2020; Liu et al., 2020), m2m_100_418M (Fan et al., 2020), and Opus-MT (Tiedemann and Thottingal, 2020), representing the state-of-the-art for publicly available neural machine translations models. We translated BUG from English to a set of eight diverse target languages with grammatical gender: Arabic, Czech, German, Spanish, Hebrew, Italian, Russian and Ukrainian, using tools developed in previous work to infer the translated gender based on morphological inflections (Stanovsky et al., 2019; Kocmi et al., 2020).[6]

Coreference resolution. We use the AllenNLP (Gardner et al., 2018) implementation of SpanBERT (Joshi et al., 2020). SpanBERT introduces contextual span representation to the the e2e-coreference model (Lee et al., 2018; Joshi et al., 2019) to achieve state-of-the-art results on the English portion of the popular CoNLL-2012 shared task coreference benchmark (Pradhan et al., 2012).

4.2 Metrics

For each tested model we compute three metrics, following Zhao et al. (2018) and Stanovsky et al. (2019), while adapting the terminology suggested recently by Mehrabi et al. (2021).

Accuracy: Denotes the F1 score of the gender prediction. For machine translation, this indicates the percentage of instances in which a correct grammatical gender inflection was produced in the target language. For example translating a female doctor as *doctor-a* in Spanish. For coreference resolution accuracy refers to the portion of instances where the entity's antecedent is correctly clustered with its pronoun, e.g., a female doctor clustered with the feminine pronoun "her".

Population bias (Δ_G):[7] denotes the difference in accuracy (F1 score) between sentences with entities which co-refer with a masculine pronoun versus those with entities which co-refer with feminine pronouns. By definition, $-100 \geq \Delta_G \geq 100$. When $\Delta_G > 0$, the model tends to perform better when the input entities co-refer with masculine pronouns, and conversely when $\Delta_G < 0$ it performs better when they co-refer with feminine ones.

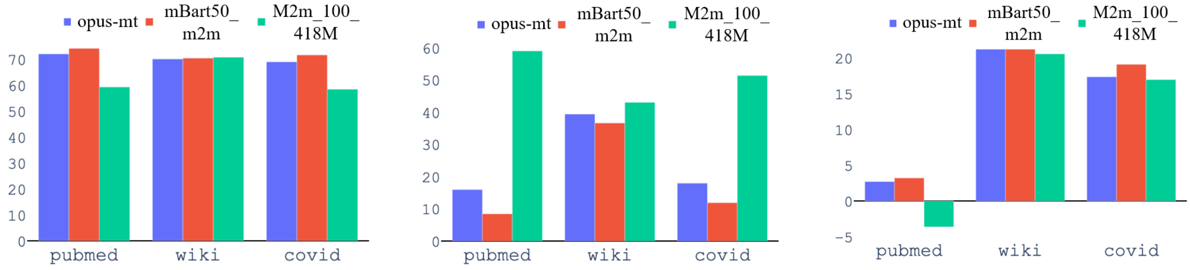
Historical Bias (Δ_S):[8] denotes the difference in accuracy (F1 score) between stereotypical sentences and anti-stereotypical sentences. Similarly to population bias, $\Delta_S \in [-100, 100]$, and positive

[5] `https://github.com/UKPLab/EasyNMT`

[6] We used the implementation provided by `github.com/gabrielStanovsky/mt_gender`

[7] According to Mehrabi et al. (2021), *Population bias* occurs when "statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population" (Olteanu et al., 2019).

[8] Mehrabi et al. (2021) defines *Historical bias* as "the already existing bias and socio-technical issues in the world" that "can seep ... from the data generation process even given a perfect sampling and feature selection." (Suresh and Guttag, 2021).



(a) Accuracy

(b) Δ_G

(c) Δ_S

Figure 6: Gender bias in machine translation across the domains in BUG.

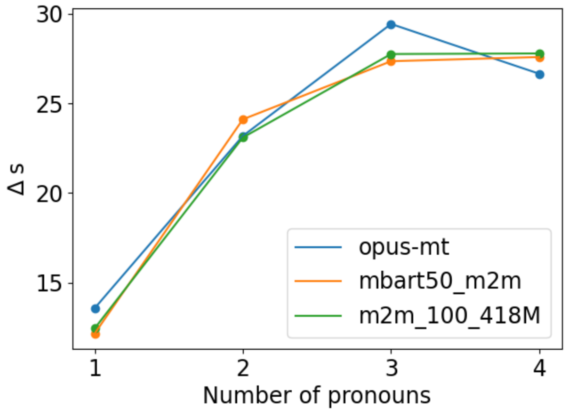


Figure 7: Historical gender bias (Δ_S) in machine translation models by the number of pronouns in the sentence. Indicating that while the bias is witnessed with a single pronoun, it is exacerbated in sentences with more pronouns.

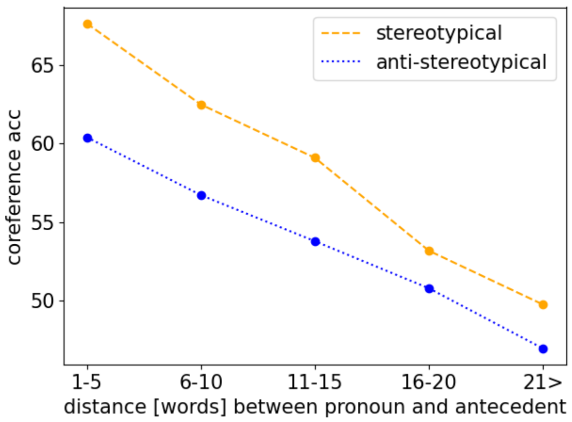


Figure 8: Coreference resolution performance as a function of the distance between pronoun and antecedent for stereotypical (orange) and anti-stereotypical (blue). The performance on both partitions deteriorates towards random choice the farther apart the two elements are.

values indicate that the model performs better on stereotypical gender role assignments.

4.3 Results

The results for gender bias in machine translation are presented in Table 3, and the results for coreference resolution are presented in the first row in Table 4. We draw various findings and observations based on these results and additional analyses.

All tested models for machine translation and coreference resolution are prone to gender bias on real-world texts. Both Δ_G and Δ_S are larger than zero across all settings, indicating that all models perform better on entities co-referring with a masculine pronoun and over-rely on gender stereotypes, even when it is in conflict with the pronouns providing contextual gender indications. To the best of our knowledge, this is the first time this phenomenon was observed and quantified at large scale on real-world instances, especially important for popular NLP services, such as machine translation and coreference resolution, which are in common use in many downstream applications.

Machine translation models do worse on sentences with many pronouns. Figure 7 breaks down Δ_S for machine translation as function of the number of pronouns in the sentence, showing that machine translation models are prone to fallback to their stereotypes the more pronouns appear in the sentence. This may be due to the increased syntactic complexity presented in such sentences.

Coreference resolution performance deteriorates towards random choice the longer the distance between pronoun and antecedent. Figure 8 shows that the larger the distance (in words) between entity and coreferring pronoun, SpanBERT's performance deteriorates towards random choice, for both stereotypical and anti-stereotypical partitions, diminishing the difference in performance between them.

Performance varies across domains. We compare gender bias across each of BUG's three domains in Figure 6. It seems that m2m_100_418M is the noisiest model in terms of gender bias, its accuracy is the lowest among all languages except Hebrew, and its Δ_G is the highest. In contrast, mBART50_m2m is the most accurate model among the three on all languages except Spanish and Hebrew, and its Δ_G is the lowest on all languages except Arabic and German. A possible explanation may be the vast difference in number of training parameters (15B for mBART50_m2m versus 418M in m2m_100_418M). Notably, m2m_100_418M achieves a negative Δ_S score on PubMed (Figure 6), indicating that it over translates entities using *anti-stereotypical* inflections (e.g., preferring to translate engineers as female). However, the model's low accuracy and high Δ_G score on the same corpus may indicate that this is mostly due to a noisy translation output, perhaps due to the scientific domain of the input texts in PubMed.

Our findings support previous work. The accuracy of the translations in this evaluation are much higher than that found by Stanovsky et al. (2019) and Zhao et al. (2018) work (69.9% in average vs. 47.6%), because of BUG's 3:1 ratio in favor of masculine entities versus feminine entities and 2:1 ratio in favor of stereotypical sentences versus anti-stereotypical sentences, representing a distribution which is closer to real-world use-cases. However, Δ_G and Δ_S are relative and their values are similar to those found in previous work, indicating that in fact all tested models were prone to gender bias. In addition, we find that all machine translation models achieve best performance on Czech as a target language, corroborating the findings of Kocmi et al. (2020), and that Russian and Hebrew have the highest Δ_G and Δ_S respectively, again confirming previous findings (Stanovsky et al., 2019). For coreference resolution, SpanBERT's gender bias Δ_S metric in Table 4 is better (i.e., smaller) than the models reported by (Zhao et al., 2018) (6.0 versus 13.5), which again may be due to the increase in number of parameters.

Coreference Model	Acc	Δ_G	Δ_S
SpanBERT	**65.1**	10.2	6.0
SpanBERT + anti-stereotypical BUG	64.1	**5.8**	**2.9**

Table 4: Results for gender bias in coreference resolution. The first row indicates the performance of off-the-shelf SpanBERT on our human validated annotations (Gold BUG), showing that it tends to overperform when clustering masculine and stereotypical gender role assignments. The second row depicts results after finetuning on the anti-stereotypical portion of BUG, showing a 50% error reduction at the cost of a 1% absolute reduction in accuracy.

5 Debiasing with BUG

Finally, we show that BUG's size and diverse instances make it amenable for finetuning, which results in more robust models, less prone to rely on gender stereotypes.

In the second row in Table 4 we report results of finetuning SpanBERT on the anti-stereotypical portion of BUG (consisting of 29.9K instances), and reevaluate its gender bias metrics on the held out human validated instances (Gold BUG, 1,720 instances). The motivation is to overexpose the coreference model to anti-stereotypical gender role assignment, where relying on stereotypes would directly hurt performance. Indeed, this yields a relative error reduction of more than 50% (3% absolute improvement).

We note however, that this comes at the cost of an absolute 1% drop in overall performance accuracy, which may be an expected side-effect due to the shift in training set distribution. Future work can explore ways to find better trade-offs between accuracy and reliance of gender bias with the help of BUG.

6 Related work

Several works created synthetic datasets to evaluate gender bias (Kiritchenko and Mohammad, 2018; González et al., 2020; Renduchintala and Williams, 2021), e.g., in the context of coreference (Rudinger et al., 2017; Zhao et al., 2018) and machine translation (Stanovsky et al., 2019; Prates et al., 2019; Kocmi et al., 2020), and some works used synthetic datasets to debias models (Saunders et al., 2020; Zhao et al., 2018).

Webster et al. (2018) and Gonen and Webster (2020), collected natural medium-scale (4.4K sentences) datasets from Wikipedia and reddit, respectively, and use them to evaluate gender bias in models of coreference resolution and machine translation. However, their datasets focused on the difference in performance between masculine and feminine entities (population bias), while in this work we also measure historical bias as the difference in performance between stereotypical and anti-stereotypical gender role assignment. In Section 3, we compare BUG to these datasets, finding it is more diverse and challenging in various respects.

7 Conclusion and Future Work

We presented BUG, a large-scale corpus of 108K diverse real-world English sentences, collected via semi-automatic grammatical pattern matching. We use BUG to evaluate gender bias in various coreference resolution and machine translation models, finding that models tend to make predictions in accordance with gender stereotypes, even when in conflict with opposite gendered pronouns in the sentence. Finally, we finetuned a coreference resolution model on BUG, finding it reduces its gender bias on a held out set. Our data and code are publicly available at `github.com/SLAB-NLP/BUG`.

Future work can extend BUG by including more patterns and by extracting sentences from corpora with gold annotations for machine translation and coreference resolution. This will allow exploration of the effect that exposure to anti-stereotypical examples during finetuning has on gender bias reduction.

Acknowledgements

We thank Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg for their help with SPIKE during our experiments, for fruitful discussions and their comments on earlier drafts of the paper, and the anonymous reviewers for their helpful comments and feedback. This work was supported in part by a research gift from the Allen Institute for AI.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

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

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Hila Gonen 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. Association for Computational Linguistics.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. The gap on gap: Tackling the problem of differing data distributions in bias-measuring datasets. In *AAAI*.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. 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 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and L. Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Adithya Renduchintala and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender. *ArXiv*, abs/2104.07838.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 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.

Danielle Saunders, 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). Association for Computational Linguistics.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23, Online. 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.

Harini Suresh and John V. Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy

Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 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.