Extending Challenge Sets to Uncover Gender Bias in Machine Translation
Impact of Stereotypical Verbs and Adjectives

Jonas Troles
Cognitive Systems
University of Bamberg
jonas.troles@uni-bamberg.de

Ute Schmid
Cognitive Systems
University of Bamberg
ute.schmid@uni-bamberg.de

Abstract

Human gender bias is reflected in language and text production. Because state-of-the-art machine translation (MT) systems are trained on large corpora of text, mostly generated by humans, gender bias can also be found in MT. For instance when occupations are translated from a language like English, which mostly uses gender neutral words, to a language like German, which mostly uses a feminine and a masculine version for an occupation, a decision must be made by the MT System. Recent research showed that MT systems are biased towards stereotypical translation of occupations. In 2019 the first, and so far only, challenge set, explicitly designed to measure the extent of gender bias in MT systems has been published. In this set measurement of gender bias is solely based on the translation of occupations. With our paper we present an extension of this challenge set, called *WiBeMT*[1], which adds gender-biased adjectives and sentences with gender-biased verbs. The resulting challenge set consists of over $70,000$ sentences and has been translated with three commercial MT systems: DeepL Translator, Microsoft Translator, and Google Translate. Results show a gender bias for all three MT systems. This gender bias is to a great extent significantly influenced by adjectives and to a lesser extent by verbs.

1 Introduction

The problem of unfair and biased models has been recognized as an important problem for many applications of machine learning (Mehrabi et al., 2019). The source of unfairness typically is based on biases in the training data. In many domains, unfairness is caused by sampling biases. In machine translation (MT), however, the main source of unfairness is due to historical or social biases when there is a misalignment between the world as it is and the values or objectives to be encoded and propagated in a model (Suresh and Guttag, 2019; Bolukbasi et al., 2016). One source of imbalance in natural language is the association of specific occupations with gender. Typically, occupations in more technical domains as well as occupations with high social status are associated with male gender (Cheryan et al., 2017).

In natural language processing (NLP), gender bias has been investigated for word embeddings (Bolukbasi et al., 2016). Analogy puzzles such as "man is to king as woman is to x" generated with *word2vec*[2] yield $x = queen$ while for "man is to computer programmer as woman is to x", the output is $x = homemaker$. Only few publications exist that directly address gender bias and MT. Although gender bias is particularly relevant for translations into gender-inflected languages, for instance from English to German, where biased models can result in translation errors (Saunders and Byrne, 2020). For the English sentence "The doctor told the nurse that she had been busy.", a human translator would resolve the co-reference of 'she' to the doctor, correctly translating it to 'die Ärztin'. However, a neural machine translation model (NMT) trained on a biased dataset, in which most doctors are male might incorrectly default to the masculine form, 'der Arzt'. Hovy et al. (2020) investigated the prediction of age and gender of a text's author before and after translation. They found that *Bing Translator*, *DeepL Translator*, and *Google Translate* all skew the predictions to be older and more masculine, which shows that NMT systems and the expression of gender interact.

According to Saunders and Byrne (2020), the first systematic analysis of gender bias in MT is from Stanovsky et al. (2019). They introduce *WinoMT* which is the first challenge set explicitly designed to quantify gender bias in MT sys-

[1] Our test set and related data is available at: www.github.com/JDtroles/WiBeMTdata.git

[2] www.code.google.com/archive/p/word2vec/

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 531–541
November 10–11, 2021. ©2021 Association for Computational Linguistics

tems. Furthermore, they introduce an automatic evaluation method for eight different languages. Evaluating six MT systems, they found a strong preference for masculine translations in all eight gender-inflected languages. The above example sentence, where 'doctor' has been errouneously translated into its male German form (*Arzt*) is one of the $3,888$ sentences used in *WinoMT*.

WinoMT focuses solely on the translation of occupations to evaluate gender bias. In this paper, we present the extended data set *WiBeMT* to uncover gender bias in neural machine translation systems. Gender stereotypical occupations are augmented by gender stereotypical adjectives and verbs and we investigate how congruent and incongruent combinations impact translation accuracy of the NMT systems DeepL, Microsoft Translator and Google Translate. In the next section, the original *WinoMT* data set is introduced together with our extensions. Afterwards, results for the three NMT systems based on 70,686 sentences are presented, followed by a discussion and an outlook.

2 Constructing the Extended Challenge Set WiBeMT

To construct a more diverse challenge set, we extend *WinoMT*, respectively its core data base *WinoBias*. Therefore, we identify verbs and adjectives of high stereotypicality with respect to gender. A gender score is determined by the cosine similarity of these words to a list of gender specific words. To calculate the similarity different pretrained word embeddings are used, where each of these words is represented by a vector. With the resulting most feminine and masculine adjectives the original sentences of *WinoBias* are extended. Furthermore, new sentences are created combining occupations with gender stereotypical verbs.

2.1 WinoBias and its Extension

WinoMT is based on two previous challenge sets, *Winogender* (Rudinger et al., 2018) and *WinoBias* (Zhao et al., 2018). Both were introduced to quantify gender bias in co-reference resolution systems. Since *WinoBias* constitutes 81.5% of *WinoMT*, we use it as basis for our extension. In total *WinoBias* consists of $3,168$ sentences, of which $1,582$ are feminine and $1,586$ are masculine sentences. The sentences are based on 40 occupations. An example sentence of *WinoBias* involving a cleaner and a developer is:

- WinoBias sentence: *The cleaner hates the* ***developer*** *because* ***she*** *always leaves the room dirty.*
- DeepL translation: *Die Reinigungskraft hasst* ***den Entwickler****, weil* ***sie*** *das Zimmer immer schmutzig hinterlässt.*

DeepL fails to correctly translate **developer** to the female inflection **die Entwicklerin**, but instead favors the stereotypical male inflection **der Entwickler**.

The *WinoBias* set is constructed such that each sentence is given in a stereotypical and an anti-stereotypical version. Stereotypical in this context means that the gender of the pronoun matches the predominant gender in the sentences occupation. The example sentence above is the anti-stereotypical version for the occupational noun 'developer', where the stereotypical version contains a 'he' instead of a 'she'.

We argue that a measurement of gender bias solely based on the translation of occupational nouns does not do justice to the complexity of language. Therefore, we want to diversify the given approach by taking gender-stereotypical adjectives and verbs into account. This is realized in two steps: First, *WinoBias* sentences are extended such that each occupational noun is preceded with an adjective which is congruent or incongruent with respect to the gender stereotypical interpretation of the occupation. For instance, a developer might be *eminent* (male) or *brunette* (female). By adding feminine and masculine adjectives to each WinoBias sentence, we create a new subset of extended WinoBias sentences.

Second, we create completely new sentences based on feminine and masculine verbs. One example with a feminine verb being: "The X **dances** in the club", and one with a masculine verb: "The X **boasts** about the new car.". Those base sentences are then extended with 99 occupations from Zhao et al. (2018), Rudinger et al. (2018), and Garg et al. (2018), resulting in a new subset of verb sentences. All 100 occupational nouns used in *WinoBias* and in the new verb sentences are listed in Table 1 which is based on numbers from the US Bureau of Labor Statistics[3]. After concatenating both subsets, the complete extended gender-bias challenge set consists of $70,686$ sentences. Overall, 100 occupa-

[3] Labor Force Statistics from the Current Population Survey: www.bls.gov/cps/cpsaat11.htm

Table 1: All 100 occupations, used for this work with the corresponding percentage of women in the occupation in the US. Numbers with an asterisk are taken from the WinoBias paper and are from 2017 (Zhao et al., 2018). All other numbers are from 2019 from the US Bureau of Labor Statistics. Occupations in bold font are used in the WinoBias challenge set. All occupations without corresponding percentage were not listed by the US Bureau of Labor Statistics.

Occupation	%	Occupation	%	Occupation	%	Occupation	%	Occupation	%
electrician	2	athlete	35	pharmacist	60	**receptionist**	**89**	examiner	–
carpenter	**3**	**lawyer**	**36**	**accountant**	**62**	**nurse**	**90***	gardener	–
firefighter	3	**janitor**	**37**	**auditor**	**62**	paralegal	90	geologist	–
plumber	3	musician	37	**editor**	**63**	dietitian	92	hygienist	–
construction-worker	**4**	**CEO**	**39***	**writer**	**63***	**hairdresser**	**92**	inspector	–
laborer	**4***	**analyst**	**41***	author	64	nutritionist	92	investigator	–
mechanic	**4***	**physician**	**41**	instructor	65	**secretary**	**93**	mathematician	–
driver	**6***	surgeon	41	veterinarian	68	administrator	–	officer	–
machinist	6	**cook**	**42**	**cashier**	**71**	advisor	–	pathologist	–
painter	9	chemist	43	**clerk**	**72***	appraiser	–	physicist	–
mover	**18***	**manager**	**43***	**tailor**	**75**	broker	–	practitioner	–
sheriff	**18**	**supervisor**	**44***	**attendant**	**76***	CFO	–	professor	–
developer	**20***	**salesperson**	**48***	**counselor**	**76**	collector	–	sailor	–
programmer	20	photographer	49	**teacher**	**78***	conductor	–	scientist	–
guard	**22***	bartender	53	planner	79	CTO	–	soldier	–
architect	25	dispatcher	53	**librarian**	**80**	dancer	–	specialist	–
farmer	**25**	judge	53	psychologist	80	doctor	–	student	–
chief	**28**	artist	54	**assistant**	**85***	economist	–	surveyor	–
dentist	34	**designer**	**54**	**cleaner**	**89***	educator	–	technician	–
paramedic	34	**baker**	**60**	**housekeeper**	**89***	engineer	–	therapist	–

tions are used in the set, 42 gender-verbs, and 20 gender-adjectives.

We hypothesize that gender-stereotypical verbs and adjectives influence the gender of the translation of the occupational nouns, when translating from English to the gender-inflected language German:

Hypothesis 1

1 Sentences with a feminine verb result in significantly more translations into the female inflection of an occupation than sentences with a masculine verb.

Hypotheses 2

2a WinoBias sentences extended with a feminine adjective result in significantly more translations into the female inflection of an occupation than original WinoBias sentences (without a preceded adjective).

2b WinoBias sentences extended with a masculine adjective result in significantly more translations into the masculine inflection of an occupation than original WinoBias sentences (without a preceded adjective).

2.2 Finding Gender-Stereotypical Verbs and Adjectives

To determine gender-stereotypicality of adjectives and verbs, large collections of these word types have been scored with respect to their similarity to a list of gender specific words given by Bolukbasi et al. (2016). As input we used a list of 3.250 verbs from patternbasedwriting.com[4] and a combined list of 4.889 adjectives from patternbasedwriting.com and from Garg et al. (2018).

Calculation of gender score is based on word embeddings and the cosine-similarity, following the work of Bolukbasi et al. (2016) and Garg et al. (2018). In their work similarity of the words to the pronouns “she” and “he” has been used. We extended scoring using a longer list of feminine and masculine words such as “mother”, “uncle”, “menopause” or “semen” to enhance robustness of the gender-score. This list, containing 95 feminine and 108 masculine words, is taken from Bolukbasi et al. (2016), who used it for their debiasing methods of word embeddings.

Two families of *word embeddings* were used: two pre-trained versions of *fastText*[5] from Mikolov et al. (2017) and two pre-trained versions of *GloVe*[6] from Pennington et al. (2014). All four word embeddings have a vector size of 300 dimensions.

[4] www.patternbasedwriting.com offers teaching materials for primary school children.

[5] Downloaded from: www.fasttext.cc/docs/en/english-vectors.html

[6] Downloaded from: www.nlp.stanford.edu/projects/glove/

Table 2: Summary of the origin of training corpora and their corresponding size for each word embedding.

	Size [billion]			
corpora	fastText-small	fastText-large	GloVe-small	GloVe-large
Wikipedia 2014	—	—	1.6	—
Gigaword 5	—	—	4.3	—
Wikipedia meta-page 2017	9.2	—	—	—
Statmt.org News	4.2	—	—	—
UMBC News	3.2	—	—	—
Common Crawl	—	630	—	840

Table 2 gives an overview of all word embeddings and their training data.

Cosine Similarity is a measure of similarity between two normalized and non-zero vectors and can take values in the range between -1 and 1. Equation 1 shows the calculation of the cosine similarity between the vectors $\mathbf{a}$ and $\mathbf{b}$ of two words:

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum\limits_{i=1}^{n} a_i b_i}{\sqrt{\sum\limits_{i=1}^{n} a_i^2}\sqrt{\sum\limits_{i=1}^{n} b_i^2}} \tag{1}$$

where a_i and b_i are components of vector $\mathbf{a}$ and $\mathbf{b}$ respectively.

Since word embeddings inherit to some extend the meaning of words, it is possible to use the cosine similarity as a measure for the similarity in meaning or, furthermore, the relationships between words. This enables mathematical operations on the vectors representing words such that: $cos(\overrightarrow{brunette} \cdot \overrightarrow{her}) \geq cos(\overrightarrow{brunette} \cdot \overrightarrow{him})$ becomes true for “brunette” and other gender-biased adjectives. If the feminine-gender value ($cos(\overrightarrow{brunette} \cdot \overrightarrow{her})$) is then subtracted from the masculine-gender value ($cos(\overrightarrow{brunette} \cdot \overrightarrow{him})$) the resulting single float value indicates whether a word is gender-biased in the word embedding with which the cosine-similarity was computed.

The total gender-score is the sum of eight single scores resulting from the combination of the four different word embeddings with the cosine similarity with “she” and “he” and with the list of feminine and masculine words. Since different word embeddings vary in the strength of the inherited gender-bias and all word embeddings should have equal impact on the overall score, the interim results were normalized to fit a range between $a = -1$ and $b = 1$, as Equation 2 shows:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)} \tag{2}$$

To validate the procedure to determine a gender score for adjectives and verbs, the same method has been applied to the occupations given in Table 1. The gender-score for these occupations shows a strong correlation to the percentage of women working in each given occupation (see Figure 1).

Gender-score has been calculated for all verbs and adjectives for which word embeddings existed. They were sorted ascending from the most negative – and therefore most feminine – score value, to the most positive, i.e., most masculine, value.

Verbs: Of the $3,250$ *verbs*, $3,210$ could be ranked ($med = 0.151$, $std = 0.165$). After sorting the verbs by their gender-score, the most stereotypical verbs which could be used in a sentence in which person P actively does action A were picked. The gender score of the 21 selected feminine verbs ranges from -0.772 to -0.233 and of the masculine verbs from 0.445 to 0.733:

- Feminine verbs: crochet, sew, accessorize, bake, embroider, primp, gossip, shriek, dance, undress, milk, giggle, marry, knit, twirl, wed, flirt, allure, shower, seduce, kiss.
- Masculine verbs: draft, tackle, swagger, trade, brawl, reckon, preach, sanction, build, boast, gamble, succeed, regard, retire, chuck, overthrow, rev, resign, apprehend, appoint, fool.

Adjectives: Of the $5,441$ *adjectives*, $4,762$ could be ranked ($med = 0.189$, $std = 0.142$). After calculating the gender-score for the $4,762$ adjectives, beginning with the most feminine, respectively, most masculine, adjectives were tested for their suitability considering the extension of existing WinoBias sentences until 10 most feminine and suitable as well as 10 most masculine and

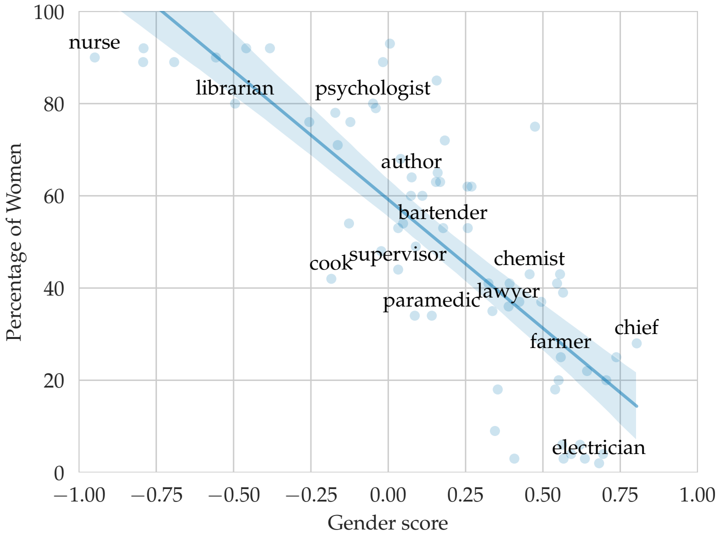

Figure 1: Scatter plot of the 99 single word occupations with the percentage of women in the profession y-axis and the gender score on the x-axis.

suitable adjectives had been selected. Words that semantically could not be combined with occupations were discarded (examples: hormonal, satin, luminous, philosophical, topographical). Furthermore, adjectives like "pregnant" – which only apply to persons with a uterus – were also discarded. The gender score of the 10 selected feminine adjectives ranges from -0.600 to -0.205 and of the masculine adjectives from 0.480 to 0.654:

- Feminine adjectives: sassy, perky, brunette, blonde, lovely, vivacious, saucy, bubbly, alluring, married.
- Masculine adjectives: grizzled, affable, jovial, suave, debonair, wiry, rascally, arrogant, shifty, eminent.

2.3 Translation of *WiBeMT* and Evaluation Design

To test the extent to which machine translation systems inherit a gender-bias, three different services were tested with *WiBeMT*: *Google Translate*, *Microsoft Translator*, and *DeepL Translator*. All three NMT systems were accessed via their API, and all translation processes took place in July 2020.

To test our hypotheses that adjectives and verbs significantly influence the gender in translations of NMT systems, translations from English to German have to be categorized as either (correctly) feminine, masculine, neutral, or wrong. Stanovsky et al. (2019) use an automated method in the form of different morphological analyzers such as *spaCy* to determine the gender of the occupation in the translated sentence. While this method is convenient for an automated approach that other researchers can use with different data, it also suffers from a certain degree of inaccuracy. To measure the accuracy of their automated evaluation method, Stanovsky et al. (2019) compared the automated evaluations with a random sample of samples evaluated by native speakers. They found an average agreement of 87% between automated and native speaker evaluations. To evaluate NMT systems with respect to *WiBeMT*, we preferred to augment automated evaluation by "manual" evaluation to gain higher accuracy.

Evaluating the gender of all translations is based on a nested list, we refer to as *classification-list*, and a set of rules for automated evaluation and manual evaluation for all remaining ambiguous translations. The classification-list contains four sublists for each occupation: a sublist of correct feminine translations, a sublist of correct masculine translations, a sublist of correct neutral translations, and a sublist of inconclusive or wrong translations. The gender of the translated occupation is then classified by checking in which sublist the occupation is listed and the gender is labeled as the sublist's category.

To control classifications for possible errors, $N = 665$ (1%) of the WinoBias sentences extended by adjectives were manually controlled for each translation system (in total $N = 1,995$ sentences). Not a single one was miscategorized. Of the verb-sentences translations, even 5% were manually controlled by the authors, and here also, not a single one of the 618 translations was miscategorized.

3 Results

After the creation of the *WiBeMT* challenge set and the translation of all $70,686$ sentences with each NMT system, the translations were categorized as *feminine*, *masculine*, *neutral*, or *inconclusive / wrong*. The latter will be referred to as "wrong". Due to these discrete categories the Chi^2 test of independence will be used for all statistical tests. As the extended *WinoBias* sentences include a pronoun that defines the gender, the results considering these data can be divided into *true* and *false*, and *feminine* and *masculine* translations. On the other hand, the verb sentences do not include any cue for a *correct* gender in the translation. Therefore, they are just analyzed for feminine and masculine translations. The calculated *feminine-ratio* ($\%TFG$) results from all feminine-translations divided by the sum of feminine- and masculine-translations. The calculated *correct-gender-ratio* ($\%TCG$) results from all translations with correct gender divided by the sum of all translations with correct and incorrect gender. All Chi^2-tests were, if necessary, Bonferroni corrected. First, the results for the verb-sentences; Second, the results for the extended WinoBias sentences; And third, further results are presented.

3.1 Hypothesis 1: Verb Sentences

The statistical tests regarding Hypothesis 1 yielded mixed results, depending on which NMT system is looked at. Of the 2.079 sentences with a feminine verb DeepL translated the occupations in 242 (11.9%) sentences into the female gender ($\%TFG$) compared to Microsoft Translator with 149 (7.5%), and Google Translate with 120 (6.0%). Of the 2.079 sentences with a masculine verb DeepL (DL) translated the occupations in 135 (6.6%) sentences into the female gender compared to Microsoft Translator (MS) with 104 (5.2%), and Google Translate (GT) with 109 (5.4%). For DeepL and Microsoft Translate the difference becomes significant with $\Delta\%TFG_{DL} = 5.3\%$, $Chi^2_{DL} = 33.7$ and $p_{DL} < 0.001$, and $\Delta\%TFG_{MS} = 2.3\%$, $Chi^2_{MS} = 8.4$ and $p_{MS} = 0.004$. For Google Translate the difference does not become significant. As two of three NMT systems inherit a gender bias that is influenced as expected by verbs, meaning that translations of sentences with feminine verbs result in a significantly higher $\%TFG$ than translations of sentences with masculine verbs, $H1$ is accepted. Figure 2 gives an overview of the translations to the female gender ($\%TFG$) for all three NMT systems and the two categories of verb sentences.

3.2 Hypothesis 2: Extended WinoBias Sentences

Both hypotheses assume that adjectives that inherit a gender-bias in word embeddings also influence the gender of translations from English to German. While the results confirm the assumptions of Hypothesis H2$_a$, they also yield unexpected results for Hypothesis H2$_b$. Of the 3,168 original WinoBias sentences DeepL translated $1,259$ (41.7%), Microsoft Translator translated $1,041$ (35.8%), and Google Translate translated 976 (33.2%) to the female inflection of the occupation. These percentages are used as the baseline to test, whether feminine and masculine adjectives skew the outcome of translations into the expected direction.

Of the sentences extended with a feminine adjective DeepL translated 49.5%, Microsoft Translator translated 42.5%, and Google Translate translated 40.1% to the female inflection. The difference to the percentage of translations to the female inflection of the original WinoBias sentences becomes significant for all three NMT systems: $\Delta\%TFG_{DL} = 7.8\%$, $Chi^2_{DL} = 66.3$ and $p_{DL} < 0.001$; $\Delta\%TFG_{MS} = 6.7\%$, $Chi^2_{MS} = 49.0$ and $p_{MS} < 0.001$; and $\Delta\%TFG_{GT} = 6.9\%$, $Chi^2_{GT} = 53.8$ and $p_{GT} < 0.001$. Therefore Hypothesis H2$_a$ is accepted.

Of the sentences extended with a masculine adjective DeepL translated 44.6%, Microsoft Translator translated 39.3%, and Google Translate translated 37.9% to the female inflection. The difference to the percentage of translations to the female inflection of the original WinoBias sentences becomes significant for all three NMT systems: $\Delta\%TFG_{DL} = 2.9\%$, $Chi^2_{DL} = 9.1$ and $p_{DL} = 0.008$; $\Delta\%TFG_{MS} = 3.5\%$, $Chi^2_{MS} = 13.3$ and $p_{MS} < 0.001$; and $\Delta\%TFG_{GT} = 4.7\%$, $Chi^2_{GT} = 25.1$ and $p_{GT} < 0.001$. While all differences are significant, they contradict our assumption that preceding masculine adjectives to Wino-

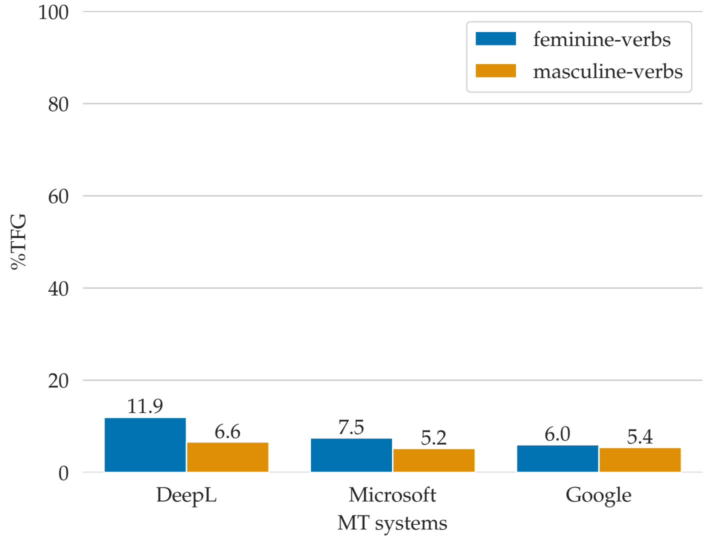

Figure 2: Percentage of translations to female gender ($\%TFG$) of occupations in verb sentences, organized by the category of verbs.

Bias sentences results in less translations to the female inflection. Instead the preceded masculine adjectives have the opposite effect. Therefore, Hypothesis H2_b is rejected. Figure 3 gives an overview of the translations to the female gender ($\%TFG$) for all three NMT systems and the three categories of sentences. Table 3 lists all numbers of different translation categories for a deeper insight.

3.3 Influence of Gender-Stereotypical Verbs and Adjectives on Translations

Our findings show that gender stereotypical verbs and adjectives influence gender bias in translations of NMT systems. In the following, we discuss our results in more detail.

Hypothesis 1: In general, the results from verb sentences differ drastically from the ones of the extended WinoBias sentences, with far fewer sentences being translated to their feminine gender. The percentage of sentences translated to their feminine gender ($\%TFG$) in the verb sentences ranges from 5.2% for masculine verb sentences translated by Microsoft Translator, to 11.9% for feminine verb sentences translated by DeepL. In comparison to that, $\%TFG$ ranges from 33.2% in original WinoBias sentences without adjective translated by Google Translate, to 49.5% in extended WinoBias sentences with feminine adjectives translated by DeepL Translator.

Two reasons could be responsible for the low $\%TFG$. Firstly, the generic masculine in German: As mostly the masculine gender is used to address all genders, this must be present in the training data of all three NMT systems. Therefore, they tend to use the male inflection whenever the gender bias for a specific occupation does not outweigh the bias by the generic masculine. Secondly, the verb-sentences lack a gender pronoun like “her” or “him”, which urges the NMT system to decide which gender would be correct in the translation. This probably leads to the strong bias towards male inflections in translations.

To check whether the bias induced by occupations and the generic masculine outweighed an existing bias of verbs, we analyzed the data of the verb sentences again, looking at the different groups of occupations[7]: feminine, neutral and masculine. For all three NMT systems the $\%TFG$ for verb sentences with neutral and masculine occupations was below 2% regardless of the stereotypical feminine verbs. The $\%TFG$ in sentences with female occupations was 25.1% in DeepL, 19.0% in Microsoft Translator and 17.6% in Google Translate. All differences to sentences with neutral and masculine occupations were significant with p-values below 0.001 and Chi^2 values reaching from 352 to 256.

Hypotheses 2a & 2b: As gender bias works both ways in word embeddings, meaning that words can be stereotypical feminine or stereotypical mascu-

[7] With the calculated gender score we split the list of occupations in three equally sized categories (each $n = 33$)

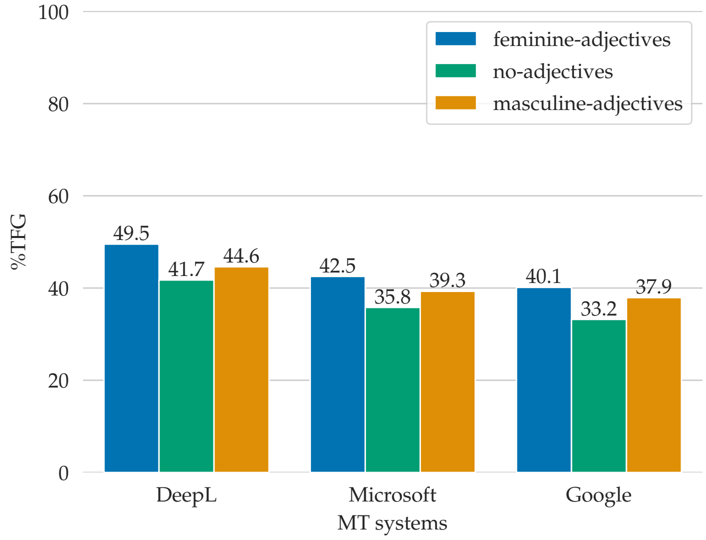

Figure 3: Percentage of translations to female gender ($\%TFG$) of occupations in all three types of EWB sentences: feminine adjective, masculine adjective, and no adjective.

line, we assumed that, depending on their gender-score derived from word embeddings, feminine adjectives would skew translations of NMT systems more often to their feminine gender and vice versa that masculine adjectives would skew translations more often to their masculine gender. This assumption was also supported by the findings of Stanovsky et al. (2019), who preceded "handsome" to occupations of sentences which would be correct, if translated to their masculine gender, and "pretty" to occupations of sentences which would be correct if translated to their feminine gender. With this measure, they could improve the accuracy of NMT systems and reduce gender bias.

Contrary to the prior assumptions, adjectives preceded to occupations led to significantly more translations with feminine gender, regardless of their gender-score derived from word embeddings. Finding that masculine adjectives also lead to more translations with feminine gender does not only result in the rejection of H2$_b$, but also weakens the acceptance of H2$_a$. Therefore, we introduce a new Hypothesis H2$_c$: "WinoBias sentences extended with a feminine adjective result in significantly more translations into the female inflection of an occupation than WinoBias sentences extended with a masculine adjective.".

The difference between the percentage of translations to the female inflection of WinoBias sentences extended with feminine and WinoBias sentences extended with masculine adjectives becomes significant for all three NMT systems: $\Delta\%TFG_{DL} = 4.9\%$, $Chi^2_{DL} = 147.6$ and $p_{DL} < 0.001$; $\Delta\%TFG_{MS} = 3.2\%$, $Chi^2_{MS} = 59.1$ and $p_{MS} < 0.001$; and $\Delta\%TFG_{GT} = 2.2\%$, $Chi^2_{GT} = 29.2$ and $p_{GT} < 0.001$. Therefore, Hypothesis H2$_c$ is accepted which strengthens H2$_a$, as stereotypical feminine adjectives preceded to occupational nouns lead to significantly more translations to the female gender than stereotypical masculine adjectives preceded to occupational nouns.

3.4 Influence of Adjectives on Correct Gender in Translations

The comparison of all three NMT systems was not one of the main research questions. Nevertheless, the extent of gender bias in the NMT systems can be of interest to users. Therefore, and because of the surprising findings considering Hypothesis 2_b a short comparison is presented in the following paragraphs.

To better understand our results regarding hypotheses H2$_a$, H2$_b$ and H2$_c$ we plotted percentage of translations with the correct gender ($\%TCG$) organized by NMT systems and the type of sentence: original WinoBias sentence (no adjective), with masculine adjective extended WinoBias sentence (M adjective), and with feminine adjective extended WinoBias sentence (F adjective). Additionally, we split each type of sentence into sentences with male pronouns (M pronouns) and sentences with female pronouns (F pronouns). Figure

Table 3: Numbers of the categorizations of the gender of translations of extended WinoBias sentences sorted by the gender-class of the adjective.

		NMT system		
adjective gender	translation category	DeepL	Microsoft	Google
feminine	true feminine	**13,341**	11,352	10,177
	false masculine	1,553	3,405	**4,593**
	true masculine	**13,833**	13,814	13,312
	false feminine	1,751	1,382	**1,826**
	neutral	**550**	440	472
	wrong	652	1,287	**1,300**
masculine	true feminine	**12,497**	8,966	8,724
	false masculine	2,378	3,405	**4,533**
	true masculine	**14,425**	11,943	12,128
	false feminine	1,028	963	**1,450**
	neutral	**483**	346	447
	wrong	869	**6,057**	4,398
no adjective	true feminine	**1,162**	916	812
	false masculine	324	503	**652**
	true masculine	**1,434**	1,365	1,313
	false feminine	97	125	**164**
	neutral	75	**77**	53
	wrong	76	**182**	174

4 shows the resulting plot.

The results are quite astonishing: while the $\%TCG$ slightly decreases for sentences with male pronouns and any adjective ($1 \leq \Delta\%TCG \leq 5$) it drastically improves for sentences with female pronouns and any adjective ($12 \leq \Delta\%TCG \leq 13$). Considering that sentences with female and male pronouns are equally prevalent in original WinoBias sentences and extended WinoBias sentences, it shows that preceding an occupational noun with any adjective likely improves the overall percentage of translations to the correct gender inflection. This reflects the findings of Stanovsky et al. (2019), who preceded "handsome" and "pretty" to occupational nouns of WinoMT sentences. With this measure, they could improve the accuracy hence $\%TCG$ of NMT systems and reduce gender bias. Our findings give more insight to this result: Stanovsky et al. (2019) very likely could have added either "handsome" or "pretty" to all sentences, regardless of their pronoun and would nonetheless have been able to record a higher accuracy.

Furthermore, Figure 4 shows that DeepL Translator performed best when it comes to $\%TCG$. Google Translate and Microsoft Translator again show more similar results, but Google Translate performed notably worse than Microsoft Translator. With the least discrepancy in $\%TCG$ between sentences with a feminine pronoun and sentences with a masculine pronoun in all three conditions (feminine adjective, masculine adjective, no adjective) DeepL Translator, therefore, inherits the lowest gender bias.

4 Conclusions and Further Work

The three neural machine translation systems evaluated with respect of their gender bias are black boxes, in so far as the architecture is not publicly available and – even more important – it is not transparent on what data these systems are trained. It is most likely that the gender bias in all three systems is inherited from the data used for training and their use of word embeddings.

To give a closer look on gender-bias of the NMT systems DeepL, Microsoft, and Google Translate, an extension of the *WinoMT* challenge set – the first challenge set designed to measure gender bias in NMT systems – has been presented. While *WinoMT* relies solely on the gender of the translation of occupations, our extended set *WiBeMT* includes gender-adjectives and gender-verbs. Thereby, a more detailed assessment of gender bias has been possible. The number of sentences in our challenge set is, with over $70,000$ sentences, nearly 20 times as large as the original WinoMT challenge set. This makes it less prone to overfitting when used to evaluate or reduce gender bias in NMT systems.

We could show that adjectives do significantly

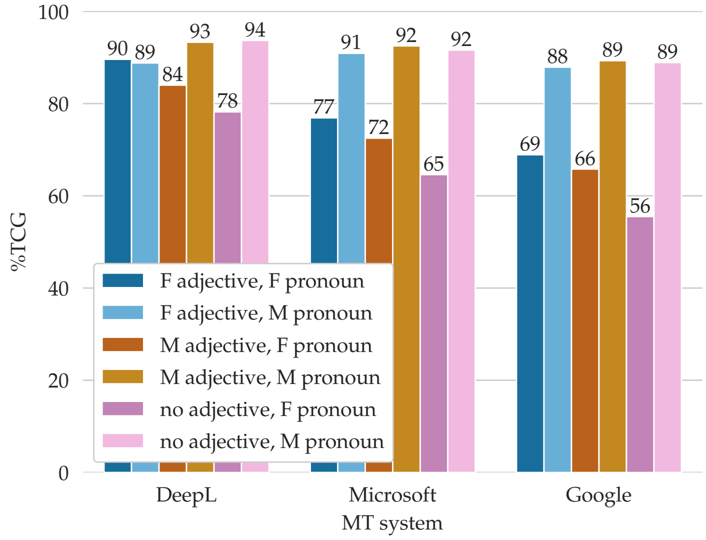

Figure 4: Bar plot of the percentage of translations into the correct gender ($\%TCG$) of original WinoBias sentences and extended WinoBias sentences.

influence the gender of translated occupations. Against the hypotheses, feminine as well as masculine adjectives skew the translations of NMT systems to more translations with the feminine gender. Nevertheless, feminine adjectives still produce significantly more translations with the feminine gender than masculine adjectives do.

All three NMT systems prefer translations to the generic masculine when no pronoun defines a correct gender for the translation. Despite this preference, 5.7% to 9.3% of all verb sentences are translated to their feminine gender by the MT systems. This can mostly be attributed to a gender bias in translating occupations, as our results regarding verb sentences and occupation categories show. The effect of gender-verbs on the gender of translations only became significant in DeepL and Microsoft Translator and was smaller than the effect of the occupations gender-categories. It can be assumed that Google Translate only performed better on the verb sentences task, as it generally has the strongest tendency to translate sentences to their masculine gender.

Surprisingly, gender-adjectives drastically improve the overall accuracy of all three NMT systems when it comes to the correct gender in translations. While the accuracy slightly drops for sentences with masculine pronouns, it drastically improves for sentences with feminine pronouns. Therefore, the discrepancy in accuracy between masculine and feminine pronoun sentences decreases, resulting in lower discrimination. One could even argue that gender-adjectives reduce gender bias in the output of NMT systems, but at the same time, it can be discriminating in itself that, as soon as you add an adjective to describe a person, the instance of this person is more likely to be translated to its feminine gender. Further research is certainly needed to find reasons for this effect and to assess the potentials of discrimination.

The sentences in our *WiBeMT* challenge set are – as the original *WinoMT* challenge set – constructed in a systematic way. While this allows for a controlled experiment environment, it might also introduce some artificial biases (Stanovsky et al., 2019). A solution could be to collect real-world examples of sentences, which are suitable for gender bias detection. Furthermore, the limitation on one source language (English) and one target language (German) does not allow for a generalization of the results.

Another limitation of our study – as most other studies – is that it does not take into account that gender should not be seen as a binary, but rather a continuous variable. Cao and Daumé (2019), for example, outline why “trans exclusionary” co-reference resolution systems can cause harm, which is probably also valid for MT systems. A further extension of the challenge set could help to shine a light on the shortcomings of the inclusion of transgender persons of NMT systems.

References

Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, pages 4356–4364.

Yang Trista Cao and Hal Daumé. 2019. Toward Gender-Inclusive Coreference Resolution. *arXiv preprint arXiv:1910.13913*.

Sapna Cheryan, Sianna A Ziegler, Amanda K Montoya, and Lily Jiang. 2017. Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, 143(1):1–35.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in Pre-Training Distributed Word Representations. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 52–55.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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, pages 8–14.

Danielle Saunders 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.

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.

Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *arXiv preprint arXiv:1804.06876*.