MT-GenEval: A Counterfactual and Contextual Dataset for Evaluating Gender Accuracy in Machine Translation

Anna Currey, Maria Nădejde, Raghavendra Pappagari, Mia Mayer,
Stanislas Lauly, Xing Niu, Benjamin Hsu, Georgiana Dinu
AWS AI Labs
ancurrey@amazon.com

Abstract

As generic machine translation (MT) quality has improved, the need for targeted benchmarks that explore fine-grained aspects of quality has increased (Freitag et al., 2021; Isabelle et al., 2017). In particular, gender accuracy in translation (Choubey et al., 2021; Saunders and Byrne, 2020) can have implications in terms of output fluency, translation accuracy, and ethics. In this paper, we introduce *MT-GenEval*, a benchmark for evaluating gender accuracy in translation from English into eight widely-spoken languages. MT-GenEval complements existing benchmarks by providing realistic, gender-balanced, counterfactual data in eight language pairs where the gender of individuals is unambiguous in the input segment, including multi-sentence segments requiring inter-sentential gender agreement. Our data and code is publicly available under a CC BY SA 3.0 license.[1]

1 Introduction

Although neural machine translation (NMT) has made great strides in quality (Hassan et al., 2018; Wu et al., 2016), evaluations on generic test sets may not tell the whole story. Indeed, NMT models are known to make systematic errors in areas like robustness to input perturbations (Niu et al., 2020), disambiguating pronouns in context (Müller et al., 2018), and translating unambiguous human gender (Stanovsky et al., 2019). In particular, gender-related issues in translation can lead to translations that are inaccurate, ungrammatical, or biased.

Accordingly, there has been increasing interest in improving machine translation for gendered entities (Bentivogli et al., 2020; Choubey et al., 2021; Saunders and Byrne, 2020; Savoldi et al., 2021; Stanovsky et al., 2019). Adequate evaluation benchmarks play an important role in supporting this line of research and improving understanding of how models perform on the task of gender translation accuracy. Existing benchmarks have limited diversity in terms of gender phenomena (e.g., focusing on professions), sentence structure (e.g., using templates to construct sentences), or language coverage (see section 4 for more information), making it difficult to gauge how systems perform in terms of both gender and quality simultaneously.

To this end, this paper releases a new **M**achine **T**ranslation **Gen**der **Eval**uation benchmark: **MT-GenEval**. We include development data (for model improvements) as well as test data and corresponding metrics (for comprehensive evaluation). MT-GenEval is realistic and diverse, includes a wide variety of contexts for gender disambiguation, and is fully balanced by including human-created gender counterfactuals. In its first release, the MT-GenEval benchmark covers translation from English into eight target languages, for two genders.

MT-GenEval focuses on the task of **gender accuracy in translation**. We define gender accuracy in translation as the extent to which a machine translation output accurately reflects the gender of the humans mentioned in the input, restricted to cases where the gender is explicitly and linguistically disambiguated in the context of the input. Thus, in our benchmark we do not consider the grammatical gender on inanimate objects, or cases where the input gender is ambiguous within the given level of context.[2]

Table 1 gives examples of the data in the MT-GenEval dataset. In all cases, the source segment contains a reference to a human and that human's gender is unambiguous based on the linguistic context, be that context intra-sentential (rows 1 and 2), or inter-sentential (row 3). Row 2 shows a

[1] https://github.com/amazon-research/machine-translation-gender-eval

[2] In this latter case, multiple translations are valid, so this falls under the purview of gender *customization* tasks (Habash et al., 2019; Nădejde et al., 2022; Saunders et al., 2020; Vanmassenhove et al., 2018).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4287–4299
December 7-11, 2022 ©2022 Association for Computational Linguistics

	Segment Pair
1	After some wrangling Blacket accepted £50 in full settlement of the fees due to him. Nach einigem Hin und Her akzeptierte Blacket 50 Pfund als vollen Ausgleich für die ihm zustehenden Gebühren.
2	Having served his apprenticeship Crookall became a *master painter* trading at Duke Street, Douglas. Tras servir como aprendiz, Crookall se convirtió en maestro pintor en la calle Duke de Douglas.
3	Serrano agreed to the restaurant contract as long as he could have a tapas restaurant. Serrano *traveled* to Spain [. . .] Serrano s'est rendu en Espagne [. . .]

Table 1: Examples of segment pairs that are included in MT-GenEval. Explicitly gendered words are underlined, while ungendered words whose translation is explicitly gendered are in *italics*.

case where although a given entity (*Crookall*) is gendered in both the source and the target, the gendered words themselves are different (*his* is marked for gender in English but absent in Spanish, while *master painter* is marked for gender in Spanish but not English). This is in large part what makes gender accuracy in translation non-trivial, especially in real, diverse, and long segments.

We provide a detailed description of the dataset in section 2 and of the associated automatic evaluation metrics in section 3. Section 4 gives an overview of existing gender accuracy benchmarks for machine translation; we hope this overview will enable researchers to assess which benchmarks are appropriate for their use cases. Finally, section 5 evaluates commercial systems as well as models trained on public data on MT-GenEval. We find that: (1) systems trained with contextual and gender-filtered data show improvements in both inter- and intra-sentential gender accuracy as measured by MT-GenEval; and (2) generic (unrelated to gender) translation quality is correlated with gender, exhibiting a new facet of gender in machine translation that is understudied in prior work.

2 MT-GenEval: Gender Translation Accuracy in 8 Language Pairs

For the initial release, MT-GenEval covers translations in two genders (female and male)[3] from English (EN) into eight diverse and widely-spoken target languages: Arabic (AR), French (FR), German (DE), Hindi (HI), Italian (IT), Portuguese (PT), Russian (RU), and Spanish (ES). The source language has limited morphological gender, with gender expressed only on some pronouns and nouns. By contrast, the target languages have extensive grammatical gender and may express gender through morphological markings on a variety of parts of speech including verbs and adjectives, as well as on inanimate objects. In the target languages, human gender often, but not always, lines up with grammatical gender. In order to facilitate evaluation and training of gender-accurate machine translation systems, we release two test subsets (*counterfactual*, Table 2; and *contextual*, Table 3) as well as a counterfactual development set. Each subset is described in more detail below.

2.1 Counterfactual Subset

Data sourcing In developing MT-GenEval, our goal was to create a realistic, gender-balanced dataset that naturally incorporates a diverse range of gender phenomena. To this end, we extracted English source sentences from Wikipedia[4] as the basis for our dataset. We automatically pre-selected relevant sentences using EN gender-referring words based on the list provided by Zhao et al. (2018). In addition to the sentence containing the relevant gendered word(s), we included the two prior sentences in the pre-selection, so as to increase the diversity of gendered words beyond the list used. In a second stage, we asked annotators to manually review these initial candidate segments to ensure that they contain (1) at least one reference to an unambiguously gendered human, (2) no references to individuals of a different gender, and (3) no first names (to avoid confounds where models associate a first name with a gender). Items (2) and (3) are necessary to enable the creation of counterfactual source segments.

Gender balance through counterfactuals In order to ensure that our dataset was fully balanced between female and male genders, as well as to eliminate correlations between gender and content, we asked annotators to manually create *counterfactual* versions of each source segment. Since each source segment refers to individuals of a single gender (in our case, female or male), we were able to create counterfactual version by changing all un-

[3] We recognize that the coverage of only two genders is a limitation of our work. As such, we plan to expand to additional genders in the future.

[4] https://www.wikipedia.org/

	Feminine Source	**Feminine Reference**	**Masculine Source**	**Masculine Reference**
1	Her family moved to the midwest where she was educated and permanently scarred by dour nuns.	Sa famille a déménagé dans le Midwest où elle a été éduquée et irrémédiablement traumatisée par des religieuses austères.	His family moved to the midwest where he was educated and permanently scarred by dour monks.	Sa famille a déménagé dans le Midwest où il a été éduqué et irrémédiablement traumatisé par des moines austères.
2	Many of her short stories have been broadcast on BBC Radio 4.	Muchos de sus relatos cortos se han emitido en BBC Radio 4.	Many of his short stories have been broadcast on BBC Radio 4.	Muchos de sus relatos cortos se han emitido en BBC Radio 4.

Table 2: Counterfactual examples from MT-GenEval. Each source segment refers to individuals of a single gender (in this case, female or male), and counterfactual segments change all unambiguous gender references. Unambiguous gendered words are underlined. In some cases, the reference translation might not be gendered (row 2).

	Context and Source	**Correct Reference**	**Contrastive Reference**
1	Paul intervenes and overpowers him, but he wriggles free. <sep> *The librarian* is then *run over* by a car in front of the library and apparently killed.	El bibliotecario es luego atropellado por un auto enfrente de la biblioteca y al parecer murió.	La bibliotecaria es luego atropellada por un auto enfrente de la biblioteca y al parecer murió.
2	After the war, she continued her career at the Boruprokat factory. <sep> Hasanova was *the* chief *brigadier* in 1970 and led four brigades at that factory.	Hasanova era la brigadiera capo nel 1970 e guidò quattro brigate nella stessa fabbrica.	Hasanova era il brigadiere capo nel 1970 e guidò quattro brigate nella stessa fabbrica.

Table 3: Contextual examples from MT-GenEval. Note that unlike counterfactual examples, reference examples are *contrastive*. Contrastive references are available for the main sentence (which comes after <*sep*>), but not for the context. Unambiguous gendered words are underlined, and their ambiguous translations are in *italics*.

ambiguous references to that gender (e.g., female) to equivalent unambiguous references to another gender (e.g., male). See Table 2 for an example of original and counterfactual sentences.

Reference translations We asked professional translators to create translations for the original segments from scratch, and to use post-editing for the corresponding counterfactual segments, where annotators had access to both the counterfactual and the original source segment during post-editing. This had the effect of eliminating spurious differences in the original and counterfactual translations that were unrelated to gender. Translators were encouraged *not* to introduce gender marking in the translation when such differences would not be natural. Thus, for HI, IT, and ES, several of the gendered inputs have gender-neutral translations.

All annotation was done by professional linguists/translators who are native speakers of the relevant language (English in the case of sourcing and counterfactual creation; target languages in the case of translations). Additionally, annotations were reviewed by professional quality assurance teams to ensure that the data was high-quality. For each step in the process, half the data was created by a female annotator and half by a male annotator. We have released the full text of the annotation instructions on GitHub.[5]

Development and test data The counterfactual **test** set consists of 600 segments (balanced by gender), all of which have gender-specific sources *and* references. Each segment in the test set also has its counterfactual in the test set, which facilitates automatic evaluation (see section 3). We additionally release **development** data, which consists of 2400 sentence-level segments. Unlike the test data, we do not enforce that all reference translations in the development set be gendered. As such, 84.7% of references are gendered for EN-HI, 89.0% for EN-IT, and 89.2% for EN-ES (for the remaining five language pairs, all references are gendered).

2.2 Contextual Subset

Data sourcing In developing the contextual subset of MT-GenEval, our goal was to create a gender-balanced dataset for evaluating gender accuracy and bias in *contextual* MT models. First, using word lists, we automatically pre-selected sentences from Wikipedia that contained at least one mention of a profession and no gendered words. The selected professions fall into one of three categories: stereotypical female, stereotypical male and neutral (Troles and Schmid, 2021). Additionally, the professions were selected to lack gender marking in English, while potentially requiring gender inflection and agreement in the target languages. To remove any further gender cues, sentences contain-

[5] `https://github.com/amazon-research/machine-translation-gender-eval`

	Person Gender	
Profession	Female	Male
Female	150	150
Male	150	150
Neutral	250	250
Total	550	550

Table 4: Number of source sentences in each of the six sub-categories of the contextual subset of MT-GenEval.

ing commonly used first names were excluded.[6] Annotators were subsequently asked to manually review the remaining segments to ensure that (1) they were indeed gender-ambiguous on the segment level and (2) they contained mentions of exactly one individual.

Context For each of the selected gender-ambiguous sentences, we extracted the two preceding context sentences. We took a semi-automatic approach to verifying whether the context sentences disambiguated the gender of the selected sentence, first checking for the presence of gendered words and subsequently asking annotators to mark which context sentence disambiguated the selected sentence (none, one of them, or both). This yielded a set of gender-ambiguous sentences referring to a single individual along with at least one preceding context sentence that linguistically disambiguated the gender of that individual. We give examples of selected source sentences in Table 3, where we use the <*sep*> token to delimit the context and main sentence.

Gender balance To ensure the dataset was balanced, we selected an equal number of source examples for both female and male genders, as well as for each profession category. As a result, the dataset contains both stereotypical and anti-stereotypical examples and covers six sub-categories in total, as shown in Table 4.

Reference translations In addition to the original source segments, we release two contrastive reference translations for each main sentence (references for context sentences are not included in the dataset). One of the reference translations correctly translates the gender of the individual, while the contrastive reference changes the gender (e.g., female to male). For example, as shown in Table 3, the ambiguous noun phrase "the librarian" is translated as "el bibliotecario" in the correct reference and as "la bibliotecaria" in the contrastive reference. We asked translators not to introduce unnecessary gender marking (similar guidelines as for the counterfactual subset) and we excluded from this subset examples where the contrastive translations were identical (no gender marking).

3 Automatic Metrics for MT-GenEval

3.1 Gender Accuracy

All segments in our test set include both correct and contrastive/counterfactual references. To automatically evaluate gender accuracy in translation on this set, we propose a straightforward accuracy metric based on the fact that, by design, the correct and contrastive references differ only in gender-specific words.

We define accuracy of gender in translation on our test set as follows.[7] Let w_{hyp}, w_{ref}, and w_{con} denote the set of words in the hypothesis, reference, and contrastive reference, respectively. First, we obtain the set of words in the contrastive reference that are not in the correct reference:

$$unique_{con} = w_{con} \setminus w_{ref} \quad (1)$$

This removes from consideration all the words that are unrelated to the gender of the individual(s) in the source, as the correct reference and contrastive reference do not have any non-gender-related differences. We consider a segment *incorrect* if:

$$unique_{con} \cap w_{hyp} \neq \emptyset \quad (2)$$

i.e., if the hypothesis contains words specific to the contrastive (incorrect) gender.

To evaluate our metric, we ran human evaluations of gender accuracy on a subset of the contextual set. We selected a stratified sample of 600 source segments, translated each with three commercial systems, and asked two professional translators to mark the gender correctness of the system outputs. Further details on the evaluation task are provided in section 5.1, including inter-annotator agreement scores. Table 5 shows average F-score of the automatic accuracy metric with respect to human annotations in this evaluation. The automatic

[6] We used publicly available US and UK census data from https://github.com/OpenGenderTracking/globalnamedata.

[7] We considered other approaches to defining gender translation accuracy, e.g., based on model score on the correct vs. contrastive references. However, we found that these had lower agreements with human scores in initial experiments.

metric matches humans reasonably well across all language pairs, with F-scores consistently at 0.80 or higher.

	F-score
EN→AR	0.84
EN→DE	0.82
EN→ES	0.89
EN→FR	0.86
EN→HI	0.80
EN→IT	0.83
EN→PT	0.86
EN→RU	0.84

Table 5: F-score between automatic accuracy metric and human accuracy labels on the contextual set.

Gender accuracy on the counterfactual subset is defined similarly.[8] However, on the counterfactual subset we have pairs of counterfactual source segments as well as counterfactual references. We consider a segment pair as correct only if *both* the original and the counterfactual segment are marked as correct. This is to reward models for cases where they are actually predicting correct gender based on the input, rather than randomly guessing.

3.2 Gender Quality Gap

While the gender accuracy metric introduced in section 3.1 evaluates gender translation at the lexical level, *generic translation quality* may also vary across the inputs and may be correlated with gender. For this reason, we complement the accuracy metric with a **gender quality gap** metric, $\boldsymbol{\Delta_{\mathrm{qual}}}$. This allows us to measure representational bias (Blodgett et al., 2020), expressed as lower quality for one of the two genders considered, on MT-GenEval.

We evaluate Δ_{qual} on the counterfactual subset, where we can abstract away non-gender-related content differences (since for a given sentence, its semantically equivalent gender-counterfactual is always in this test set). We define Δ_{qual} as:

$$\Delta_{\mathrm{qual}} = \mathrm{BLEU}_{\mathrm{male}} - \mathrm{BLEU}_{\mathrm{female}} \qquad (3)$$

where $\mathrm{BLEU}_{\mathrm{gender}}$ is the BLEU score of the $gender$ subset of the counterfactual test set.

4 Gender Evaluation Benchmarks for MT

In this section, we review existing benchmarks on gender accuracy in machine translation, in order contextualize MT-GenEval with respect to similar work. Table 6 summarizes these benchmarks. Note that we focus our analysis on evaluation of gender translation accuracy in this section; see Sun et al. (2019) for a more general review of gender in natural language processing, and Savoldi et al. (2021) for a summary of work on gender bias in MT.

WinoMT One of the most widely-used datasets for evaluating gender accuracy in machine translation is WinoMT (Stanovsky et al., 2019; Kocmi et al., 2020), which was created by combining the Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) coreference test sets. As such, WinoMT contains synthetic Winograd-style sentences where gender is associated with pro-stereotypical and anti-stereotypical professions.[10] While the dataset does not contain reference translations, it instead includes a target language-specific alignment-based automatic metric. WinoMT is one of the gender translation accuracy benchmarks with the largest language coverage: the metric covers translation from English into AR, Czech (CS), DE, ES, FR, Hebrew (HE), IT, Polish (PL), RU, and Ukrainian (UK).

MuST-SHE MuST-SHE (Bentivogli et al., 2020; Savoldi et al., 2022) is a dataset of roughly 1,000 gender-specific segments for each of EN→ES, FR, and IT. Segments include both text and audio, and are extracted from TED talks (Cattoni et al., 2021). The dataset contains segments where gender is disambiguated by the intra-sentential context, as well as segments where gender is only present as speaker metadata. It is curated to only include segments containing at least one gender-neutral source word that requires gender marking in the translation. The dataset also provides contrastive references for each segment. MuST-SHE was also extended with annotations of agreement chain and part of speech by Savoldi et al. (2022), and with source-side gender annotations by Vanmassenhove and Monti (2021).

GeBioCorpus and Translated Wikipedia Biographies GeBioCorpus (Costa-jussà et al.,

[8] Initial human evaluations on this set were unreliable, with very low inter-annotator agreements. As such, we leave evaluation of the metric on the counterfactual subset for future work, as this is a relatively difficult task for which annotators need more training.

[10] Recently, Troles and Schmid (2021) extended WinoMT to include sentences with specific gender-stereotypical adjectives and verbs.

Benchmark	Languages	Size	Data Type	Advantages	References?	Metric?
MT-GenEval (**ours**)	AR, DE, EN, ES, FR, HI, IT, PT, RU	4,008	natural	counterfactual sources and references; inter-sentential context; gender quality gap	✓	✓
GeBioCorpus (Costa-jussà et al., 2020)	CA, EN, ES	2,000	natural	extraction pipeline	✓	×
MuST-SHE (Bentivogli et al., 2020)	EN, ES, FR, IT	1,095	natural	counterfactual references; aligned audio	✓	✓
SimpleGEN (Renduchintala et al., 2021)	DE, EN, ES	1,332	synthetic	stereotype annotations	×	✓
Translated Wikipedia Bios[9]	DE, EN, ES	138	natural	inter-sentential context	✓	×
WinoMT (Stanovsky et al., 2019)	AR, CS, DE, EN, ES, FR, HE, IT, PL, RU, UK	3,888	synthetic	stereotype annotations; easily extensible to new target languages	×	✓

Table 6: Summary of benchmarks for gender accuracy in machine translation. Dataset sizes listed are per-language (for benchmarks with different sizes per language, we take the smallest) and per-segment (sentence pair or document).

2020) and Translated Wikipedia Biographies[11] are closely related efforts that extract gender-related machine translation benchmarks from Wikipedia biographies. GeBioCorpus, which covers EN↔ES↔Catalan (CA), consists of gender-balanced parallel sentences that are automatically extracted and aligned using the GeBioToolkit.[12] A subset of 2,000 of the automatically extracted segments were reviewed by humans to yield an evaluation set. By contrast, Translated Wikipedia Biographies consists of professional human translations of Wikipedia biographies. It covers EN→DE and EN→ES and contains 138 documents, each with 8-15 sentences. This allows for evaluation of gender disambiguation in inter-sentential context. Note that both of these biography-based sets may contain irrelevant segments that have no gender information in either the source or the target.

Other benchmarks Renduchintala et al. (2021) introduced SimpleGEN, which consists of relatively short, synthetic source phrases focusing on professions and verbs. The dataset is annotated for pro- and anti-stereotypicalness and probes for translations with ungrammatically mixed gender. Also based on professions is the occupations test set introduced by Escudé Font and Costa-jussà (2019), which consists of 1,000 human-translated EN→ES sentence pairs that follow a simple pattern.

5 Benchmarking Models with MT-GenEval

In this section, we benchmark both commercial and research-scale models using MT-GenEval. In addition to giving initial baseline numbers for MT-GenEval, through these experiments we aim to show that MT-GenEval is a useful new benchmark. Specifically, we show that:

1. MT-GenEval data is high-quality, and the benchmark is difficult even for state-of-the-art (SOTA) systems (section 5.1).

2. MT-GenEval measures a novel aspect of gender in translation that is absent from existing benchmarks (section 5.2).

3. MT-GenEval is able to discriminate between models that are trained to improve contextual translation and translation of gender (section 5.3).

5.1 Context-Level Gender Accuracy in Commercial Systems

In this section, we evaluate three industrial systems on the contextual subset of the MT-GenEval benchmark. Our goal is to show that the dataset is sufficiently diverse and challenging even for systems trained on web-sized corpora.

We used human evaluations to measure the gender accuracy of the translation outputs for each system for all eight target languages in the benchmark. To anonymize the commercial systems, we label them A, B, and C. In order to anchor the eval-

[11] https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html
[12] https://github.com/PLXIV/Gebiotoolkit

	Online Systems				Reference	
EN→	Acc-A	Acc-B	Acc-C	IAA	Acc	IAA
AR	51.5	51.3	50.5	0.93	95.8	0.72
FR	56.1	55.9	56.2	0.82	98.3	0.25
DE	52.3	51.3	54.0	0.96	92.2	0.84
HI	59.7	61.1	61.3	0.70	97.1	0.25
IT	55.9	54.9	55.7	0.86	97.4	0.65
PT	50.9	52.3	52.7	0.89	91.4	0.68
RU	56.9	57.2	57.7	0.77	97.0	0.16
ES	57.0	58.2	58.5	0.94	97.6	0.89

Table 7: Gender accuracy scores (*Acc*) and inter-annotator agreement (*IAA*) measured on the contextual subset for the three anonymized commercial systems and the correct reference translation.

uations, we additionally included reference translations in the evaluation.[13]

We asked two professional translators to judge gender accuracy in context[14] using a stratified sample of 600 source segments (100 for each subcategory in Table 4). We show the accuracy scores and inter-annotator agreement (IAA) computed with Krippendorff's alpha (Hayes and Krippendorff, 2007) in Table 7. We first note that IAA for the system outputs is high (> 0.85) for the majority of target languages, with the exception of Hindi, Russian and French. For these three languages we also observe poor agreement (≤ 0.25) on judging the correct reference translation. This indicates that some evaluators may have exhibited bias regarding gender accuracy judgments and over-penalized correct reference translations. For computing the system-level accuracy scores, we use the judgments of the "preferred" evaluator, which is the evaluator who judged the reference translation as correct more times than the other annotator.

Across languages, the accuracy of these industrial systems is close to 50%, ranging from 50.5% for AR to a high of 61.3% for HI. This indicates that they are largely not able to effectively take context into account to disambiguate the gender of the input. Thus, MT-GenEval is difficult even for SOTA industrial systems, despite the fact that they are trained on very large data that might actually include Wikipedia (which was used as the source for MT-GenEval).

Additionally, for the majority of languages, evaluators found the correct reference translations $>$ 95% accurate, confirming the high quality of the dataset. However, for PT and DE, we observe a slightly lower reference accuracy (91.4% and 92.2%).

EN→	Δ_{qual} **A**	Δ_{qual} **B**	Δ_{qual} **C**
AR	0.3	0.4	0.2
DE	13.0	12.3	13.0
ES	11.0	11.0	11.3
FR	9.2	11.0	10.9
HI	6.0	6.4	7.5
IT	9.2	9.0	9.1
PT	12.3	12.9	13.2
RU	13.8	12.9	15.3

Table 8: Gender quality gap (lower magnitude is better) for the three anonymized commercial systems. Gender quality gap is defined as the difference in quality on the male and the female subsets (see section 3.2).

5.2 Evaluating Commercial Systems for Gender Quality Gap

Next, we evaluate the gender quality gap Δ_{qual} on the counterfactual test set for the same three industrial systems. These results are shown in Table 8.

With the exception of EN→AR, the results show a clear pattern where the overall quality on masculine inputs is much higher (9.6 points on average) than the overall quality on feminine inputs (even though the inputs are identical aside from gender). Based on these results, as well as on initial observations regarding examples such as the one shown in Table 9, we ran a pilot analysis to see whether the automatically computed gender quality gap is indeed visible in generic (non-gender-related) quality as judged by humans. For EN→DE, we extracted the 50 sentences from the test set that had the largest gap (i.e., $BLEU_{male} > BLEU_{female}$) and the smallest gap (i.e., $BLEU_{female} > BLEU_{male}$). A native German speaker manually checked whether the quality and gender translation accuracy differed between the female-referring and male-referring outputs. We found that most of the segments with a gender quality gap did indeed have meaningful differences in translation quality on portions of the segment unrelated to gender, even though those portions were (by design) identical in the source. Additionally, for the segments where male-referring translations were better, the gap in human-perceived quality was much larger than for the segments where female-referring translations were better.

To our knowledge, the observation that there can

[13] Evaluators were not aware one of the translations was a human reference, and the order of the translations was shuffled.

[14] We showed translators both the context and the main sentence, but asked them to evaluate only the translation of the latter.

src:	We had to repair our relationship **because I wanted my mother/father back**.
fem:	Wir mussten unsere Beziehung reparieren, **weil ich meine Mutter pflegte**.
msc:	Wir mussten unsere Beziehung reparieren, **weil ich meinen Vater wollte**.

Table 9: Model output from the GFST-CTX system (section 5.3) where the gender accuracy is correct, but the feminine (fem) input leads to a lower-quality output than the counterfactual masculine (msc) input. In the feminine translation, "I wanted my mother back" is translated incorrectly as "I took care of my mother", whereas the masculine translation is closer to the source: "I wanted my father".

be gaps in quality beyond gender-related words for otherwise equivalent inputs referring to different genders is novel.[15] Since MT-GenEval contains realistic and counterfactual data, it is now possible to evaluate models for these quality differences while controlling for content.

5.3 Contextual Gender Accuracy with Contextual and Gender-Balanced Models

In this section, we use MT-GenEval to benchmark both contextual and gender-balanced NMT models trained on publicly available data. This helps us understand how existing methods for training these models affect performance on MT-GenEval. We focus on three language pairs: EN→DE, EN→FR, and EN→RU. For each pair, we build four models:

- BASE: Non-contextual baseline
- CTX: Model trained with additional contextual data in the 2+2 format (Tiedemann and Scherrer, 2017), following Majumder et al. (2022)
- GFST: Model trained with additional gender-filtered self-training (GFST) data from Choubey et al. (2021)[16]
- GFST-CTX: Model trained with both the GFST data and the 2+2 CTX data

The Transformer-base architecture (Vaswani et al., 2017) is used to train all the NMT models, with tied weight matrices for the source embeddings, target embeddings, and output layer. However, we use 8 decoder layers instead of 6, following the recommendation of Majumder et al. (2022). Training is done using Sockeye 3 (Hieber et al., 2022). For training data, we use WMT19 (Barrault et al., 2019) for EN→DE and OpenSubtitles (Lison and Tiedemann, 2016) for EN→FR and EN→RU, all of which contain document-level data (used to train CTX models). We use the dev sets from WMT19 (DE, RU) and IWSLT 2019 (FR) (Niehues et al., 2019) for development.

	Contextual			Counterfactual		
EN→	DE	FR	RU	DE	FR	RU
BASE	66.7	66.1	62.5	71.0	63.0	79.7
CTX	73.6	**69.3**	65.0	71.0	63.0	80.7
GFST	65.5	65.9	61.7	70.3	72.0	87.0
GFST-CTX	**77.1**	68.8	**68.1**	**76.0**	**75.3**	**91.0**

Table 10: Automatic accuracy scores on MT-GenEval for the systems trained on public data.

Table 10 shows automatic accuracy scores for each system on both subsets of MT-GenEval (contextual and counterfactual).[17] On the contextual subset, as expected, we see much higher accuracy when a model is trained to take inter-sentential context into account: CTX is better than BASE and GFST-CTX is better than GFST. This confirms that our test set is both sensitive to gender in context and challenging for vanilla contextual models (accuracy is below 75%). We find that we can improve the accuracy of contextual models by combining gender-filtered and contextual data: GFST-CTX is better than CTX for German and Russian. On the other hand, gender balancing somewhat decreases the performance of non-contextual models (BASE vs. GFST) on the contextual subset. Considering that the BASE accuracy is higher than 50%, this result indicates that non-contextual systems may be inferring gender from some source words that correlate with gender although they do not mark it explicitly (also observed for commercial systems in Table 7). Thus, the lower accuracy on the contextual set for GFST compared to BASE could indicate a less gender-stereotypical model.

On the counterfactual subset, the GFST data improves gender translation accuracy significantly overall, supporting the findings of Choubey et al. (2021) on WinoMT and MuST-SHE. This confirms that our counterfactual subset is also sensitive to changes in gender balance in the training data. A

[15] A similar observation has been made for translation quality based on the demographics of the *author* rather than the referred individual (Hovy et al., 2020; Rabinovich et al., 2017; Vanmassenhove et al., 2018), although such work could not enforce the inputs being otherwise equivalent.

[16] `https://github.com/amazon-research/gfst-nmt`

[17] Quality scores are shown in Appendix C.

surprising finding is that the contextual data improves the performance on non-contextual inputs in the counterfactual subset (GFST-CTX is better than GFST). This aligns with prior work showing that adding contextual training data can introduce noise that acts as a regularizer (Kim et al., 2019), and that adding contextual data can help reduce gender bias in MT models (Basta et al., 2020).

6 Conclusions

In this paper, we have introduced **MT-GenEval**, a counterfactual and contextual benchmark for evaluating gender accuracy in translation from EN into AR, DE, ES, FR, HI, IT, PT, and RU. In addition to the test data and evaluation metrics, we are releasing 2400 segments of development data.[18] We have shown that this benchmark is useful for evaluating both commercial and research systems, including contextual machine translation models and gender-balanced models, in terms of gender accuracy as well as quality. We hope that this benchmark and development data will spur more research in the field of gender accuracy in translation on diverse languages.

Acknowledgments

We would like to thank Tanya Badeka and Jen Wang for their help in creating the MT-GenEval data, Kathleen McKeown and Brian Thompson for their feedback on earlier drafts of this paper, and the anonymous reviewers for their suggestions.

7 Limitations

Dataset coverage The main limitation of this benchmark is that it only covers two human genders: female and male. Additionally, the language pairs covered in the benchmark all exhibit a similar pattern: language with limited grammatical gender (English) → language with morphological grammatical gender. It is not clear whether this benchmark could be expanded to more language pairs or used in the reverse direction. Finally, due to the counterfactual nature of the set, we excluded data containing individuals with different genders, as well as data with first names, which could bias the evaluation.

Annotator bias and errors In dataset creation, we relied heavily on human annotators, both for generating the counterfactual versions of the original sentences, and for translation into the target languages. Although we endeavored to mitigate biases in annotators by providing explicit instructions and examples, as well as by drawing annotators from a diverse population, it was not possible to eliminate such biases completely. For example, in the source annotation phase, an annotator created a male counterfactual of the sentence "Pekgul trained as a nurse, a profession in which she worked both before and after her election as a politician." by changing "nurse" to "male nurse". This was prohibited by the instructions as the word "nurse" is already gender-neutral in English, and it exhibits the annotator's subtle bias that nurses are by default female. MT-GenEval could contain additional annotator errors, both in sources and in references.

Limitations of the accuracy metric Our proposed accuracy metric relies on overlap with the reference, and as such it will not necessarily be reliable when translations of gendered words use synonyms that are not present in the reference. Additionally, this metric has yet not been compared to human scores for the counterfactual set due to unreliable human evaluations on that set.

8 Ethical Considerations

In this paper, we release MT-GenEval, a benchmark for evaluating gender accuracy and quality in machine translation. We hope that this benchmark will be useful in evaluating representational harms related to gender in machine translation, particularly lower quality and accuracy in translation based on gender. The benchmark focuses on inputs that contain unambiguously gendered references to humans, and as such does not infer or assign gender in any way. Additionally, gender is not used as a variable in our work and we do not work with human subjects. We sourced our data from Wikipedia articles, which are publicly available and have a relatively low risk of inclusion of private information.

As discussed in section 7, the main limitation of our work is that evaluations are limited to two genders (female and male). We hope to be able to expand this work to more genders in the future.

In creating our annotations, we worked with a language service provider to contract with professional translators and ensure suitable working conditions for them. Annotators were compensated in accordance with translation industry standards.

[18] We give suggested applications for the development set in Appendix B.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

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

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Andrew Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77–89.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with "pytorch". *arXiv preprint arXiv:2207.05851*.

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, Online. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

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.

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, Portorož, Slovenia. European Language Resources Association (ELRA).

Suvodeep Majumder, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Maria Nădejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, USA. Association for Computational Linguistics.

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

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. 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 and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. 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.

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

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

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

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

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

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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.

A Descriptive Statistics of MT-GenEval

Table 12 shows representative data statistics comparing MT-GenEval (development subset) with WinoMT and MuST-SHE v1.0. We select four languages for illustration. As can be observed, MT-GenEval is more diverse than the other two datasets, both when it comes to overall vocabulary as well as to the diversity of gendered phrases, particularly in the target langauges.

B Potential Uses for MT-GenEval Development Set

In creating and releasing the MT-GenEval set, we hope to cover several potential use cases for improving and evaluating gender accuracy in translation. The counterfactual and contextual test sets allow gender translation accuracy to be evaluated on both the sentential and inter-sentential levels for translation from English into eight languages, with data releases for additional languages planned for the future. Additionally, translation evaluation in the reverse direction (i.e., *→English) is possible for the counterfactual set, since this set was constructed so that most sentence pairs are marked for gender on both the source and the target sides.

We anticipate that the 2,400 development sentences released for each language pair will be useful in training models to improve gender translation accuracy. Since the development set consists of gender-balanced counterfactual sentences, it can be used in gender fine-tuning as introduced by Saunders and Byrne (2020), with the added advantage that the MT-GenEval development data is naturally occurring and more complex than artificially constructed segments used in their original work. As an alternative, this data can be potentially used to train a model that generates counterfactuals automatically, instead of relying on rule-based gender counterfactuals as in prior work. Other prior work on improving gender in translation used wordlists and morphological taggers to extract gender-specific data from a generic corpus (Choubey et al., 2021); the counterfactual MT-GenEval development data that we are releasing could generalize this process by being used to train a classifier that automatically detects gender-specific segments.

C Quality Scores on Contextual and Gender-Balanced Models

Table 11 shows gender-specific BLEU scores on the models trained on public data from section 5.3. Unlike for the commercial systems (section 5.2), we do not see a large gender quality gap in these models. However, BLEU scores are quite low, particularly for EN→RU, possibly due to domain mismatch (Wikipedia vs. WMT/OpenSubtitles).

	$BLEU_{female}$			$BLEU_{male}$		
EN→	DE	FR	RU	DE	FR	RU
BASE	12.9	12.9	8.0	12.7	13.4	8.7
CTX	11.1	12.8	6.8	11.7	13.8	7.8
GFST	12.7	13.8	8.5	13.6	14.4	9.5
GFST-CTX	12.8	14.5	7.6	13.2	15.2	8.6

Table 11: Automatic accuracy scores on the contextual and counterfactual subsets for the systems trained on public data.

			Source		**Target**	
Dataset	EN →	# instances	# distinct words	# distinct gendered phrases	# distinct words	# distinct gendered phrases
MT-GenEval	AR	1,200	6,690	350	11,194	3,890
	DE		6,604	328	8,053	1,091
	FR		6,575	369	7,640	1,480
	RU		6,619	348	10,121	2,253
WinoMT	–	1,944	1,883	–	–	–
MuST-SHE	FR	1,113	4,605	–	5,792	1,456

Table 12: Representative data statistics of MT-GenEval, WinoMT, and MuST-SHE v1.0.