

GATE: A Challenge Set for Gender-Ambiguous Translation Examples

Spencer Rarrick Ranjita Naik Varun Mathur Sundar Poudel Vishal Chowdhary^{*†}

Abstract

Although recent years have brought significant progress in improving translation of unambiguously gendered sentences, translation of ambiguously gendered input remains relatively unexplored. When source gender is ambiguous, machine translation models typically default to stereotypical gender roles, perpetuating harmful bias. Recent work has led to the development of "gender rewriters" that generate alternative gender translations on such ambiguous inputs, but such systems are plagued by poor linguistic coverage. To encourage better performance on this task we present and release GATE, a linguistically diverse corpus of gender-ambiguous source sentences along with multiple alternative target language translations. We also provide tools for evaluation and system analysis when using GATE and use them to evaluate our translation rewriter system.

1 Introduction

Gender is expressed differently across different languages. For example, in English the word *lawyer* could refer to either a male or female individual, but in Spanish, *abogada* and *abogado* would be used to refer to a female or a male lawyer respectively. This frequently leads to situations where in order to produce a single translation, a translator or machine translation (MT) model tends to choose an arbitrary gender to assign to an animate entity in translation output where it was not implied by the source. In this paper, we refer to this phenomenon as *arbitrary gender marking* and to such entities as Arbitrarily Gender-Marked Entities (AGMEs).

Translation with arbitrary gender marking is a significant issue in MT because these arbitrary gender assignments often align with stereotypes, perpetuating harmful societal bias (Stanovsky et al.,

2019; Ciora et al., 2021). For example, MT models will commonly translate the following (from English to Spanish):

The surgeon \xrightarrow{MT} *El cirujano* (m)

The nurse \xrightarrow{MT} *La enfermera* (f)

Progress has been made to remedy this using a "gender rewriter" – a system that transforms a single translation with some set of gender assignments for AGMEs into a complete set of translations that covers all valid sets of gender assignments for a source sentence into the target language (Prates et al., 2018). Using a rewriter:

The surgeon

⇓ MT

El cirujano (m)

⇓ rewriter

La cirujana (f)

El cirujano (m)

Although a step in the right direction, these rewriters often have poor linguistic coverage and only work correctly in simpler cases. Google Translate has publicly released such a system for a subset of supported languages, and we observe two error cases¹:

1. It does not rewrite when necessary: *The director was astonished by the response of the community.* produces only one translation corresponding to masculine director.
2. It rewrites partially, or incorrectly: *I'd rather be a nurse than a lawyer* produces two translations but only lawyer is reinflected for gender (nurse is feminine in both).

^{*}All authors are affiliated with Microsoft.

[†]Contact author at vishalc@microsoft.com.

¹as observed on Mar 6, 2023

To facilitate improvement in coverage and accuracy of such rewriters and reduce bias in translation, we release GATE², a test corpus containing gender-ambiguous translation examples from English (en) into three Romance languages ((Vincent, 1988)): Spanish (es), French (fr) and Italian (it). Each English source sentence³ is accompanied by one target language translation for each possible combination of masculine and feminine gender assignments of AGMEs⁴:

*I know a **Turk** who lives in Paris.*
 ↓↓ it
*Conosco **una turca** che vive a Parigi. (f)*
*Conosco **un turco** che vive a Parigi. (m)*

GATE is constructed to be challenging, morphologically rich and linguistically diverse. It has ~ 2000 translation examples for each target language, and each example is annotated with linguistic properties (coreferent entities, parts of speech, etc.). We additionally propose a set of metrics to use when evaluating gender rewriting modules.

This corpus was developed with the help of bilingual linguists with significant translation experience for each of our target languages (henceforth *linguists*). Each is a native speaker in their respective target language. We spoke in depth with our linguists about the nuances of gender-related phenomena in our focus languages and we share our analysis of the relevant aspects and how they impact our work and the task of gender rewriting.

Along with the corpus, we also provide tools for evaluation and system analysis when using GATE and use them to evaluate our own translation rewriter system.

2 GATE Corpus

We present GATE corpus, a collection of bilingual translation examples designed to challenge source-aware gender-rewriters. The linguists were asked to compile roughly 2,000 examples for each target language, with the hope that this would be sufficient for good variety along several dimensions: sentence lengths, sentence structures, vocabulary diversity, and variety of AGME counts.

²Data and evaluation code available at <https://github.com/MicrosoftTranslator/GATE>

³A few non-sentence utterances are also included as well, such as noun-phrases and sentence fragments

⁴The majority of source sentences contain only one AGME and thus two translations

2.1 Anatomy of an Example

Each example in the data set consists of an English sentence with at least one AGME, and a set of alternative translations into the given target language corresponding to each possible combination of male/female gender choices for each AGME. Variation among the alternative translations is restricted to the minimal changes necessary to naturally and correctly indicate the respective gender-markings.

We also mark several category features on each example, such as what class of animate noun AGMEs belong to (profession, relationship, etc), what grammatical role they play in the sentence (subject, direct object, etc), sentence type (question, imperative, etc) and several other phenomena. These are discussed in more detail in Appendix A, as well as statistics over each language’s corpus.

Additionally, each example is accompanied by a list of AGMEs as they appear in the English source, as well as their respective masculine and feminine translations found in the translated sentences. For multi-word phrases, we asked annotators to enclose the head noun in square brackets. For example, if *police officer* is translated to *policía* in Spanish, the English field could include *police [officer]*.

The same entity may be referred to multiple times in the same sentence through coreference. We asked annotators to indicate coreferent mentions of AGMEs are by joining them with ‘=’. For example, in the following *en-es* example, the English AGME field would contain "nurse=lawyer".

*I’d rather be a **nurse** than a **lawyer**.*
 ↓↓ es
*Prefiero ser **enfermera** que **abogada**. (f)*
*Prefiero ser **enfermero** que **abogado**. (m)*

Finally, In cases where an AGME is represented by a pronoun that is elided in the translation, it will be represented by the nominative case form and be enclosed in parentheses. For example, in the following example, the Spanish AGME field would contain (*yo*):

*I am **tired**.*
 ↓↓ es
*Estoy **cansada**. (f)*
*Estoy **cansado**. (m)*

2.2 Arbitrarily Gender-Marked Entities

In this paper, we use *animate entity* (or just *entity*) to refer to an individual or group for which a refer-

Data Set	< 10	10-19	20-29	>= 30	Total
Spanish 1 AGME	477	722	197	105	1,501
Spanish 2+ AGMEs	70	176	56	21	323
French 1 AGME	704	661	171	14	1,550
French 2+ AGMEs	177	222	41	4	444
Italian 1 AGME	397	867	195	48	1,507
Italian 2+ AGMEs	93	500	139	30	762

Table 1: Distribution of lengths (words) of English utterance per target language and AGME count

ential gender could be implied in either the source or target language⁵. Usually this will refer to humans, but may also be extended to some animals and mythical or sentient beings. For example, *cat* is generally translated into Spanish as *gato*, but *gata* is also frequently used to refer to a female cat. Following Dahl (2000), we use *referential gender* to refer to an entity’s gender as a concept outside of any linguistic considerations.

To qualify as an AGME, an entity’s referential gender must be ambiguous in the source sentence, but implied by one or more words in the target translation. Compared to Romance languages, there are relatively few ways that gender is denoted through word-choice in English. Most notably, English uses a handful of gendered pronouns and possessive adjectives (*she, her, hers, he, him, his*), as well as a relatively small number of animate nouns that imply a gender (e.g. *mother, father, masseuse, masseur*, etc). There is also often a correlation between certain proper names and referential gender (e.g. *Sarah* is traditionally a female name and *Matthew* is traditionally male), but we do not consider this a reliable enough signal for gender determination unless they are a well known public figure (e.g. *Barrack Obama* is known to be male). We follow Vanmassenhove and Monti (2021) in this.

Additionally, an AGME must have some gender marking in the translation. In the following English-Italian example,

I heard the thief insult his interlocutor.

⇓ *it*

Io ho sentito il ladro insultare la sua interlocutrice.

Io ho sentito il ladro insultare il suo interlocutore.

interlocutor → *interlocutrice* (f) / *interlocutore* (m) is an AGME, while *thief* → *ladro* and *I* → *Io* are not. *Thief* is unambiguously male because of its

⁵For simplicity, we limit our discussion of gender and linguistics to masculine and feminine within the scope of this paper, but we do not intend to imply that gender is limited in this way.

coreference with *his* in the source, while *I* has ambiguous gender which is not marked in the target.

2.3 Corpus Development Process

The linguists were asked to aim for a distribution of sentences lengths ranging from very short (< 10 words) to complex (> 30) words. Actual example counts are shown in Table 1. Of the 2,000 examples for each language, linguists were asked to include roughly the following breakdown:

- 1,000 single animate noun AGME
- 500 single pronoun AGME
- 500 with two or more AGMEs

Linguists were given details of the various categories and attributes listed in section A and asked to find sentences such that each such category is well represented (depending on the relative ease of finding such sentences). Linguists were also asked to prioritize diversity of animate nouns where possible. They were allowed to pull examples sentences from natural text or construct them from scratch as they saw fit. However, except for a small number of toy examples, we asked that they include only sentences that were natural in both English and their target language, and could reasonably appear in some imaginable context.

We provided samples of web-scraped data that had been filtered with various heuristics to help identify sentences fitting some of the harder-to-satisfy criteria. For example, we used Stanza (Qi et al., 2020) to filter some web-scraped data for those containing an animate noun marked as an indirect object and provided this to the linguists. In some cases these sentences were used directly, and in others they were modified slightly to fit the requirements.

Throughout the process, we prioritized diversity of sentence structure, domain and vocabulary. Rather than produce a representative sample, our intention was to produce a corpus that would chal-

lence any tested systems on a wide range of phenomena.

3 Evaluation with GATE

3.1 Gender Rewriting

Our goal in developing this corpus is to facilitate the generation of multiple translations covering all valid gender assignments. One strategy for producing such a set of translations is to first use an MT model to produce a default translation and then use a rewriter to generate one or more alternative translations with other gender assignments (Prates et al., 2018).

$$\text{source} \xrightarrow{\text{MT}} \text{translation} \xrightarrow{\text{rewriter}} \{\text{all translations}\}$$

3.2 Evaluation Methodology

We formalize the task of gender rewriting on a single-AGME sentence as follows: given the source sentence src , target translations corresponding to male and female referent entities, and a rewrite direction (M to F or F to M), produce an output target translation with the alternative gender from the original translation. We will refer to the original input translation as tgt_0 , the desired/reference translation as tgt_1 and the output generated by the rewriter as hyp :

$$\text{rewriter}(src, tgt_0) = hyp \sim tgt_1$$

For this task, we consider looking at exact full-sentence matches between hyp and tgt_1 to be the most sensible approach for evaluation. We do not give partial credit for changing the gender markings on only a subset of the words to those found in tgt_1 . Doing so will generally result in a sentence that is either grammatically incorrect due to newly introduced agreement errors, or for which the semantics has changed in an unacceptable way, such as a changed coreference. Because of this, we find sentence-similarity measures such as BLEU (Papineni et al., 2002) and words error rate not to be reflective of a user’s experience.

The rewriter may also produce a null output, meaning that only the default translation will be produced. This is necessary because in real-world scenarios, many sentences will not contain AGMEs. When AGMEs are present, it may still be preferable to produce null output over a low confidence rewrite if accuracy errors are judged to be more costly than coverage errors.

We calculate precision as the proportion of correct alternatives among those attempted, i.e. that were non-null outputs. Because there are no true negatives in GATE, recall can be calculated as the proportion of correct alternatives produced among all sentences, including null outputs. Using these definitions of precision and recall, we also find $F_{0.5}$ to be a useful overall metric, prioritizing precision while still incorporating coverage.

While we have focused our discussion of evaluation on sentences containing a single AGME, which typically should produce exactly two alternative translations, GATE also includes a smaller number of examples with more than one AGME. These have more than two alternative translations and thus more than one correct output for a rewriter. We do not formalize evaluation on this subset here but believe that the data set will be useful in evaluating rewriting systems capable of producing multiple outputs for multiple sets of gender assignments.

3.3 System Overview

We use GATE to evaluate our translation gender rewriter, which follows a pipeline approach, roughly similar to Habash et al. (2019).

The system receives as input the original source sentence (src) and a default translation (tgt_0) with the specified language pair. The following components are then applied:

AGME Identifier – We first attempt to find AGMEs in the sentence pair to determine whether rewriting is appropriate. We leverage an AllenNLP coreference model to detect ambiguously gendered entities in the source sentence (Lee et al., 2018). We use a dependency parse generated by Stanza (Qi et al., 2020) and a gendered vocab list to identify gender-marked animate entities in the target sentence.

Candidate Generator - For each word position in tgt_0 , we use a lookup table to find all possible alternate gender variants for the word in that position. We compose the word-level variant sets to build a set of sentence-level hypotheses, while applying grammatical constraints to prune incoherent hypotheses. This yields a set of candidate rewrites.

Translation Scorer - Finally, we use a Marian translation model (Junczys-Dowmunt et al., 2018) to score each rewrite candidate as a translation of source sentence. If no candidates have scores close to the tgt_0 , We return a null output. Otherwise we choose .

3.4 Experimental Results

We evaluate our system for rewriting quality on GATE in both masculine-to-feminine and feminine-to-masculine directions. To simulate runtime efficiency constraints, we impose a cutoff of 20 maximum source words. Any input sentence longer than this is treated as a null output and therefore a false negative.

Language	Direction	P	R	F0.5
Spanish	F→M	0.97	0.50	0.82
Spanish	M→F	0.95	0.40	0.74
French	F→M	0.97	0.28	0.65
French	M→F	0.91	0.27	0.61
Italian	F→M	0.96	0.47	0.79
Italian	M→F	0.91	0.32	0.67

Table 2: Our rewriter’s scores on GATE for each target language and rewrite direction

From these results we can see that our system performs best for Spanish in both directions, and in the female-to-male direction across all language pairs. Both trends can be explained to an extent by the properties of the translation models. High quality training data for English-Spanish is more plentiful than for the other two languages, leading to a higher quality model in general. As noted earlier, translation models have been shown to skew towards stereotypical gender assignments, which are more heavily weighted towards masculine forms. Therefore, it is not too surprising that when rewriting in this direction, the translation model is more likely to prefer an incorrect rewrite candidate.

3.5 End-to-End Evaluation

In our envisioned scenario, a gender rewriter would operate on the output of an MT system. It is unlikely, however, that direct MT output will consistently match GATE’s translations word-for-word. As a result, references cannot be directly utilized, and human annotation is required to assess the output of a rewriter alongside machine translation (MT) or any integrated system that generates a series of gender alternative translations from a single source sentence. One consideration is that a parallel sentence from GATE may no longer contain an AGME when machine translated, as the MT output may be unmarked for gender.

In order to test our combined system end-to-end, we sampled 200 source sentences from GATE and used our production MT models to translate

them into Spanish, and then pass that output to our rewriter. We then ask annotators to examine the source sentence and all translation outputs, and to provide the following annotations:

- If two translations are produced, mark true positive if the following are true (otherwise false positive):
 - Is the target gender-marked for an ambiguous source entity?
 - Were all the words marking gender of AGME changed correctly?
 - Were only the words marking gender of AGME changed?
- If only one translation is produced, is the target gender marked for an ambiguous source entity?
- If there are multiple AGMEs:
 - If two valid translations are produced mark as a true positive.
 - If only one translation is produced mark as a true negative.
 - Otherwise mark as a false positive.

We also retrieve translations for these sentences from an online English-Spanish translation system that can produce masculine and feminine alternative translations for this translation direction. We asked annotators to annotate these translations in the same manner.

Finally, we also asked annotators to mark source sentences for which the speaker is reasonably likely to know the referent’s gender, and therefore use of a masculine generic should be less likely (see 4.5). We evaluate quality on that subset as well for each system, in rows marked NG (*non-generic*). Results are presented in Table 3 and visualized in Figure 1.

	P	R	F0.5
Our System	0.97	0.41	0.76
Our System (NG)	1.00	0.50	0.84
Online system	0.96	0.14	0.45
Online system (NG)	1.00	0.21	0.56

Table 3: end-to-end scores for our system and an online translation system. NG rows are calculated only on *non-generic* sentences

Both systems heavily favor precision over recall, and recall is somewhat higher on the *non-generic*

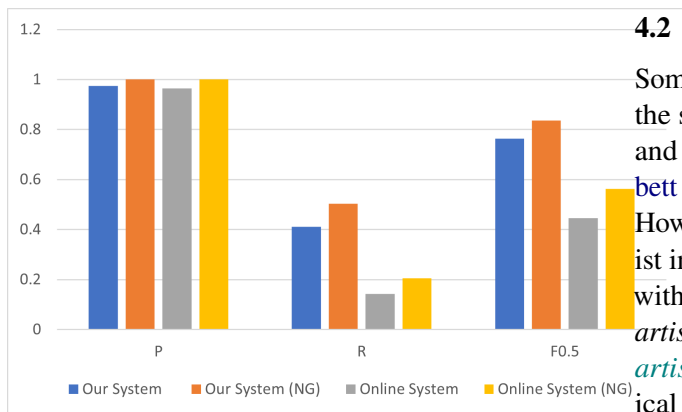


Figure 1: End-to-end scores for our system and an online translation system.

portion of the data. Overall, our system demonstrates significantly better coverage.

A full, end-to-end evaluation should include testing both sentences with and without AGMEs. As each instance in GATE involves at least one AGME, we suggest enhancing GATE with instances from Renduchintala and Williams (2021) and Vanmassenhove and Monti (2021), which feature unequivocally gendered source entities. In future work, we intend to develop a supplemental data set for GATE containing various types of negative examples: unambiguous source entities, entities that are unmarked in both source and target, and inanimate objects whose surface forms are distractors (e.g. depending on context, *player* and *cleaner* may refer to either objects or people).

4 Linguistic Background

4.1 Gender in Romance Languages

In Spanish, French and Italian, all nouns have a grammatical gender – either masculine or feminine. For inanimate objects, this gender is fixed and often arbitrary; for example, in French, *chaise* (chair) is feminine, while *canapé* (couch) is masculine. When a noun or pronoun refers to an animate entity, its grammatical gender will, with some notable exceptions, match the referential gender of that entity. (Vincent, 1988)

In these languages, referential gender of entities is frequently marked through morphology of an animate noun (e.g. *en-es*: *lawyer* \Rightarrow *abogada*(f), *abogado*(m)) or through agreement with gendered determiners, adjectives and verb forms.

4.2 Dual Gender and Epicene

Some animate nouns are *dual gender*, meaning that the same surface form is used for both masculine and feminine, such as French *artiste* (artist) (Corbett 1991 as cited in Hellinger and Bussmann 2002). However, other clues to the artist’s gender may exist in a French sentence through gender agreement with other associated words. For example, *The tall artist* could be translated into French as *La grande artiste*(f) or *Le grand artiste*(m). Here, grammatical gender of translations of *the* (*la* (f) / *le* (m)) and *tall* (*grande* (f), *grand* (m)) must match the referential gender of the referent noun.

Dual-gender determiners and adjectives exist as well, such as Spanish *mi* (my) and *importante* (important). So, for example, Spanish *mi huésped importante* (My important guest) has no gender marking. Similarly, in French and Italian, some determiners may contract before vowels to lose their gender marking. Feminine and masculine forms of *the* in French, *le* and *la*, both contract before vowels (and sometimes *h*) to become *l’*, so *l’artiste* (the artist) is not marked for gender.

While typically an entity’s referential gender will align with its grammatical gender, these languages each contain a handful of *epicene* nouns. These are nouns whose grammatical gender is fixed, regardless of the referential gender of the referent (Grevisse 2016 as cited in Hellinger and Bussmann 2003). Most notable among these is the direct translation of *person* into each of the target languages, which is always grammatically feminine: *La persona* (*es,it*) or *La personne* (*fr*). We also find some language-specific epicene nouns. For example, these Italian words are always grammatically feminine: *la guardia* (guard), *la vedetta* (sentry), *la sentinella* (sentry), *la recluta* (recruit), *la spia* (spy).⁶

4.3 Pronouns

Similarly to English, some pronouns in Romance languages are inherently gendered, while others are not. Entities referred to by gender-neutral pronouns, such as Spanish *yo* (I) and *tú* (you) commonly become gender-marked through predicative gender-inflecting adjectives. Further complicating these cases, subject pronouns are frequently omitted in Spanish and Italian (but notably not in

⁶Color-coding in this paragraph corresponds only to grammatical gender, while referential gender is ambiguous in these expressions.

French) as the subject can be inferred from verb morphology (Hellinger and Bussmann 2002, pp. 189, 252). This means that in some cases, the AGME in a sentence pair may be a zero-pronoun, such as English *I am tired* being translated to Spanish as *estoy cansada* (f) or *estoy cansado* (m). There is no overt subject in these translations corresponding to *I*, but the subject is implied by the verb form *estoy*.

4.4 Coreference

Another common pattern is that of coreferent mentions of a single entity, which must by definition have the same referential gender, and usually but not always the same grammatical gender. For example, in the following sentence, *friend* and *nurse* are the same individual and we would typically expect them to share the same referential gender in a direct translation into any of the target languages.

My best friend is a nurse

In cases where one coreferent mention is an epicene noun as described in 4.2, the grammatical genders of those mentions may in fact differ. In the following sentence, the described individual is unambiguously male. The phrase *una buena persona* (a good person) is grammatically female, while *un mal amigo* (a bad friend) and *él* (he) are grammatically male.⁷

He is a good person but a bad friend.
 ↓↓ es
Él es una buena persona, pero un mal amigo.

4.5 Masculine Generics

Traditionally, many languages, including Spanish, French and Italian, employ a paradigm known as masculine generics. Under this paradigm, feminine forms are considered to be explicitly gender-marked, while masculine forms should be used in situations where referential gender is unclear. Specifically, when referential gender is unknown by the speaker, or a mixed-gender group is known to contain at least one male individual, defaulting to grammatically masculine forms is generally considered correct in the language standard⁸. In this sense, masculine gender marking does not imply

⁷In this example, color-coding indicates grammatical gender of each mention as it appears the Spanish translation

⁸In recent years there is some explorations of using novel, gender-neutral forms in these contexts

the exclusion of female-identifying individuals, but a feminine gender marking would imply the exclusion of male-identifying individuals. (Hellinger and Bussmann 2002, 2003)

In most cases where a masculine generic might be used, we nonetheless ask our linguists to provide an alternative translation with feminine gender-marking. Language critics have noted that the use of masculine generics can evoke an association with 'male' (Hellinger and Bussmann 2003, pp. 101), and so we believe that inclusion of a feminine generic variant fits our mission of promoting inclusive language use. Our linguists were asked to annotate such generic mentions with the label INDF (*indefinite gender*), so that users who wish to follow a stricter interpretation can exclude these examples in their evaluations. However, upon analysis of our corpus we noted that this annotation was only consistently applied to the Italian data.

5 Related Work

A slew of challenge sets has been proposed for evaluating gender bias in Machine Translation.

MuST-SHE (Bentivogli et al. (2020) ; Savoldi et al. (2022) comprises approximately 1000 triplets consisting of audio, transcript, and reference translations for en-es, en-fr, and en-it languages. Each triplet is classified based on the gender of the speaker or explicit gender markers, such as pronouns, as either masculine or feminine. Furthermore, the dataset contains an alternative incorrect reference translation for every correct reference translation that alters the gender-marked words.

WinoMT Stanovsky et al. (2019) is a challenge set that comprises English sentences containing two animate nouns, one of which is coreferent with a gendered pronoun. Based on the context provided in the sentence, a human can easily identify which animate noun is coreferent and thus deduce the gender of the person described by that noun. By evaluating the frequency with which an MT system generates a translation with the correct gender for that animate noun, one can measure the extent to which the system depends on gender stereotypes rather than relevant context.

SimpleGEN Renduchintala et al. (2021) on the English-Spanish (en-es) and English-German (en-de) language pairs. It includes a test set consisting of short sentences with straightforward syntactic structures. Each source sentence includes an occupation noun and a clear indication of the gender of

the person described by that noun. In other words, the source sentence provides all the necessary information for a model to generate occupation nouns with the correct gender.

The Translated Wikipedia Biographies⁹ dataset comprises 138 documents containing human translations of Wikipedia biographies from English to Spanish and German. Each document comprises 8-15 sentences, providing a context for gender disambiguation evaluation across sentences.

MT-GenEval Currey et al. (2022) is a dataset that includes gender-balanced, counterfactual data in eight language pairs. The dataset ensures that the gender of individuals is unambiguous in the input segment, and it comprises multi-sentence segments that necessitate inter-sentential gender agreement.

Regarding the work on addressing ambiguously gendered inputs, Habash et al. (2019) tackle translation of ambiguous input by treating it as a gender classification and reinflection task when translating English into Arabic. Their approach focuses on the first-person singular cases. Given a gender-ambiguous source sentence and its translation, their system generates an alternative translation using the opposite gender. Additionally, they create a parallel corpus of first-person singular Arabic sentences that are annotated with gender information and reinflected accordingly. Alhafni et al. (2021) expand on the work of Habash et al. (2019) by adding second person targets to the Arabic Parallel Gender Corpus, as well as increasing the total number of sentences.

Google Translate announced¹⁰ an effort to address gender bias for ambiguously gendered inputs by showing both feminine and masculine translations. They support this feature for English to Spanish translation, as well as several gender-neutral languages into English.

Regarding debiasing in the monolingual context, (Zmigrod et al., 2019) propose a generative model capable of converting sentences inflected in masculine form to those inflected in feminine form, and vice versa, in four morphologically rich languages. Their work focuses on animate nouns.

In terms of rewriting text in English, Vanmassenhove et al. (2021) and Sun et al. (2021) propose rule-based and neural rewriting models, respec-

tively, that are capable of generating gender-neutral sentences.

6 Conclusion

We have presented GATE, a corpus of hand-curated test cases designed to challenge gender rewriters on a wide range of vocabulary, sentence structures and gender-related phenomena. Additionally, we provide an in-depth analysis of many of the nuances of grammatical gender in Romance languages and how it relates to translation. We also suggest metrics for gender rewriting and provide tools to aid with their calculation. Through this work we aim to improve the quality of MT output in cases of ambiguous source gender, as well as facilitate the development of better and more inclusive natural language processing (NLP) tools in general.

We look forward to future work in improving GATE and related projects. We aim to add additional languages pairs to GATE and investigate translation directions into English. We also hope to supplement with additional data, including negative examples. Finally, we plan to explore use of gender-neutral language use in various languages and how it can be incorporated into NLP applications.

7 Bias Statement

In this work, we propose a test set to evaluate translation of ambiguously gendered source sentences by NMT systems. Our work only deals with English as the source and is currently scoped to Romance languages as the target. To construct our test set, we have worked with bilingual linguists for each target language. We plan to increase scope of both source and target languages in future work.

Through this work, we hope to encourage and facilitate more inclusive use of natural language processing technology, particularly in terms of gender representation. In recent years, there is significant ongoing movement in the way gender manifests in languages use. One form that this takes is in new gender-neutral language constructs in Romance languages such as French, Spanish and Italian to accommodate gender underspecificity and non-binary gender identities. We support the development of this more representative and inclusive language, and endeavor to find ways to support it through technology. In this work, however, for the sake of simplicity we restrict our scope to language as used to express gender along more conventionally binary

⁹<https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html>

¹⁰<https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html>

lines, and we therefore do not consider non-binary language or word forms. We are working with both language experts and non-binary-identifying individuals to expand the scope to include non-binary and gender-underspecified language in future work.

References

- Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2021. [The arabic parallel gender corpus 2.0: Extensions and analyses](#).
- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Matia 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.
- Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. [Examining covert gender bias: A case study in turkish and english machine translation models](#). *CoRR*, abs/2108.10379.
- Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. [MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Östen Dahl. 2000. [Animacy and the notion of semantic gender](#), pages 99–116. De Gruyter Mouton, Berlin, New York.
- Maurice Grevisse. 2016. *Le bon usage : Grevisse langue française*, 16e édition. edition. De Boeck Supérieur, Louvain-La-Neuve.
- 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.
- Marlis Hellinger and Hadumod Bussmann. 2002. *Gender Across Languages*, volume 2. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Marlis Hellinger and Hadumod Bussmann. 2003. *Gender Across Languages*, volume 3. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Kenton Lee, Luheng He, and L. Zettlemoyer. 2018. [Higher-order coreference resolution with coarse-to-fine inference](#). In *NAACL-HLT*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. [Assessing gender bias in machine translation - A case study with google translate](#). *CoRR*, abs/1809.02208.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- 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 Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.
- Adithya Renduchintala and Adina Williams. 2021. [Investigating failures of automatic translation in the case of unambiguous gender](#). *CoRR*, abs/2104.07838.
- 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, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. [They, them, theirs: Rewriting with gender-neutral english.](#)

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

Nigel Vincent. 1988. *The Romance Languages*. Croom Helm, London.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. [Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Category Labels

There are a wide range of linguistic phenomena that can interact with gender in translation. We have devised several category labels that can be applied to examples. In order to promote diversity within the corpus, linguists were asked to ensure that a certain minimum number of examples are included for each such label. This also has the benefit of helping pinpoint weaknesses in an evaluated system. For example, a rewriting system may perform well when the ambiguous noun is the subject of the sentence, but do poorly when it is a direct object. We hope to include per-category evaluation and analysis of our system in a future version of this work.

Unless otherwise stated, category labels are determined based on the target sentence set rather than the source sentence, as this is generally the more important input to the rewriter. A single example will typically have multiple labels.

- **Grammatical Role categories:** An AGME is a subject (SUBJ), direct object (DOBJ) indirect object (IOBJ), subject complement (SCMP), or object of a preposition (OPRP, excluding indirect objects)
- **Animate Noun categories:** profession (PROF, e.g. doctor), Religion (REL, e.g.

Bhuddist), Nationality (NAT, e.g. Italian), Family and other relationships (REL, e.g. neighbor), Non-Human (NHUM, e.g. cat, vampire), Other (OTH, e.g. winner, accused)

- **Adjectives and past participles:** attributive (AATR), predicative (APRD), past-participle form as an adjective (PPA), past-participle form not as an adjective (PPNA), Adjective modifies non-ambiguous noun (ANAN). Most of these distinctions are included to test a rewriter’s ability to distinguish between adjective surface forms that should be modified along with key nouns and those that should not.
- **Sentence Types categories:** Headline (HEAD), sentence fragment (FRAG), question (QUES), imperative (IMPR), Ambiguous noun in a subordinate clause (SUBC)
- **Other categories:** Plural ambiguous noun (PLUR), indefinite i.e. does not refer to an entity concretely known by the speaker, e.g. "Where can I find a good doctor?" (INDF), Requires agreement across different clauses with noun that was ambiguous in source (DFCL), Distinct animate nouns behave as a single group and are *gender-linked* (GLNK)

Label	es	fr	it	description
Semantic Type				
PROF	1168	490	1208	Profession word
NAT	118	249	157	Nationality or locality membership
REL	25	150	29	Religious affiliation
FAM	327	250	192	Family or other relationship
NHUM	2	40	–	Non-Human
OTH	580	941	708	Other
Grammatical role				
SUBJ	1638	1221	1573	Subject
SCMP	118	185	121	Subject complement
DOBJ	181	328	399	Direct object
IOBJ	136	275	165	Indirect object
OPRP	250	279	518	Object of preposition
VOC	3	–	4	Vocative
POSC	80	–	289	Possessive complement
Sentence Type				
QUES	124	–	–	Question
FRAG	49	101	–	Sentence Fragment
IMPR	14	135	–	Imperative
Adjective-related				
APRD	82	359	213	Predicative adjective agreeing with AGME
AATR	293	190	315	Attributive adjective agreeing with AGME
ANAN	97	1026	–	Adjective modifying a word other than AGME
PPA	361	172	290	Adjective has same surface form as a past participle
APPS	–	35	22	post-positive adjective – remove??
Pronoun subtype				
PERS	–	219	146	Personal pronoun
RELA	–	15	13	Relative pronoun
DEMO	–	64	28	Demonstrative pronoun
POSS	80	–	–	Possesive pronoun
DROP	157	–	–	AGME is a dropped/zero pronoun
IPRO	–	369	53	Indefinite pronoun
Other				
PLUR	991	1110	1042	Plural
INDF	–	–	229	Indefinite/masculine generic could apply
DFCL	136	113	–	Changed words in alternatives cross clause boundaries
GLNK	–	94	–	"gender-link" – AGMEs are not coreferent but conceptually linked, different genders would be unnatural

Table 4: Counts of sentences with each category label per language. '–' indicates that this language was not annotated for this label