Automatically Identifying Gender Issues in Machine Translation using Perturbations

Hila Gonen
Bar-Ilan University
hilagnn@gmail.com

Kellie Webster
Google Research NYC
websterk@google.com

Abstract

The successful application of neural methods to machine translation has realized huge quality advances for the community. With these improvements, many have noted outstanding challenges, including the modeling and treatment of gendered language. While previous studies have identified issues using synthetic examples, we develop a novel technique to mine examples from real world data to explore challenges for deployed systems. We use our method to compile an evaluation benchmark spanning examples for four languages from three language families, which we publicly release to facilitate research. The examples in our benchmark expose where model representations are gendered, and the unintended consequences these gendered representations can have in downstream application.

1 Introduction

Machine translation (MT) has realized huge improvements in quality from the successful application and development of neural methods (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Vaswani et al., 2017; Johnson et al., 2017; Chen et al., 2018). As the community has explored this enhanced performance, many have noted the outstanding challenge of modeling and handling gendered language (Kuczmarski, 2018; Escudé Font and Costa-jussà, 2019). We extend this line of work, which identifies issues using synthetic examples manually curated for a target language (Stanovsky et al., 2019; Cho et al., 2019), by analyzing real world text across a range of languages to understand challenges for deployed systems.

In this paper, we explore the class of issues which surface when a neutral reference to a person is translated to a gendered form (e.g. in Table 1, where the English *counselor* and *nurse* are translated into the French *conseiller* (masculine) and *infirmière* (feminine). For this class of examples, the MT task requires a system to produce a single translation without source cues, thus exposing a model's preferred gender for the reference form.

With this scope, we make two key contributions. First, we design and implement an automatic pipeline for detecting examples of our class of gender issues in real world input, using a BERT-based perturbation method novel to this work. A key advantage of our pipeline beyond previous work is its extensibility: a) beyond word lists; b) to different language pairs and c) parts of speech. Second, using our new pipeline, we compile a dataset that we make publicly available to serve as a benchmark for future work. We focus on English as the source language, and explore four target gendered languages across three language families (French, German, Spanish, and Russian). Our examples expose where MT encodings are gendered, finding new issues not covered in previous manual approaches, and the unintended consequences of this for translation.

2 Gender Marking Languages

Gender-marking languages have rich grammatical systems for expressing gender (Corbett, 1991). To produce a valid sentence in a gender-marking language, gender may need to be marked not only on pronouns (*he*, *she*), as it is in English, but also nouns and even verbs, as well as words linked to these gendered nouns and verbs. This means that translating from a language like English, with little gender marking, to a gender-marking language like Spanish, requires a system to produce gender markings that may not have explicit evidence in the source. For instance, *The tall teacher* from English could be translated into the Spanish ***La maestra alta*** (feminine) or ***El maestro alto*** (masculine).

Source Sentence (En)	Translation (Fr)	M/F
so is that going to affect my chances of becoming a **counselor**?	Alors, est - ce que cela va affecter mes chances de devenir *conseiller*?	M
so is that going to affect my chances of becoming a **nurse**?	Alors, est - ce que cela va affecter mes chances de devenir infirmière?	F

Table 1: An example from our dataset of a minimal pair of English gender-neutral source sentences, translated into two different genders in French. Red (italic) stands for masculine, cyan (normal) stands for feminine.

3   Automatic Detection of Gender Issues

The class of issues we are interested in are those where translation to a gender-marking language exposes a model's gender preference for a personal reference. The examples we find that demonstrate this are English sentence-pairs, a minimal pair differing by only a single word, e.g. *doctor* being replaced by *nurse*. In each of our examples, this minimal perturbation does not change the gender of the source but gives rise to gender differences upon translation, e.g. *doctor* becoming masculine and *nurse* feminine.

In this section, we present a simple, extensible method to mine such examples from real-world text. Our method does not require expansive manually-curated word lists for each target language, which enables us to discover new kinds of entities that are susceptible to model bias but are not usually thought of this way. Indeed, while we demonstrate its utility with nouns with four target languages, our method is naturally extensible to new language pairs and parts of speech with no change in design.

Filtering source sentences   Our first step is to identify sentences that are gender neutral and that include a single human entity, e.g. *A doctor works in a hospital*. We focus on human entities since these have been the target of previous studies and present the largest risk of gender issues in translation.

We use a BERT-based Named Entity Recognition (NER) model that identifies human entities, and exclude sentences that have more than one token tagged as such. We also remove sentences in which the entity is a gendered term in English[1] (e.g. *mother, nephew*), a name, or not a noun.

Note that all the sentences we get are naturally occurring sentences, and that we do not use any templates or predefined lists of target words that we want to handle.

[1] https://github.com/tolga-b/debiaswe/blob/master/data/gender_specific_full.json https://github.com/uclanlp/gn_glove/tree/master/wordlist

Perturbations using BERT   We use BERT as a masked language model to find words which can substitute for the human entity identified in the previous filtering step, e.g. *doctor* → *nurse*. We aim to get natural-sounding output and maintain extensibility, and thus do not use predefined substitutions. We cap our search to the first 100 candidates BERT returns, accepting the first 10 which are tagged as person, and for which the resulting sentences also pass the filtering step.

Translation   We translate each of the generated sentences into our target languages using Google Translate[2]. *A doctor/nurse works in a hospital* → *Un doctor/Una enfermera trabaja en un hospital*.

Alignment   We align tokens in the original and translated sentences using fast-align (Dyer et al., 2010). This is needed in order to know which token in the translation output is the focus entity in the source sentence, whose gender we want to analyze.

Gender Identification   We use a morphological analyzer, implemented following Kong et al. (2017), to tag the gender of the target word.

Identifying Examples   The final step of our pipeline is identifying pairs of sentences to include in our dataset, pairs where different genders are assigned to the human entity. Our example would be included since *doctor* is translated with the masculine form *Un doctor* while *nurse* is translated with the feminine form *Una enfermera*.

4   Challenge Dataset

We compile our final dataset from the output of this pipeline, and explore its properties to understand the issues it represents for deployed systems.

4.1   Random Sampling

In our final dataset, we include both examples that passed the final example identification step above (pairs referred to as "at risk"), as well as a random

[2] https://translate.google.com/

selection that did not ("not at risk"). We do this in order to not be constrained too heavily by our choice of translation model; if we did not, we would have no chance of inspecting examples that our system did not spot as at risk but other models might have.

4.2 Fixed Grammatical Gender Rating

When we inspected the examples identified as at risk by our pipeline, the major source of error we found pertained to the issue of fixed grammatical gender. Consider the example in Figure 1:

Sentence 1:
  En: you don't have to be the **victim** in whatever.
  Fr: vous ne devez pas être la victime de quoi que ce soit.
Sentence 2:
  En: you don't have to be the **expert** in whatever.
  Fr: vous ne devez pas être *l'expert* en quoi que ce soit.

Figure 1: An example from our dataset, with fixed grammatical gender. Red (italic) stands for masculine, cyan (normal) stands for feminine.

In this example, the word *victim* in the first English sentence is identified by our tagger as a human entity. However, its French translation *victime* is feminine by definition, and cannot be assigned another gender regardless of the context, causing a false positive result.

We attempted to filter these examples automatically but came across a number of challenges. Most critically, we found no high-quality, comprehensive dictionary that included the required information for all languages, and heuristics we applied were noisy and not reliable.[3] We observed that the underlying reason for these challenges was that there is no closed list of grammatically-fixed words as languages are evolving to be more gender-inclusive. In order to maximize and guarantee data quality, and to be sensitive to the nuances of language change, we decided to add a manual filtering step after our pipeline to select the positive (at risk) examples.

We note that the problem of fixed grammatical gender is particular to nouns. Our pipeline is naturally extensible across parts of speech and we would not expect the same issues in future work perturbing adjectives or verbs.

4.3 Dataset Statistics

To create our dataset we mine text from the subreddit "career".[4] From 29,330 sentences, we found 4,016 which referred to a single, non-gendered human entity. Introducing perturbations with BERT into these 4,016 sentences yielded 40,160 pairs. Out of those, 592 to 1,012 pairs are identified as at risk by our pipeline, depending on the target language. We asked humans to manually identify 100 true at risk examples for the final dataset, which was achieved for all languages except Russian, where we have 59 pairs.[5] To this 100, we add a further 100 randomly sampled negative examples for each language. Table 2 shows a representative example for each language-pair.

4.4 Exploratory Analysis

Table 3 lists the most frequent focus personal references in each language-pair among the positive (at risk) and negative (not at risk) examples, along with the ratio between times the reference form was translated as masculine compared to feminine. Words with extreme values of this ratio indicate cases where a model has a systematic preference for one gender over another, i.e. a gendered representation.

Among the negative examples, we see a prior for masculine translations across all terms. Positive examples break from this prior by exposing reference forms with a feminine preference: *nurse* and *secretary* are the most consistently feminine forms, consistent with the Bureau of Labor statistics used in previous work (Caliskan et al., 2017).

Figure 2 shows two sentence pairs that appear as positive examples across all four language-pairs. Two of the three forms, *nurse* and *mechanic*, are consistent with the gender statistics of Caliskan et al.; the association of *fighter* with the masculine gender is a new discovery of our method.

5 Related Work

Our study builds on the literature around gender bias in machine translation. Cho et al. (2019) use sentence templates to probe for differences in Korean pronouns. Prates et al. (2019) and Stanovsky et al. (2019) also use sentence templates, but filled with word lists, of professions and adjectives in the

---

[3]We tried both using a morphological lexicon and a predefined word list in English. Both methods performed poorly, filtering too many or too few sentence pairs, respectively.

[5]Out of the pairs identified using our pipeline, between 1/3 - 1/10 were selected by the annotators, depending on the language.

	Source Sentence (En)	Translation	M/F
Fr	also should i ask the **manager** what the pay would be if i got the job prior to flying out?	De plus, devrais - je demander au *gestionnaire* quel serait le salaire si je obtenais le poste avant de prendre l'avion?	M
	also should i ask the **secretary** what the pay would be if i got the job prior to flying out?	De plus, devrais - je demander à la secrétaire quel serait le salaire si je obtenais le poste avant mon départ?	F
De	currently thinking about learning a trade (mostly a **electrician**).	Derzeit über das Erlernen eines Gewerbes nachdenken (meistens *Elektriker*).	M
	currently thinking about learning a trade (mostly a **cook**).	Derzeit über das Erlernen eines Gewerbes nachdenken (meistens eine Köchin).	F
Es	- decided to become a **teacher**: spent a year working 2 jobs and doing prerequisites for a masters in education.	- Decidí ser maestra: pasé un año trabajando en 2 trabajos y haciendo requisitos previos para una maestría en educación.	F
	- decided to become a **lecturer** : spent a year working 2 jobs and doing prerequisites for a masters in education.	- Decidí ser *profesor*: pasé un año trabajando en 2 trabajos y haciendo requisitos previos para una maestría en educación.	M
Ru	i read about a **psychologist** who upgraded into becoming a m.d.	Я читал о *психологе*, который превратился в Md.	M
	i read about a **nurse** who upgraded into becoming a m.d.	Я читал о медсестре, которая превратилась в доктора медицины.	F

Table 2: Examples from our dataset of a minimal pair of English gender-neutral source sentences, translated into two different genders in all target languages. Red (italic) stands for masculine, cyan (normal) stands for feminine.

	Positive	M:F	Negative	M:F
Fr	nurse	0:36	manager	685:1
	secretary	0:17	employee	406:0
	teacher	7:1	employees	364:0
	assistant	1:7	parents	353:0
	manager	8:0	teacher	337:0
De	secretary	0:27	manager	594:0
	nurse	0:21	employees	409:1
	teacher	3:7	friends	359:0
	receptionist	0:9	employee	320:0
	manager	7:0	students	316:0
Es	teacher	4:29	manager	691:0
	nurse	0:31	employee	446:0
	secretary	0:26	friends	380:0
	writer	8:0	parents	374:0
	employee	5:0	supervisor	345:0
Ru	nurse	0:32	manager	713:0
	babysitter	0:13	employees	519:0
	nurses	0:5	friends	439:0
	dishwasher	0:4	students	417:0
	technician	3:0	employee	392:0

Table 3: Top five human reference forms in our dataset, and their ratio of times they are translated as masculine compared to feminine. Positive indicates that the examples were taken from the at-risk group from our pipeline, and negative from the random sample among the not at-risk group.

Sentence pair 1:
 Original: you need to have experience working with hydraulic lifts, & they like to see that you've worked or trained as a **mechanic**.
 Substitution: you need to have experience working with hydraulic lifts, & they like to see that you've worked or trained as a **nurse**.

Sentence pair 2:
 Original: in fact, probably not even as a seasoned **nurse**.
 Substitution: in fact, probably not even as a seasoned **fighter**.

Figure 2: Two sentence pairs from our dataset that found to be shared between all four target languages.

former, and professions in the latter. A separate but related line of work focuses on generating correct inflections when translating to gender-marking languages (Vanmassenhove et al., 2018; Moryossef et al., 2019).

6 Conclusion

The primary contribution of our work is a novel, automatic method for identifying gender issues in machine translation. By performing BERT-based perturbations on naturally-occurring sentences, we are able to identify sentence pairs that behave differently upon translation to gender-marking languages. We demonstrate our technique over human reference forms and discover new sources of risk beyond the word lists used previously. Furthermore, the novelty of our approach is its

natural extensibility to new language pairs, text genres, and different parts of speech. We look forward to future work exploring such applications.

Using our new method, we compile a dataset across four languages from three language families. By publicly releasing our dataset, we hope to enable the community to work together towards solutions that are inclusive and equitable to all.

Acknowledgements

References

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

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Greville G Corbett. 1991. Gender.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*.

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*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Lingpeng Kong, Chris Alberti, Daniel Andor, Ivan Bogatyy, and David Weiss. 2017. Dragnn: A transition-based framework for dynamically connected neural networks. *arXiv preprint arXiv:1703.04474*.

James Kuczmarski. 2018. Reducing gender bias in google translate.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*.

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*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.