



# GENDER BIAS IN MACHINE TRANSLATION

Pedro H.C. Avelar

Luis C. Lamb

October 2019

This presentation is based on our work “Assessing Gender Bias in Machine Translation – A Case Study with Google Translate”, (PRATES; AVELAR; LAMB, 2019).

(BOLUKBASI et al., 2016) identified Biases in Word Embeddings and argued that debiasing was necessary before applying these methods in real world applications.

(BOLUKBASI et al., 2016) identified Biases in Word Embeddings and argued that debiasing was necessary before applying these methods in real world applications.

*There have been hundreds of papers written about word embeddings and their applications (...). However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.*

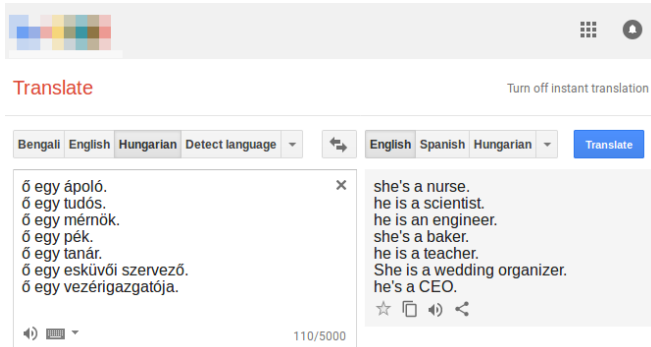
*(...)*

(BOLUKBASI et al., 2016) identified Biases in Word Embeddings and argued that debiasing was necessary before applying these methods in real world applications.

*There have been hundreds of papers written about word embeddings and their applications (...). However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.*

*(...)*

*One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings. However, by reducing the bias in today's computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society. At the very least, machine learning should not be used to inadvertently amplify these biases, as we have seen can naturally happen*



The screenshot shows the Google Translate interface. At the top, there is a header with a colorful grid icon and a notification bell. Below the header, the word "Translate" is displayed in red, followed by a link to "Turn off instant translation". The main interface features a language selection bar with "Bengali", "English", "Hungarian", and "Detect language" options, along with a bidirectional arrow icon. To the right, there are buttons for "English", "Spanish", and "Hungarian", and a blue "Translate" button. The input text on the left is in Hungarian, and the output on the right is in English. The output text shows a clear gender bias, where all professions are translated as male.

Translate [Turn off instant translation](#)

Bengali English Hungarian Detect language ↕ English Spanish Hungarian Translate

ő egy ápoló.  
ő egy tudós.  
ő egy mérnök.  
ő egy pék.  
ő egy tanár.  
ő egy esküvői szervező.  
ő egy vezérigazgatója.

she's a nurse.  
he is a scientist.  
he is an engineer.  
she's a baker.  
he is a teacher.  
She is a wedding organizer.  
he's a CEO.

110/5000

Figure: Example translations which were trending in social media

- There was a social media uproar on MT gender bias for professions in the translation from languages with gender neutral
- Expand on this, providing a transparent way of assessing gender bias in MT systems

- There was a social media uproar on MT gender bias for professions in the translation from languages with gender neutral
- Expand on this, providing a transparent way of assessing gender bias in MT systems
- Provide a case-study with a widely used system and compare it with real-world distributions of gender



- There was a social media uproar on MT gender bias for professions in the translation from languages with gender neutral
- Expand on this, providing a transparent way of assessing gender bias in MT systems
- Provide a case-study with a widely used system and compare it with real-world distributions of gender
- Extra: Provide a similar study for adjectives.

- Languages

- Languages
  - With Gender Neutral Pronouns and supported by GT

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated
  - Most occupations had data on gender distribution



- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated
  - Most occupations had data on gender distribution
  - Missing data imputed as category aggregate

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated
  - Most occupations had data on gender distribution
  - Missing data imputed as category aggregate
- Adjectives

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated
  - Most occupations had data on gender distribution
  - Missing data imputed as category aggregate
- Adjectives
  - Extracted from CoCA <<https://corpus.byu.edu/coca/>>

- Languages
  - With Gender Neutral Pronouns and supported by GT
  - Didn't include some (Nepali, Korean and Persian) due to difficulties in providing template/processing the data
- Labour
  - Extracted from the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2017)
  - Manually curated
  - Most occupations had data on gender distribution
  - Missing data imputed as category aggregate
- Adjectives
  - Extracted from CoCA <<https://corpus.byu.edu/coca/>>
  - Manually Curated from the top 1000 most frequent adjectives

# RESULTS – OCCUPATION CATEGORY

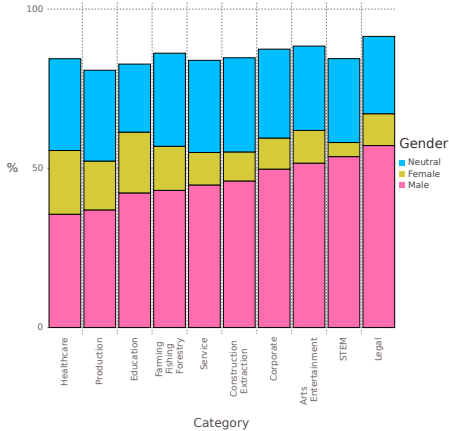


Figure: Plot showing how different Occupation Categories have different distributions of translation pronouns

# RESULTS – LANGUAGE

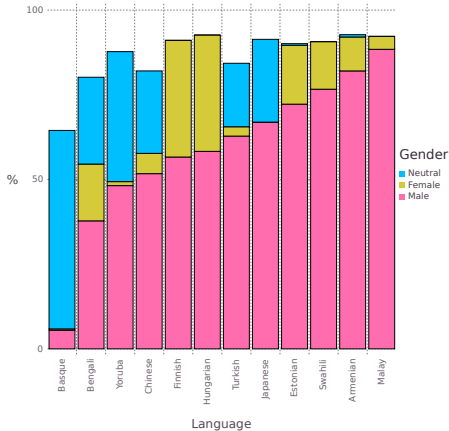


Figure: Plot showing how different Languages have different distributions of translation pronouns

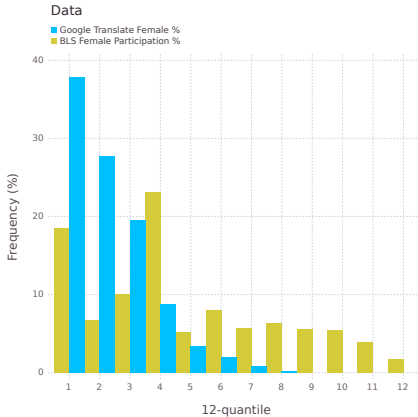


Figure: Plot showing severe underestimation of female participation

# RESULTS – ADJECTIVES

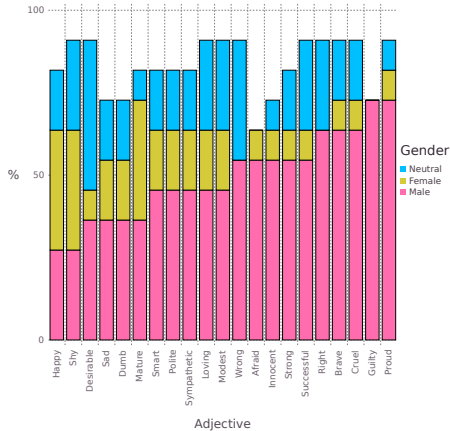


Figure: Most adjectives seem to adopt male defaults, but some specific words show certain trends, as “Guilty”, while some adjectives such as shy and happy seem to skew less towards male translations.



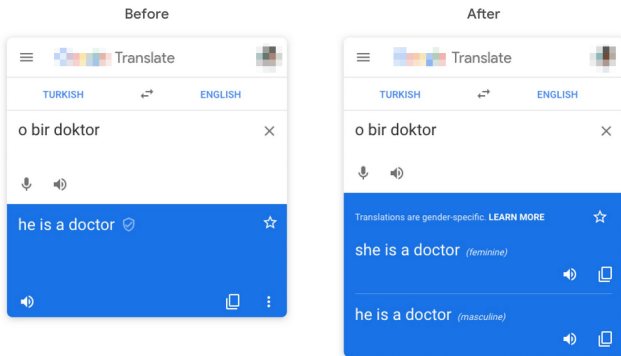


Figure: GT provided translation alternatives shortly after our paper

- None of us were speakers of the gender-neutral languages

- None of us were speakers of the gender-neutral languages
- GT doesn't provide confidence scores for words in the API

- None of us were speakers of the gender-neutral languages
- GT doesn't provide confidence scores for words in the API
- Our work was limited to a single template translation per word (except for Bengali)

- None of us were speakers of the gender-neutral languages
- GT doesn't provide confidence scores for words in the API
- Our work was limited to a single template translation per word (except for Bengali)
- The occupation list is from a single source (BLS)

- None of us were speakers of the gender-neutral languages
- GT doesn't provide confidence scores for words in the API
- Our work was limited to a single template translation per word (except for Bengali)
- The occupation list is from a single source (BLS)
- Occupations were forward translated to be back-translated again

- MT tools could provide alternative translations

- MT tools could provide alternative translations (GT has been updated to include this)
- MT tools could provide confidence scores for individual words



- MT tools could provide alternative translations (GT has been updated to include this)
- MT tools could provide confidence scores for individual words
- Automatic evaluation can help detect bias in a system and call for further action
- Datasets could have a curated subset to enforce parity

- MT tools could provide alternative translations (GT has been updated to include this)
- MT tools could provide confidence scores for individual words
- Automatic evaluation can help detect bias in a system and call for further action
- Datasets could have a curated subset to enforce parity

- We have done another work metrifying ethics engagement in AI research

- We have done another work metrifying ethics engagement in AI research
- Searched for ethics related keywords in flagship conference abstracts and titles

- We have done another work metrifying ethics engagement in AI research
- Searched for ethics related keywords in flagship conference abstracts and titles
- Although ethics is being more and more commonly discussed in workshops,

- We have done another work metrifying ethics engagement in AI research
- Searched for ethics related keywords in flagship conference abstracts and titles
- Although ethics is being more and more commonly discussed in workshops, It is far from being in the main tracks

- We have done another work metrifying ethics engagement in AI research
- Searched for ethics related keywords in flagship conference abstracts and titles
- Although ethics is being more and more commonly discussed in workshops, It is far from being in the main tracks
- For more, read our paper (PRATES; AVELAR; LAMB, 2018)

- Work already done



- Work already done
  - (CHO et al., 2019) performed a similar evaluation for Korean on three different translation tools, using multiple sentence templates.

- Work already done
  - (CHO et al., 2019) performed a similar evaluation for Korean on three different translation tools, using multiple sentence templates.
  - (STANOVSKY; SMITH; ZETTLEMOYER, 2019) evaluated gender bias for 8 languages and 6 MT systems for correct translation alignments

- Work already done
  - (CHO et al., 2019) performed a similar evaluation for Korean on three different translation tools, using multiple sentence templates.
  - (STANOVSKY; SMITH; ZETTLEMOYER, 2019) evaluated gender bias for 8 languages and 6 MT systems for correct translation alignments
  - (KUCZMARSKI; JOHNSON, 2018) proposed techniques to produce both translations in all genders in the target language.

- Work already done
  - (CHO et al., 2019) performed a similar evaluation for Korean on three different translation tools, using multiple sentence templates.
  - (STANOVSKY; SMITH; ZETTLEMOYER, 2019) evaluated gender bias for 8 languages and 6 MT systems for correct translation alignments
  - (KUCZMARSKI; JOHNSON, 2018) proposed techniques to produce both translations in all genders in the target language.
  - (ZHAO et al., 2018; RUDINGER et al., 2018; WEBSTER et al., 2018) provided corpora for pronoun resolution and assessing gender bias

- **Korean speakers**

- **Korean speakers**
- Provided a way to test MT systems for the Korean language

- **Korean speakers**
- Provided a way to test MT systems for the Korean language
- Tested on **3 different MT systems**

- **Korean speakers**
- Provided a way to test MT systems for the Korean language
- Tested on **3 different MT systems**
- Used **multiple sentence templates per pair**



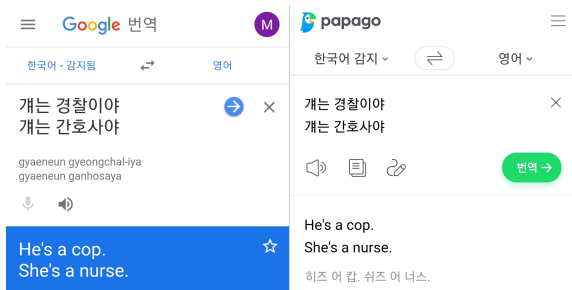


Figure: Cho et al. tested on different systems, including GT and Naver Papago (NP). Reuse of this image was kindly permitted by Cho et al.

- Based their studies in previous studies regarding Gender bias in coreference resolution (ZHAO et al., 2018; RUDINGER et al., 2018)

- Based their studies in previous studies regarding Gender bias in coreference resolution (ZHAO et al., 2018; RUDINGER et al., 2018)
- Tested on **6 different MT systems**, 4 commercial ones

- Based their studies in previous studies regarding Gender bias in coreference resolution (ZHAO et al., 2018; RUDINGER et al., 2018)
- Tested on **6 different MT systems**, 4 commercial ones
- Tested sentences based on automatic tools and checking for gender alignment between the source and target sentences

- Based their studies in previous studies regarding Gender bias in coreference resolution (ZHAO et al., 2018; RUDINGER et al., 2018)
- Tested on **6 different MT systems**, 4 commercial ones
- Tested sentences based on automatic tools and checking for gender alignment between the source and target sentences
- Also performed manual annotation for a small subset of 100 sentences with 2 **native annotators**

- Proposed techniques to produce both translations in all genders in the target language.

- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:

- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:
  - Identify if a translation query may need gendered translation



- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:
  - Identify if a translation query may need gendered translation
  - If so, translate the sentence forcing all possible genders in the target language

- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:
  - Identify if a translation query may need gendered translation
  - If so, translate the sentence forcing all possible genders in the target language
  - Post-process to see if produced sentences are appropriate

- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:
  - Identify if a translation query may need gendered translation
  - If so, translate the sentence forcing all possible genders in the target language
  - Post-process to see if produced sentences are appropriate
  - Present gendered tuple to user if so, otherwise translate as normal

- Proposed techniques to produce both translations in all genders in the target language.
- In Summary:
  - Identify if a translation query may need gendered translation
  - If so, translate the sentence forcing all possible genders in the target language
  - Post-process to see if produced sentences are appropriate
  - Present gendered tuple to user if so, otherwise translate as normal
- Similar to what GT seems to have adopted.

- (ZHAO et al., 2018; RUDINGER et al., 2018; WEBSTER et al., 2018) provided corpora for gendered pronoun resolution

- (ZHAO et al., 2018; RUDINGER et al., 2018; WEBSTER et al., 2018) provided corpora for gendered pronoun resolution
- Can be used to benchmark MT tools

- (ZHAO et al., 2018; RUDINGER et al., 2018; WEBSTER et al., 2018) provided corpora for gendered pronoun resolution
- Can be used to benchmark MT tools
- Also identified and called our biases

- Future Work



- Future Work
  - We are not aware of a study similar to (CHO et al., 2019) for the Persian or Nepali languages

- Future Work
  - We are not aware of a study similar to (CHO et al., 2019) for the Persian or Nepali languages
  - Cho et al. are looking to expand their work to multiple languages


- Future Work
  - We are not aware of a study similar to (CHO et al., 2019) for the Persian or Nepali languages
  - Cho et al. are looking to expand their work to multiple languages
  - We are expanding some of our experiments on bias in MT


- Future Work
  - We are not aware of a study similar to (CHO et al., 2019) for the Persian or Nepali languages
  - Cho et al. are looking to expand their work to multiple languages
  - We are expanding some of our experiments on bias in MT
  - Both of us are open to collaborations and suggestions


THANK YOU  
THANK YOU


GENDER BIAS IN MACHINE  
TRANSLATION


Thank You!


 BOLUKBASI, T. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: **NIPS**. [S.l.: s.n.], 2016. p. 4349–4357.


 Bureau of Labor Statistics. "**Table 11: Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity, 2017**". [S.l.], 2017.


 CHO, W. I. et al. On measuring gender bias in translation of gender-neutral pronouns. In: **Proceedings of the First Workshop on Gender Bias in Natural Language Processing**. Florence, Italy: Association for Computational Linguistics, 2019. p. 173–181. Disponível em: <<https://www.aclweb.org/anthology/W19-3824>>.


 KUCZMARSKI, J.; JOHNSON, M. Gender-aware natural language translation. 2018.


 PRATES, M. O. R.; AVELAR, P. H.; LAMB, L. C. Assessing gender bias in machine translation: a case study with google translate. **Neural Computing and Applications**, Mar 2019. ISSN 1433-3058. Disponível em: <<https://doi.org/10.1007/s00521-019-04144-6>>.

 PRATES, M. O. R.; AVELAR, P. H. C.; LAMB, L. C. On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals. In: **GCAI**. [S.l.]: EasyChair, 2018. (EPiC Series in Computing, v. 55), p. 188–201.

 RUDINGER, R. et al. Gender bias in coreference resolution. In: **NAACL-HLT (2)**. [S.l.]: Association for Computational Linguistics, 2018. p. 8–14.

 STANOVSKY, G.; SMITH, N. A.; ZETTLEMOYER, L. Evaluating gender bias in machine translation. In: **ACL (1)**. [S.l.]: Association for Computational Linguistics, 2019. p. 1679–1684.

 WEBSTER, K. et al. Mind the gap: A balanced corpus of gendered ambiguous pronouns. In: **Transactions of the ACL**. [S.l.: s.n.], 2018. p. to appear.

 ZHAO, J. et al. Gender bias in coreference resolution: Evaluation and debiasing methods. In: **NAACL-HLT (2)**. [S.l.]: Association for Computational Linguistics, 2018. p. 15–20.