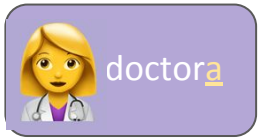
Evaluating Gender Bias in Machine Translation

Gabriel Stanovsky, Noah Smith and Luke Zettlemoyer

ACL 2019

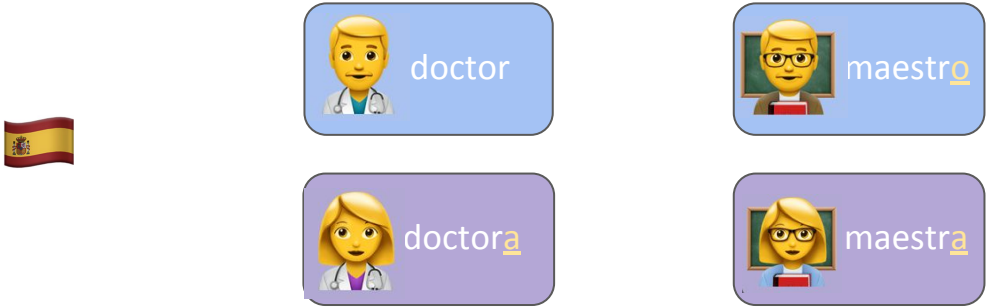
Grammatical Gender

- Some languages encode *grammatical gender* (Spanish, Italian, Russian, …)

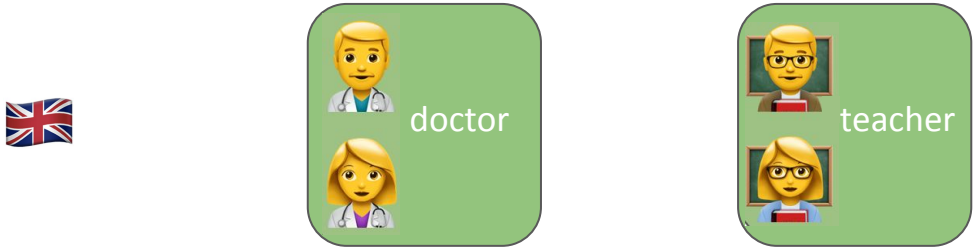


Grammatical Gender

- Some languages encode *grammatical gender* (Spanish, Italian, Russian, …)



- Other languages *do not* (English, Turkish, Basque, Finnish, …)



Translating Gender

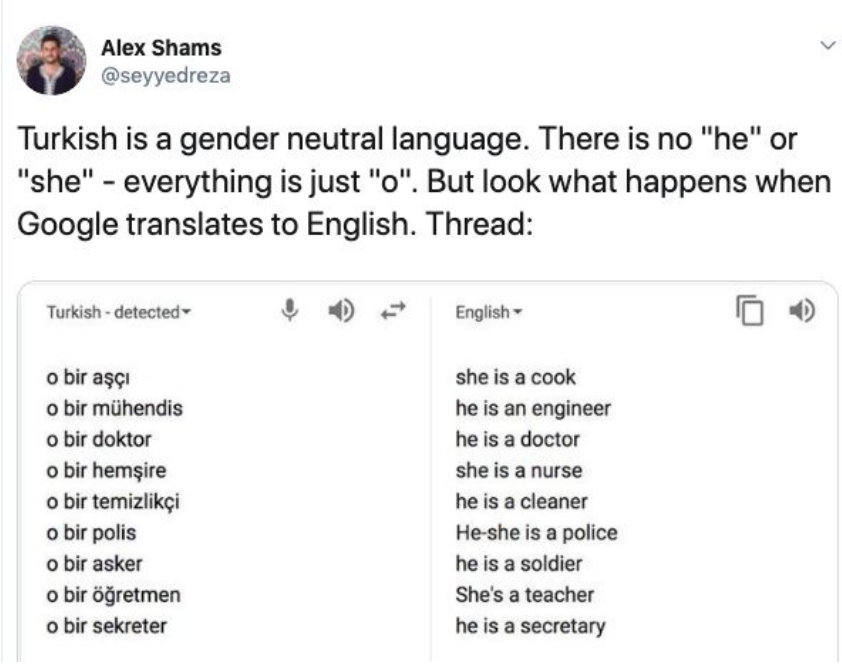
- Variations in gender mechanisms **prohibit one-to-one translations**

The **doctor** asked the nurse to help her in the procedure.

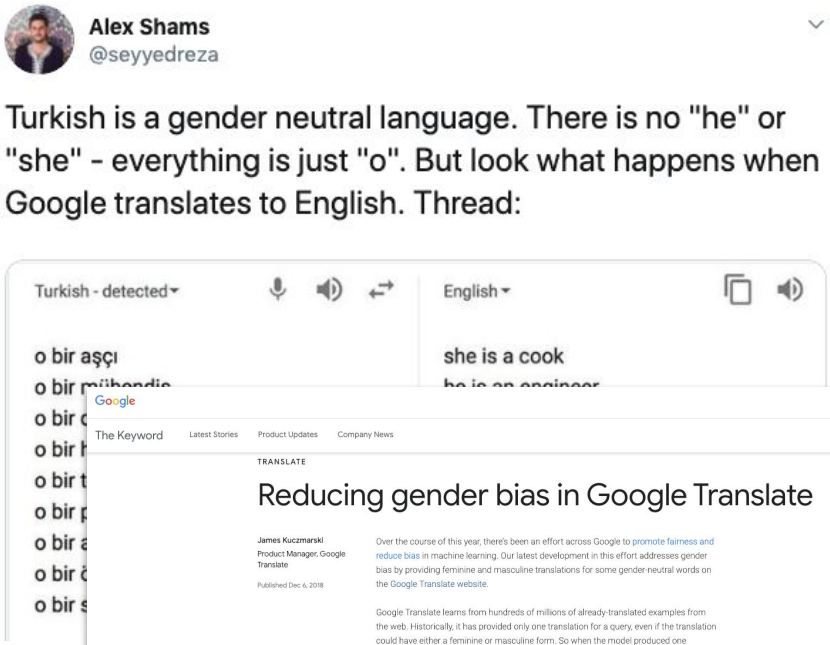
La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Is MT gender biased?

Is MT gender biased?



Is MT gender biased?



Is MT gender biased?

Alex Shams
@seyyedreza

Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:

Turkish - detected

o bir aşçı

English

she is a cook

Google

The Keyword | Latest Stories | Product Updates | Company News

TRANSLATE

Reducing gender bias in Google Translate

James Kuczmarski
Product Manager, Google Translate

Published Dec 6, 2018

Over the course of this year, there's been an effort across Google to promote fairness and reduce bias in machine learning. Our latest development in this effort addresses gender bias by providing feminine and masculine translations for some gender-neutral words on the Google Translate website.

Google Translate learns from hundreds of millions of already-translated examples from the web. Historically, it has provided only one translation for a query, even if the translation could have either a feminine or masculine form. So when the model produced one translation, it inadvertently replicated gender biases that already existed. For example: it would skew masculine for words like "strong" or "doctor," and feminine for other words, like "nurse" or "beautiful."

Now you'll get both a feminine and masculine translation for a single word—like "surgeon"—when translating from English into French, Italian, Portuguese or Spanish. You'll also get both translations when translating phrases and sentences from Turkish to English. For example, if you type "o bir doktor" in Turkish, you'll now get "she is a doctor" and "he is a doctor" as the gender-specific translations.

Gretchen McCulloch @GretchenAMcC · 6 Dec 2018
I actually like this a lot.

Introduces the human back into the equation by acknowledging ambiguity, letting us decide which translation fits a particular circumstance better.

Google @Google
To reduce gender bias in #GoogleTranslate, we're providing feminine and masculine translations for queries that include gender neutral words → goo.gl/vgvt7v

GIF

7 78 289

(((ل)()ل() 'yoav))))
@yoavgo

Follow

Replying to @GretchenAMcC

almost working...

(but really, a very hard and interesting problem to solve.)

DETECT LANGUAGE | SPANISH | ENGLISH | FRENCH | ITALIAN | SPANISH | FRENCH

the smart surgeon met the smart nurse

le chirurgien intelligent a rencontré l'infirmière intelligente

37/5000

Research Questions

1. Can we *quantitatively* evaluate gender translation in MT?

Research Questions

1. Can we *quantitatively* evaluate gender translation in MT?

2. How much does MT rely on *gender stereotypes* vs. meaningful context?

Research Questions

1. Can we *quantitatively* evaluate gender translation in MT?

2. How much does MT rely on *gender stereotypes* vs. meaningful context?

3. Can we reduce gender bias by rephrasing source texts?

Research Questions

1. **Can we *quantitatively* evaluate gender translation in MT?**

2. How much does MT rely on *gender stereotypes* vs. meaningful context?

3. Can we reduce gender bias by rephrasing source texts?

English Source Texts

- Winogender (Rudinger et al., 2018) & WinoBias (Zhao et al., 2018)
  - 3888 English sentences designed to test gender bias in *coreference resolution*
  - Following the **Winograd schema**

The **doctor** asked the nurse to help her in the procedure.

The **doctor** asked the nurse to help him in the procedure.

English Source Texts

- Winogender (Rudinger et al., 2018) & WinoBias (Zhao et al., 2018)
  - 3888 English sentences designed to test gender bias in *coreference resolution*
  - Following the **Winograd schema**

The **doctor** asked the nurse to help her in the procedure.

The **doctor** asked the nurse to help him in the procedure.

- **Observation**: These are very useful for evaluating gender bias in MT!

English Source Texts

- Winogender (Rudinger et al., 2018) & WinoBias (Zhao et al., 2018)
  - 3888 English sentences designed to test gender bias in *coreference resolution*
  - Following the **Winograd schema**

The **doctor** asked the nurse to help her in the procedure.

The **doctor** asked the nurse to help him in the procedure.

- **Observation**: These are very useful for evaluating gender bias in MT!
  - Equally split between stereotypical and non-stereotypical role assignments
  - Gold annotations for gender

Methodology: Automatic evaluation of gender bias

Input: MT model + target language
Output: Accuracy score for gender translation

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

The **doctor** asked the nurse to help her in the procedure.

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

The **doctor** asked the nurse to help her in the procedure.

La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)

Input: MT model + target language
Output: Accuracy score for gender translation

The **doctor** asked the nurse to help her in the procedure.

La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)

3. **Identify** gender in target language
   - Using off-the-shelf morphological analyzers or simple heuristics in the target languages

The **doctor** asked the nurse to help her in the procedure.

La **doctora** le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender
2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)
3. **Identify** gender in target language
   - Using off-the-shelf morphological analyzers or simple heuristics in the target languages

Input: MT model + target language
Output: Accuracy score for gender translation

Quality estimated at > 85% vs. 90% IAA
Doesn't require reference translations!

The **doctor** asked the nurse to help her in the procedure.

La **doctora** le pidió a la enfermera que le ayudara con el procedimiento.

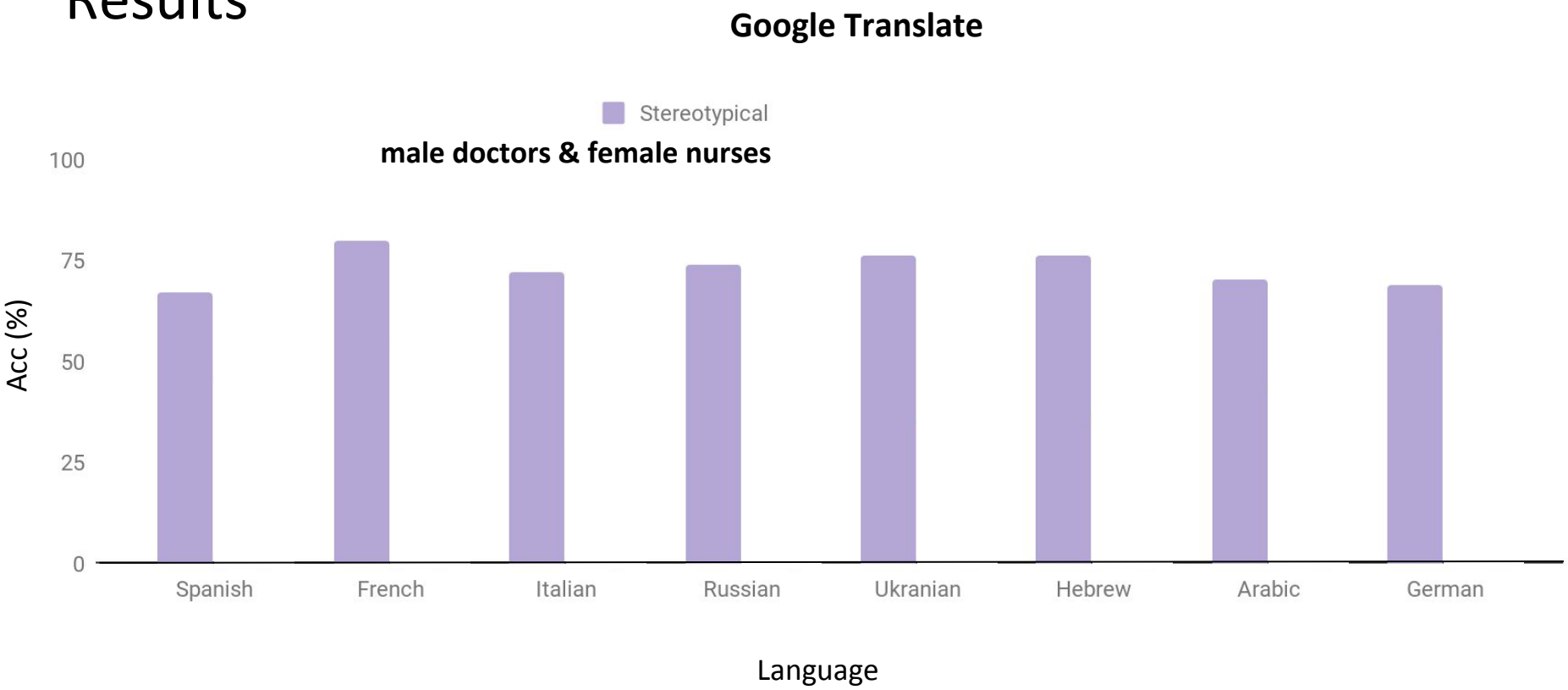
Research Questions

1. How well does machine translation handle gender?

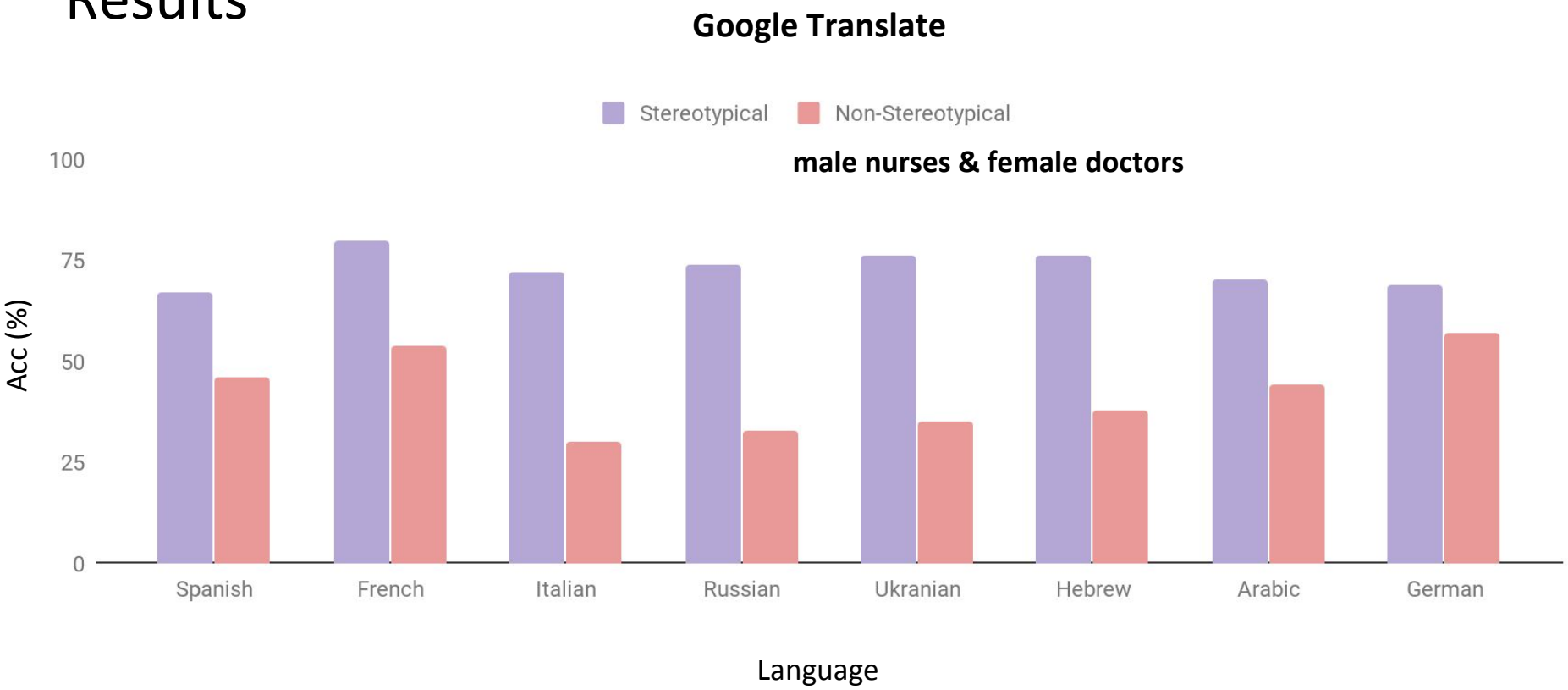
2. **How much does MT rely on *gender stereotypes* vs. meaningful context?**

3. Can we reduce gender bias by rephrasing source texts?

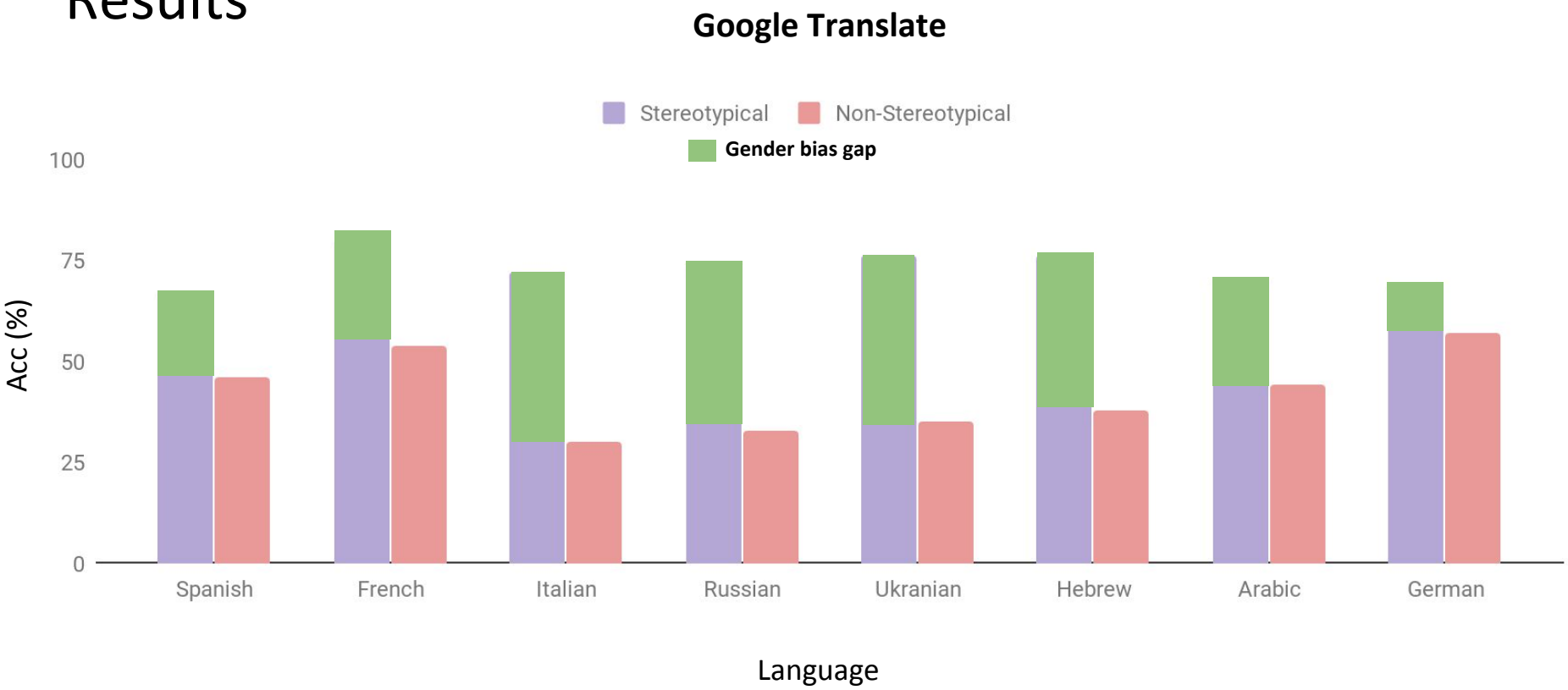
Results


Google Translate
Stereotypical
male doctors & female nurses
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

Results


Google Translate
Stereotypical
Non-Stereotypical
male nurses & female doctors
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

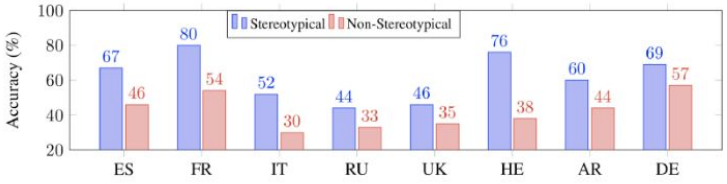
Results


Google Translate
Stereotypical
Non-Stereotypical
Gender bias gap
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

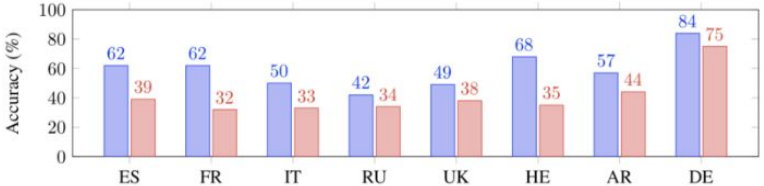
Results

- **MT struggles with non-stereotypical roles across languages and systems**
  - Often doing significantly worse than *random coin-flip*

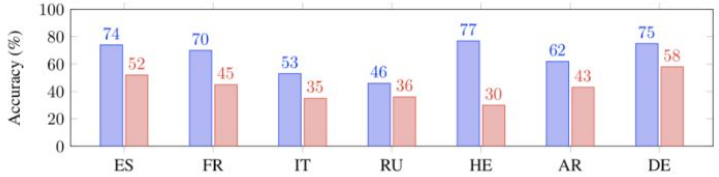
- **Academic models** (Ott et al., 2018; Edunov et al., 2018) **exhibit similar behavior**



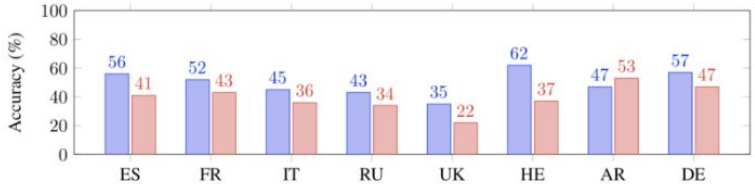
Google Translate performance on gender prediction.



Microsoft Translator

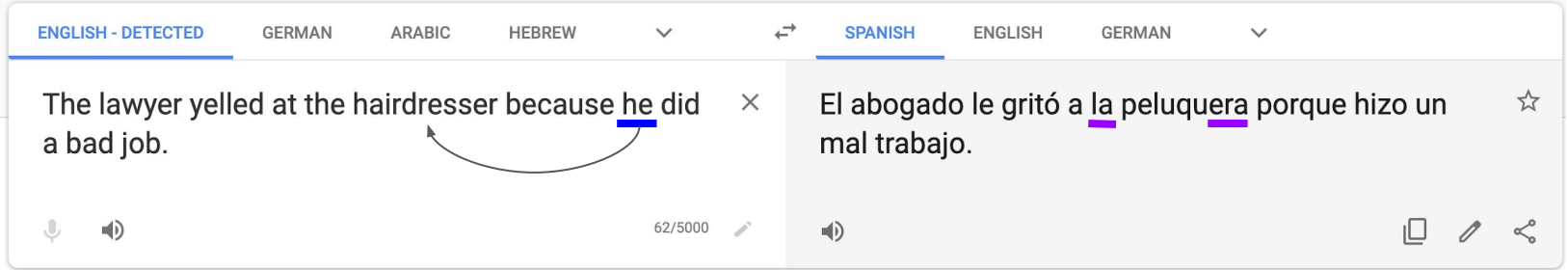


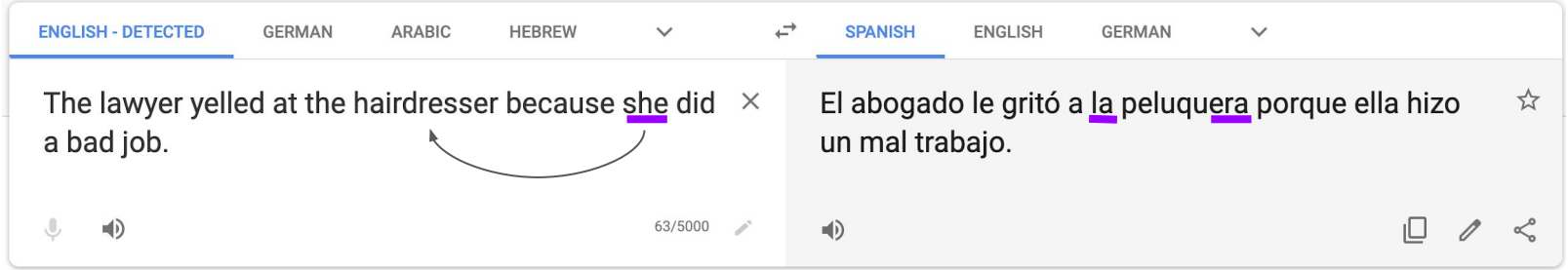
Amazon Translate



SYSTRAN

Examples


ENGLISH - DETECTED
GERMAN
ARABIC
HEBREW
SPANISH
ENGLISH
GERMAN
The lawyer yelled at the hairdresser because he did a bad job.
62/5000
El abogado le gritó a la peluquera porque hizo un mal trabajo.


ENGLISH - DETECTED
GERMAN
ARABIC
HEBREW
SPANISH
ENGLISH
GERMAN
The lawyer yelled at the hairdresser because she did a bad job.
63/5000
El abogado le gritó a la peluquera porque ella hizo un mal trabajo.

Research Questions

1. How well does machine translation handle gender?

2. How much does MT rely on gender stereotypes vs. meaningful context?

3. **Can we reduce gender bias by rephrasing source texts?**

Do Gendered Adjectives Affect Translation?

- Black-box injection of gendered adjectives (similar to Moryossef et al., 2019)
  - *the **pretty** doctor asked the nurse to help her in the operation*
  - *the **handsome** nurse asked the doctor to help him in the operation*

Do Gendered Adjectives Affect Translation?

- Black-box injection of gendered adjectives (similar to Moryossef et al., 2019)
  - *the **pretty** doctor asked the nurse to help <u>her</u> in the operation*
  - *the **handsome** nurse asked the doctor to help <u>him</u> in the operation*

- Improved performance for most tested languages and models [mean +8.6%]
  - + 10% on Spanish and Russian

Do Gendered Adjectives Affect Translation?

- Black-box injection of gendered adjectives (similar to Moryossef et al., 2019)
  - *the **pretty** doctor asked the nurse to help <u>her</u> in the operation*
  - *the **handsome** nurse asked the doctor to help <u>him</u> in the operation*

- Improved performance for most tested languages and models [mean +8.6%]
  - + 10% on Spanish and Russian

- Requires oracle coreference resolution!
  - Attests to the relation between coreference resolution and MT

Limitations & Future Work

- Artificially-created dataset
  - Allows for controlled experiment
  - Yet, might introduce its own annotation biases

- Medium-size
  - Easy to overfit - not good for training

Limitations & Future Work

- Artificially-created dataset
  - Allows for controlled experiment
  - Yet, might introduce its own annotation biases

- Medium-size
  - Easy to overfit - not good for training

- Future work
  - **Collect naturally occurring samples on a large scale**

Conclusion

- First quantitative automatic evaluation of gender bias in MT
  - 6 SOTA MT models on 8 diverse target languages
  - Doesn’t require reference translations

- **Significant gender bias found in all models in all tested languages**

- Code and data: https://github.com/gabrielStanovsky/mt_gender
  - Easily extensible with more languages and MT models

Conclusion

Come to the the **Gender Bias Workshop!** (Friday)

- First quantitative automatic evaluation of gender bias in MT
  - 6 SOTA MT models on 8 diverse target languages
  - Doesn't require reference translations
- **Significant gender bias found in all models in all tested languages**
- Code and data: https://github.com/gabrielStanovsky/mt_gender
  - Easily extensible with more languages and MT models

Спасибі за слухання! Grazie per aver ascoltato! Danke fürs Zuhören!

תודה על ההקשבה!

Thanks for listening!

¡Gracias por su atención!

Merci pour l'écoute!

شكرا على الإنصات!

Спасибо за внимание!