

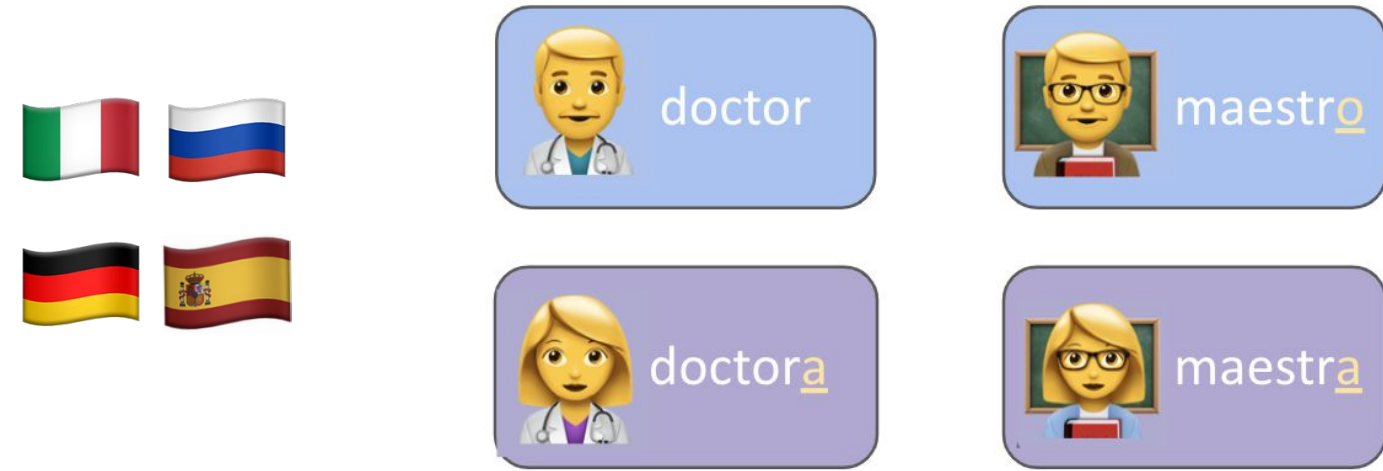


Evaluating Gender Bias in Machine Translation

Gabriel Stanovsky, Noah Smith and Luke Zettlemoyer

Grammatical Gender

Some languages encode grammatical gender



... others don't



Translating Gender

Variations in gender mechanisms across languages prohibit one-to-one translations

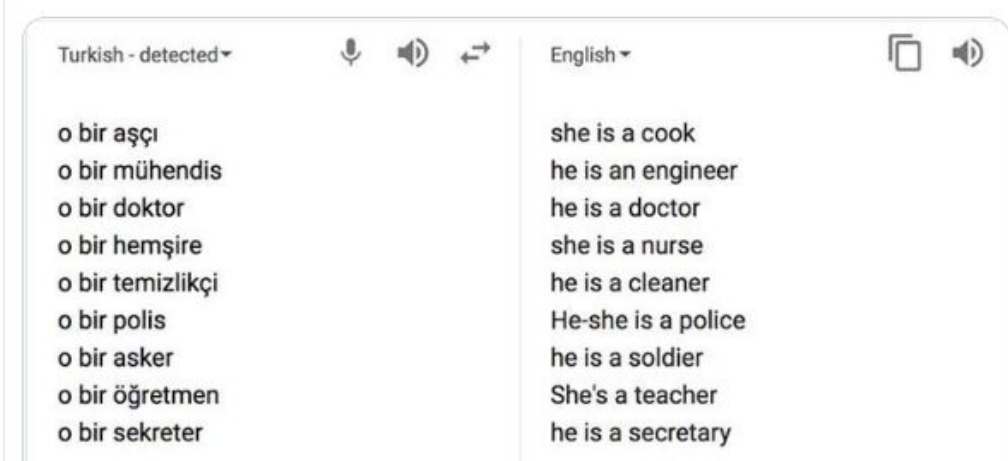
The **doctor** asked the nurse to help her in the procedure.

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

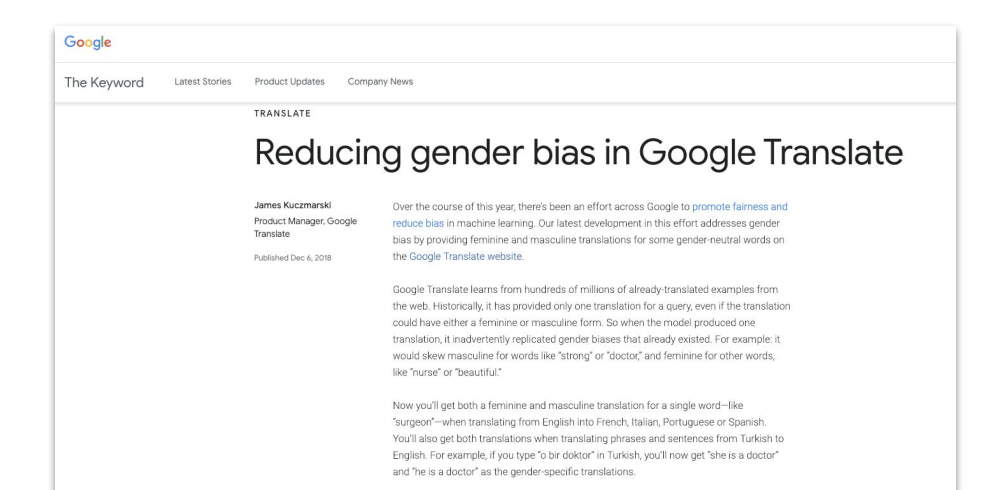
→ Translating between English Spanish, for example, requires **making decisions about gender**

Is MT Gender-Biased?

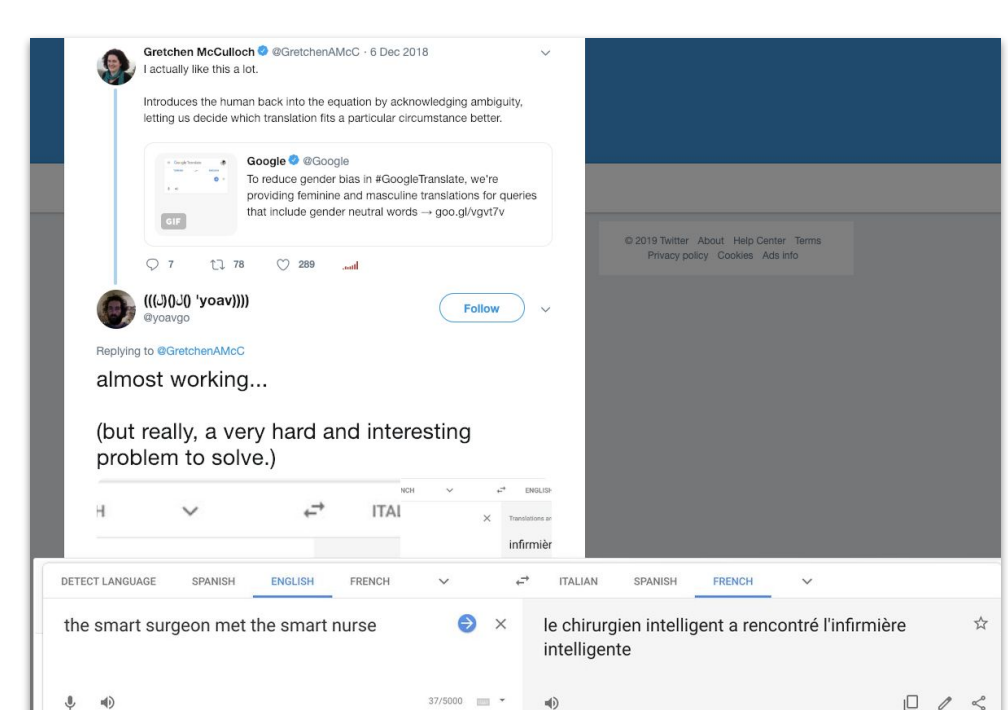
Alex Shams @sheyredreza
Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:



- Anecdotal evidence that popular MT services resort to stereotypical role assignments



- Google Translate introduces a feature to control gender inflection for individual words



- Not yet a viable solution for complete sentences

Estimating MT Gender Accuracy

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

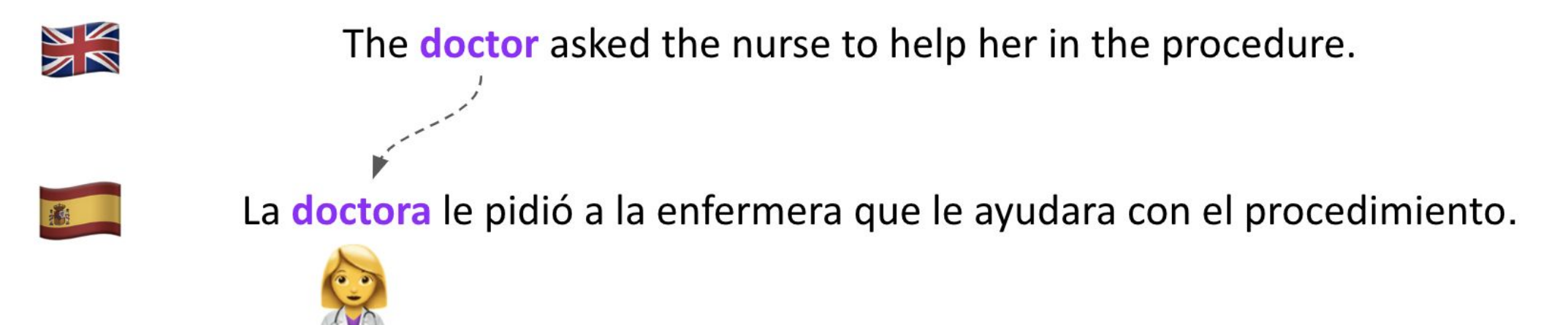
The **doctor** asked the nurse to help her in the procedure.

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

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

Automatic evaluation

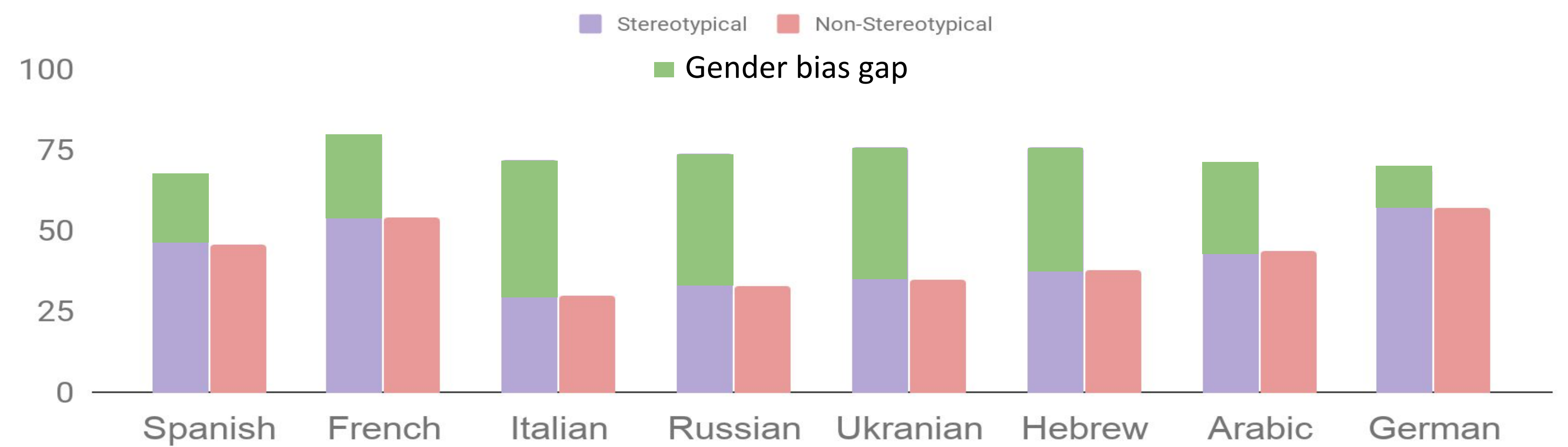
1. **Translate the coreference bias datasets**
To target languages with grammatical gender
2. **Align between source and target**
Using fast align (Dyer et al., 2013)
3. **Identify gender in target language.**
Using off-the-shelf morphological analyzers or simple heuristics in the target languages



In: MT model + target language
Out: Accuracy score for gender translation
Quality estimated at > 85% vs. 90% IAA
Doesn't require reference translations!

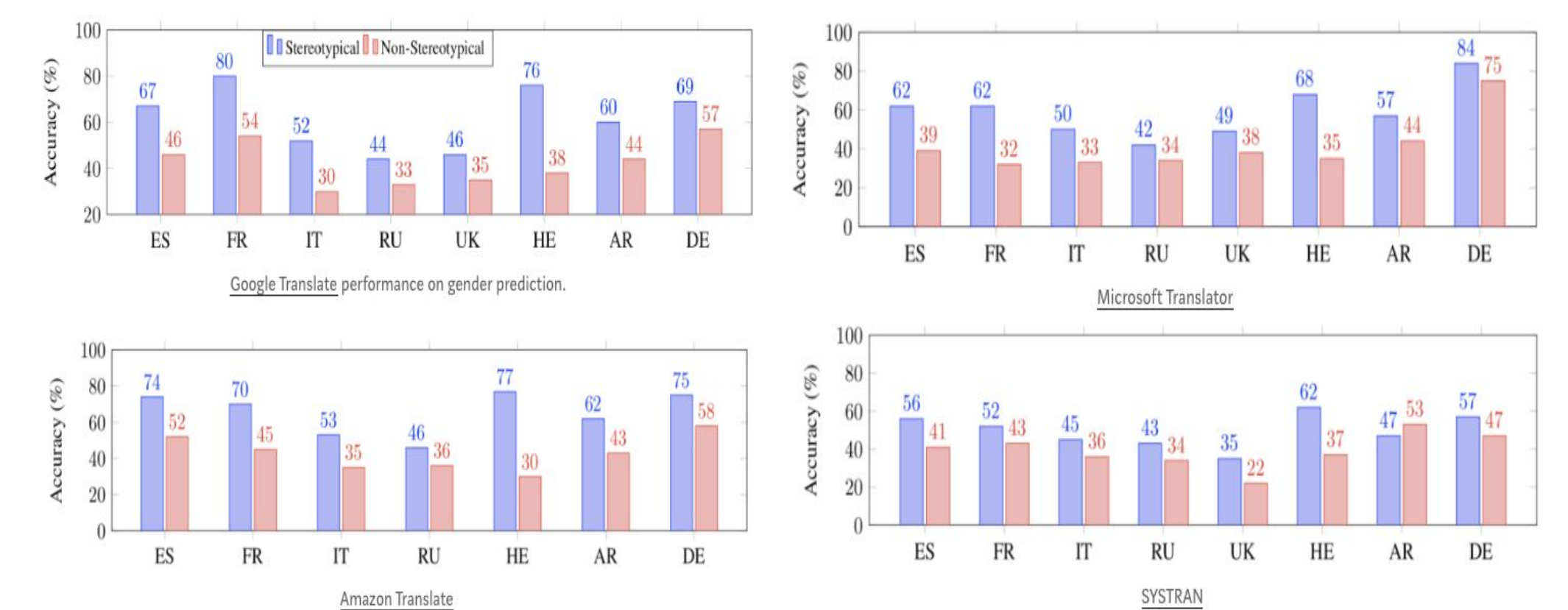
Main Findings

Google Translate



All models were significantly gender-biased

- 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



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
- Improved performance for most tested languages and models
 - + 10% on Spanish and Russian [mean +8.6%]
- **Requires oracle coreference resolution!**
 - Attests to the relation between coreference resolution and machine translation
 - Improvement in coreference resolution directly improves the accuracy of gender translation

