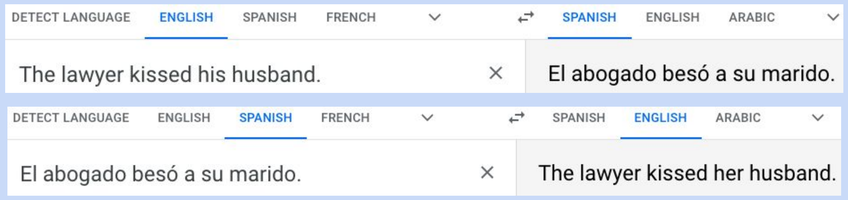
Whose wife is it anyway? Bias in machine translation of same-gender relationships

Ian Stewart (ianbstew@umich.edu)
Computer Science and Engineering, University of Michigan

Motivation

- Modern assessments of gender bias in NLP focus on **isolated** gender stereotypes, e.g. sim("woman", "nurse") > sim("man", "nurse") (Bolukbasi et al. 2016).
- Many stereotypes are encoded in **relationships**! E.g. does a language model predict that "the man" will have "a boyfriend" or "a girlfriend"?
- Important for **Machine Translation**, from languages with rich morphological gender (Sp. *el esposo*) to limited gender (Eng. *his husband*).



- Based on surface behavior: **English LM** seems to have strong different-gender (DG) bias, at the cost of same-gender (SG) relationships.

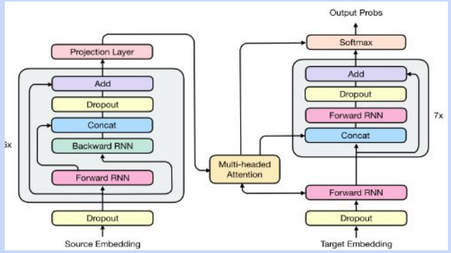
Does Google Translate exhibit SG bias across variety of contexts?

Experiment design

- Goal: test whether SG relationships can be translated as accurately as DG relationships (metric = % correct).
- Source domain: gender-NP languages w/ **gender-ambiguous possessives** (French, Italian, Spanish)
  - High-resource, similar grammar, long-term cultural exposure to SG relationships.
  - Known gender bias in FR, ES (Zhao et al. 2020)
- Target domain: no-gender-NP language w/ **gendered possessives** (English)

Occupation (Gonen and Goldberg 2019)	el abogado (M; "lawyer") la abogada (F)	100
Relationship **template**	X besó a Y ("X kissed Y")	5
Relationship **target**	el novio (M; "boyfriend") la novia (F; "girlfriend")	6
Sentence	El abogado besó a su novio. ("The lawyer kissed his boyfriend.")	3000

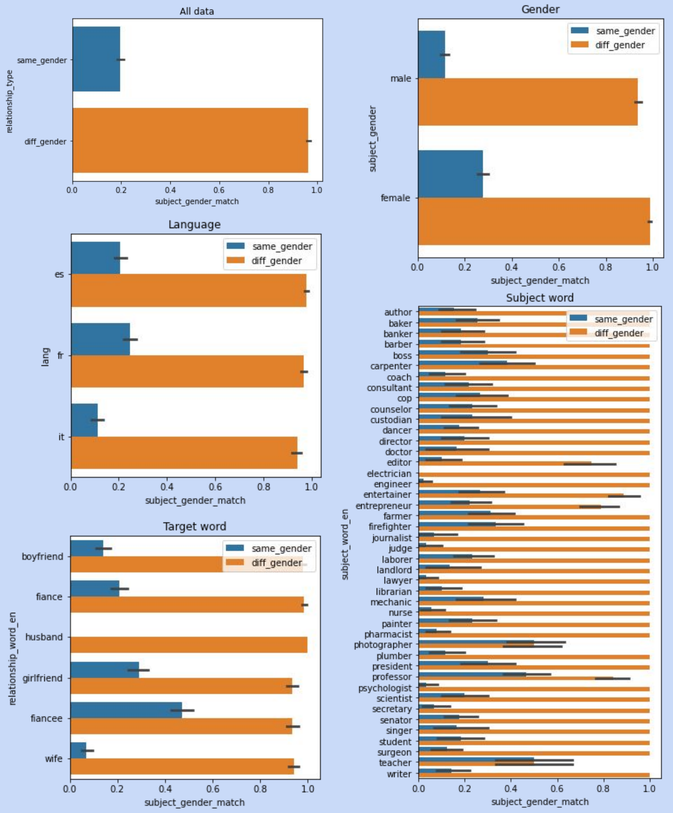
El abogado besó a su novio.



The lawyer kissed his boyfriend.

RNN-based Neural Machine Translation (Chen et al. 2018)

Results



- Potential sources of bias:
  - Unbalanced training data
  - Grammatical "freezing"?
    (cf. coreference resolution with gender-neutral pronouns; Cao and Daumé 2020)
- Potential **harms** in bias:
  - Invalidate someone's relationship
  - Reinforce heteronormative standards
- Need to test representation of **basic relationships** in language models
  - Romantic, family, power status (Prabhakaran et al. 2012), social roles
- Text-as-data **extensions**
  - What is the correlation between SG % correct vs. **representation of LGBTQ people** in occupation?
  - Can we retrain MT model to recognize SG relationships?
  - Can we find a **latent SG dimension** in sentence representations, and use this to identify other latent SG sentences?