Social NLP: Model gender bias

CS 685, Spring 2021

Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

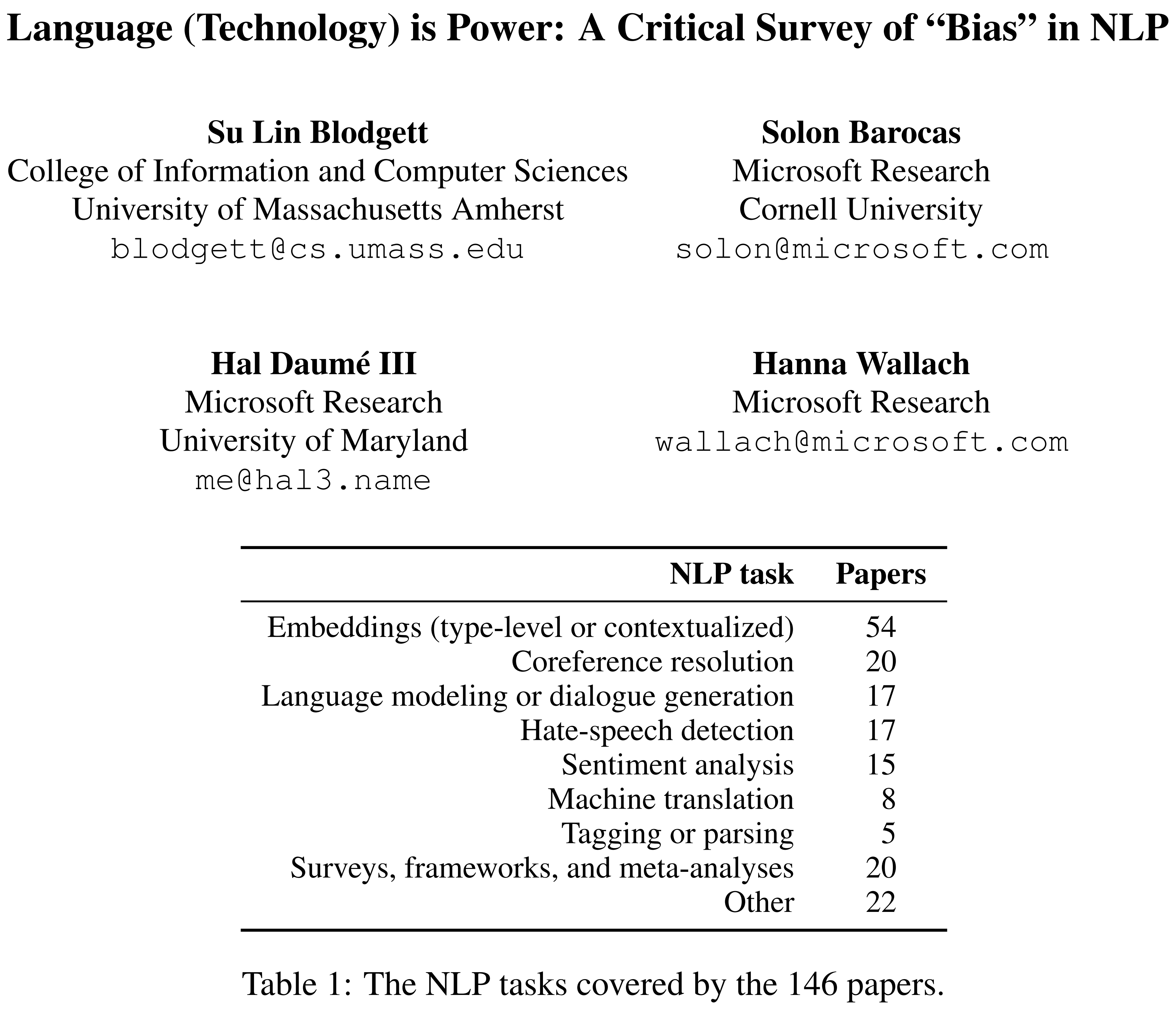
https://people.cs.umass.edu/~brenocon/cs685_s21/

Brendan O’Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- Language is socially situated
  - **By** and **for** communicators
    - Wednesday; focus on *dialect*
  - often, it's also **about** people
    - Today; focus on *gender*



Language (Technology) is Power: A Critical Survey of “Bias” in NLP

Su Lin Blodgett
College of Information and Computer Sciences
University of Massachusetts Amherst
blodgett@cs.umass.edu

Solon Barocas
Microsoft Research
Cornell University
solon@microsoft.com

Hal Daumé III
Microsoft Research
University of Maryland
me@hal3.name

Hanna Wallach
Microsoft Research
wallach@microsoft.com

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Table 1: The NLP tasks covered by the 146 papers.

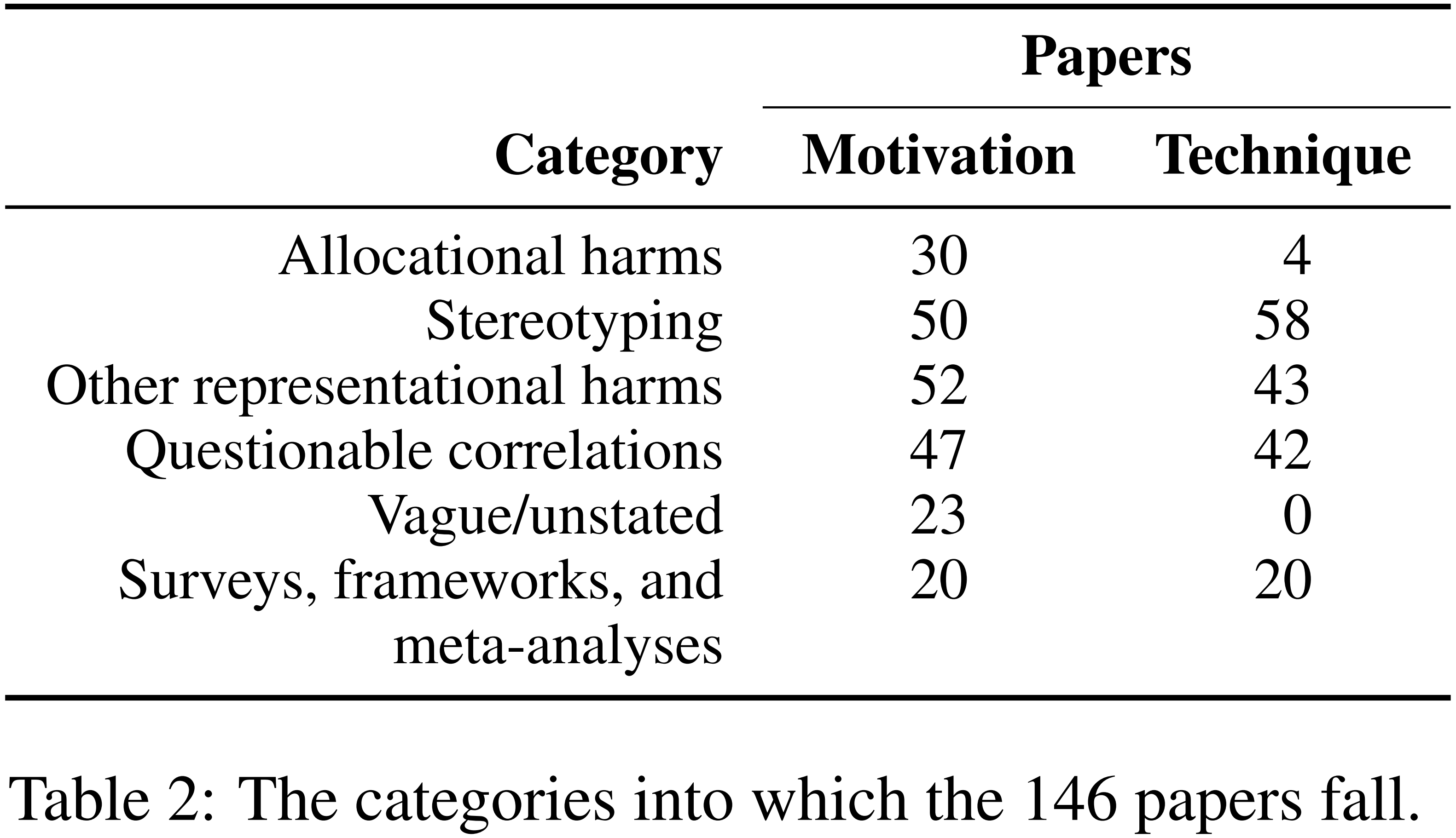
- *[Blodgett et al., ACL 2020]*

From applications
- credit scoring, resume screening
- historical bias in training data

- **Allocational harms**: "when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups"

- Lang generation? MT?

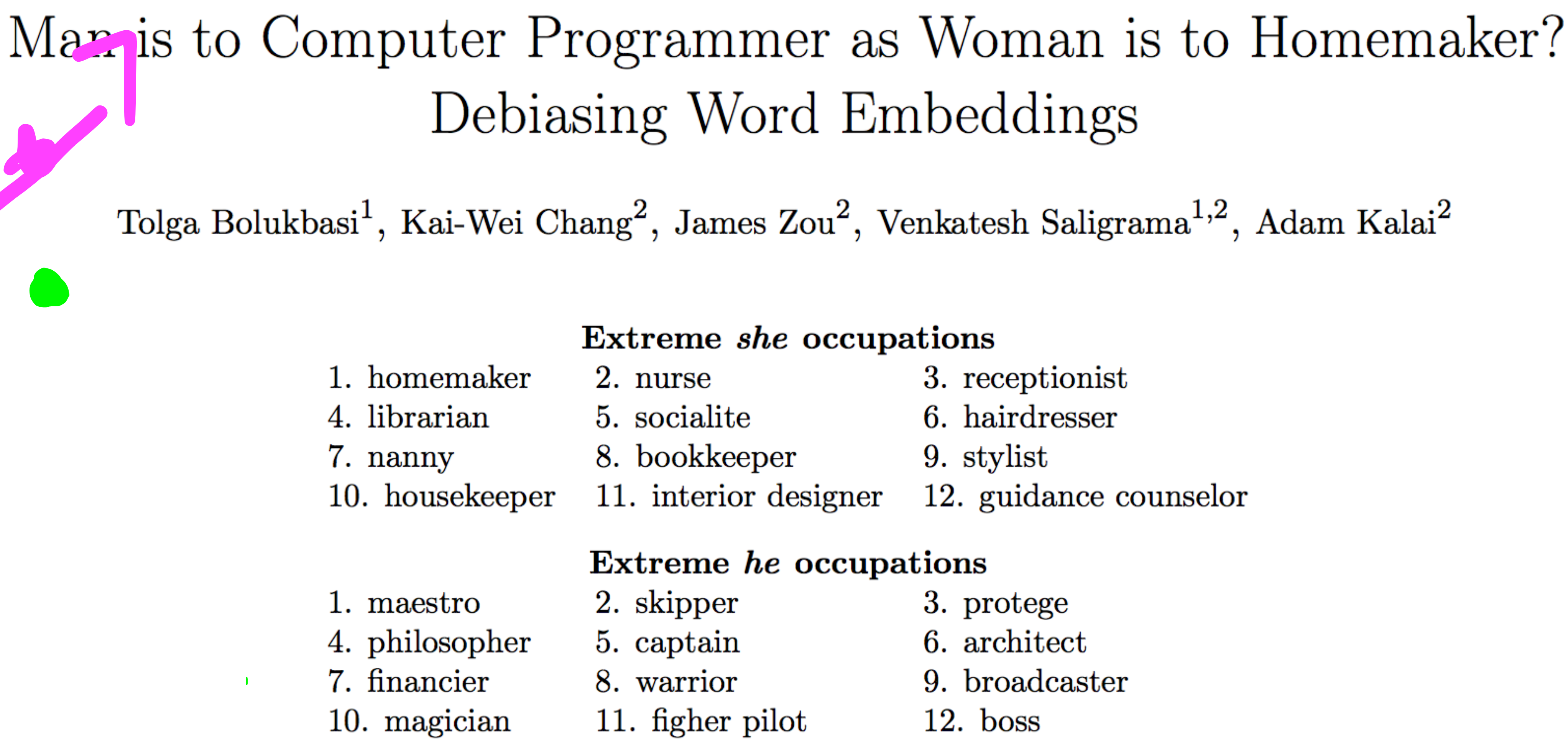
- **Representational harms**: "when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether."



	Papers	
Category	**Motivation**	**Technique**
Allocational harms	30	4
Stereotyping	50	58
Other representational harms	52	43
Questionable correlations	47	42
Vague/unstated	23	0
Surveys, frameworks, and meta-analyses	20	20

Table 2: The categories into which the 146 papers fall.

Biases in word embeddings



Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

Extreme *she* occupations

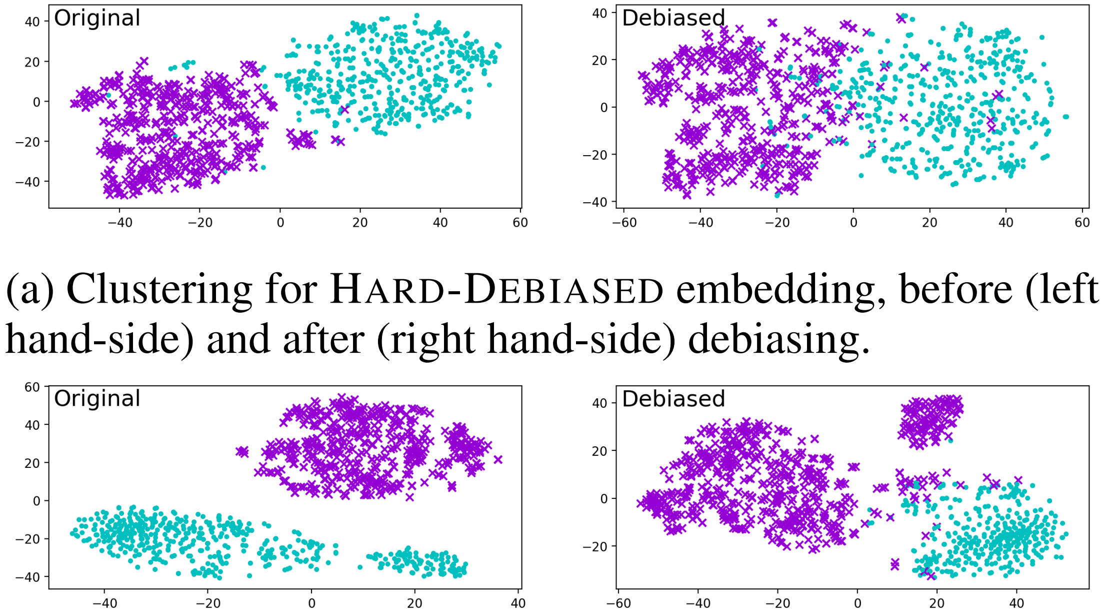
1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper
11. interior designer
12. guidance counselor

Extreme *he* occupations

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician
11. figher pilot
12. boss

- Does this have implications as either allocational or representational harms (or otherwise)?

- Can you "de-bias"? Bolukbasi et al. (2016) proposed a linear projection postprocessing step to de-bias embeddings. But Gonen and Goldberg (2019) showed the nearest-neighbor / clustering structure still encodes lots of gender information!



(a) Clustering for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.

(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Figure 1: Clustering the 1,000 most biased words, before and after debiasing, for both models.

Gender Bias in Coreference Resolution

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme
Johns Hopkins University

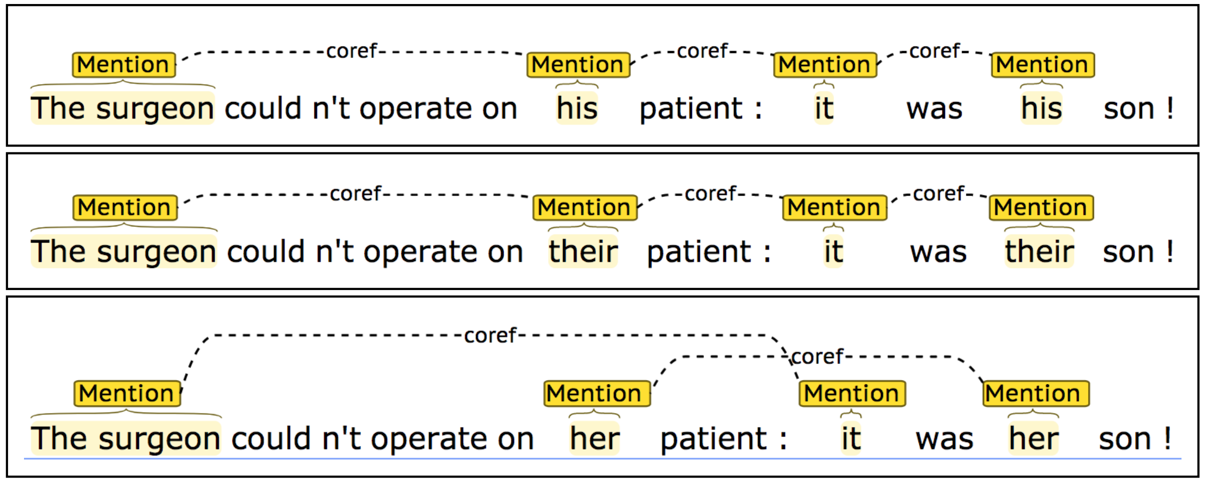
There is a classic riddle: *A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, “I can’t operate on this boy, he’s my son!”* **How can this be?**

Gender Bias in Coreference Resolution

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme
Johns Hopkins University

Abstract

We present an empirical study of gender bias in coreference resolution systems. We first introduce a novel, Winograd schema-style set of minimal pair sentences that differ only by pronoun gender. With these *Winogender schemas*, we evaluate and confirm systematic gender bias in three publicly-available coreference resolution systems, and correlate this bias with real-world and textual gender statistics.



The surgeon could n't operate on his patient : it was his son !

The surgeon could n't operate on their patient : it was their son !

The surgeon could n't operate on her patient : it was her son !

Figure 1: Stanford CoreNLP rule-based coreference system resolves a male and neutral pronoun as coreferent with “The surgeon,” but does not for the corresponding female pronoun.

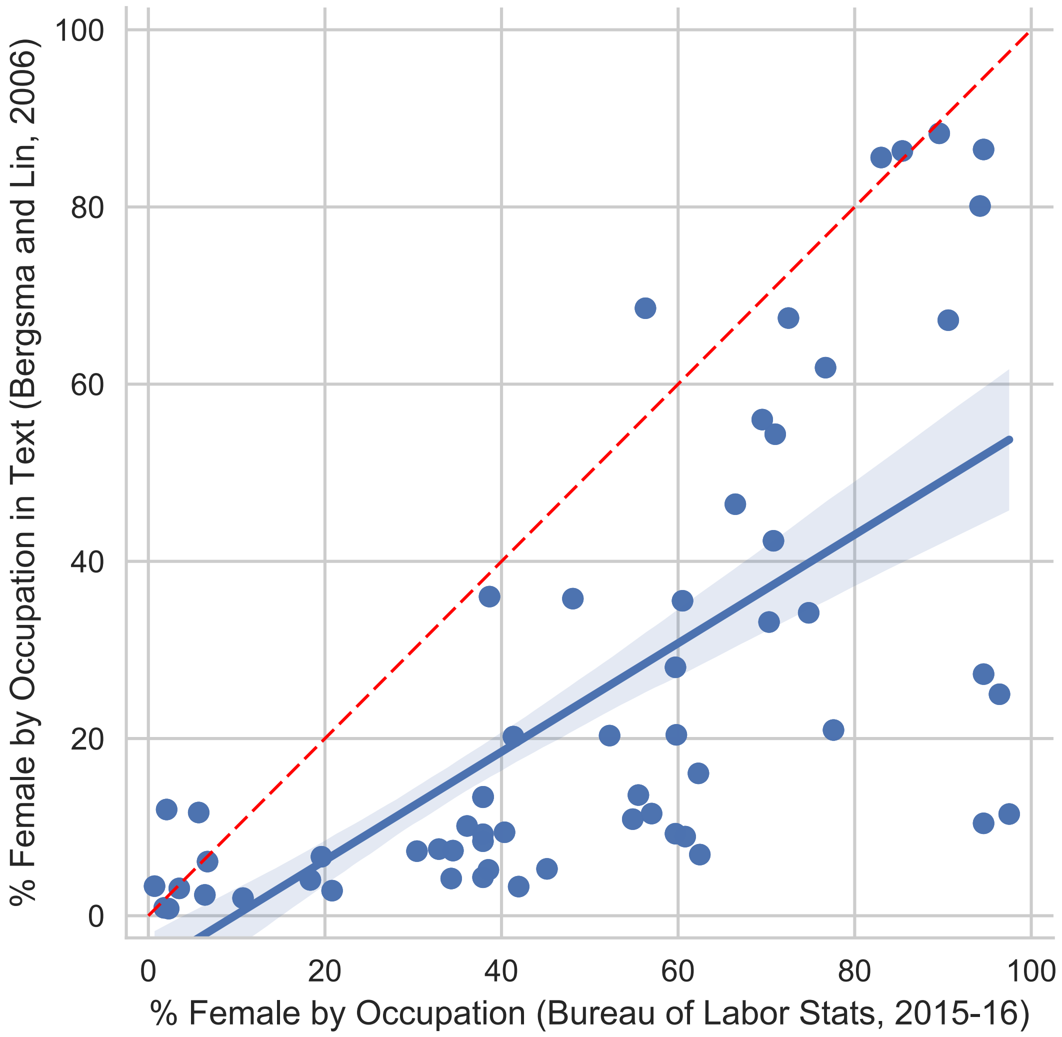
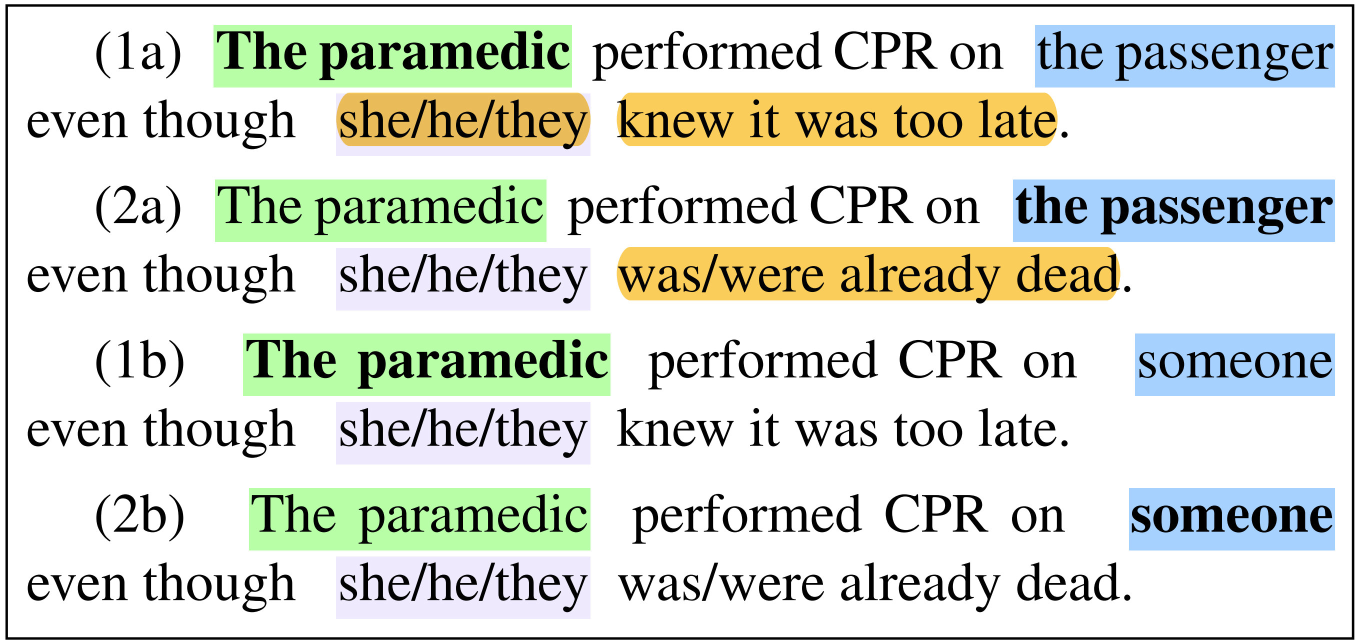


Figure 3: Gender statistics from Bergsma and Lin (2006) correlate with Bureau of Labor Statistics 2015. However, the former has systematically lower female percentages; most points lie well below the 45-degree line (dotted). Regression line and 95% confidence interval in blue. Pearson r = 0.67.

- Bergsma and Lin: infer nouns' gender and number from distributional syntactic path stats from unlabeled corpus



(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

Figure 2: A "Winogender" schema for the occupation *paramedic*. Correct answers in bold. In general, OCCUPATION and PARTICIPANT may appear in either order in the sentence.

By multiple measures, the Winogender schemas reveal varying degrees of gender bias in all three systems. First we observe that these systems do not behave in a gender-neutral fashion. That is to say, we have designed test sentences where correct pronoun resolution is not a function of gender (as validated by human annotators), but system predictions do exhibit sensitivity to pronoun gender: 68% of male-female minimal pair test sentences are resolved differently by the RULE system; 28% for STAT; and 13% for NEURAL.

Overall, male pronouns are also more likely to be resolved as OCCUPATION than female or neutral pronouns across all systems: for RULE, 72% male vs 29% female and 1% neutral; for STAT, 71% male vs 63% female and 50% neutral; and for NEURAL, 87% male vs 80% female and 36% neutral. Neutral pronouns are often resolved as neither OCCUPATION nor PARTICIPANT, possibly due to the number ambiguity of "they/their/them."

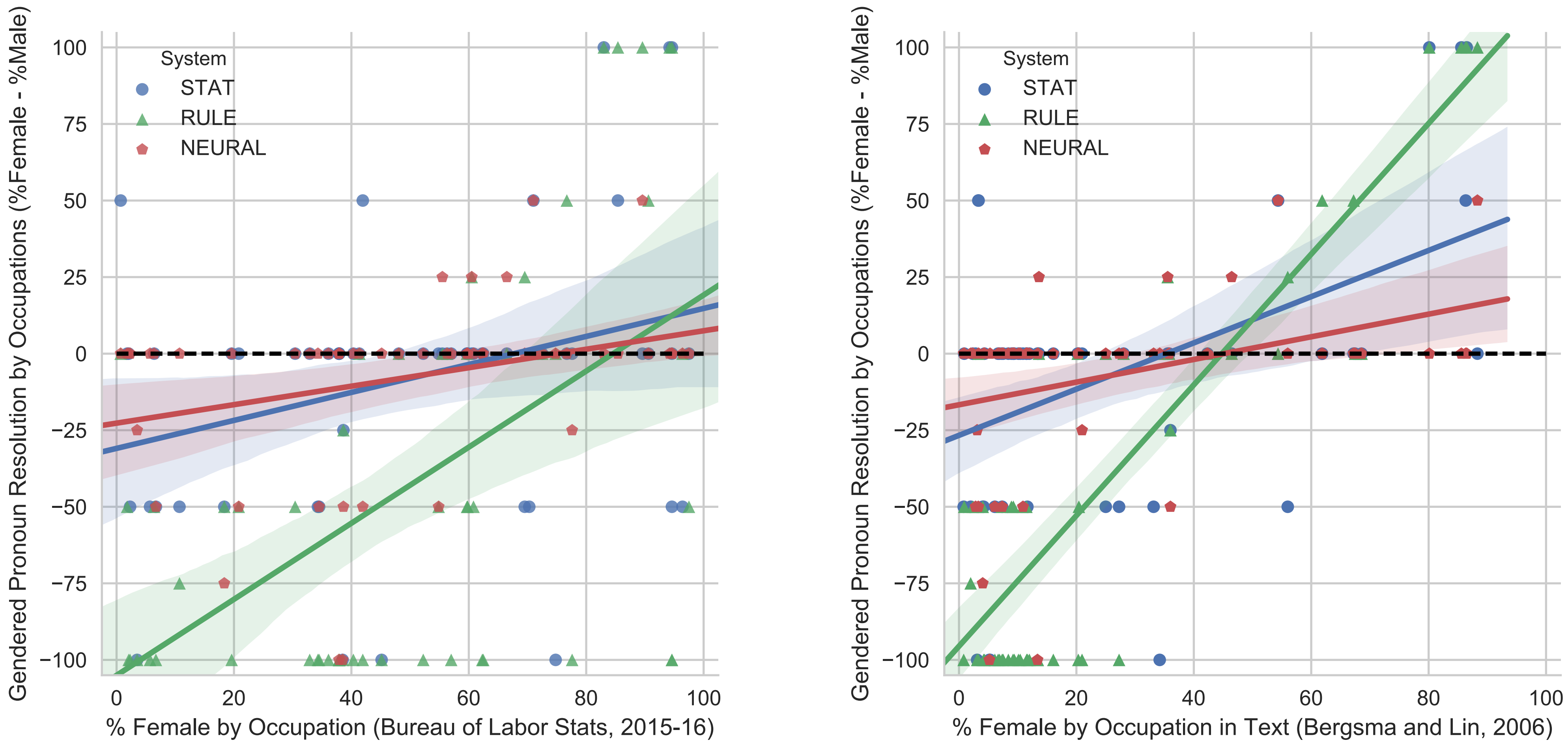


Figure 4: These two plots show how gender bias in coreference systems corresponds with occupational gender statistics from the U.S Bureau of Labor Statistics (left) and from text as computed by Bergsma and Lin (2006) (right); each point represents one occupation. The y-axes measure the extent to which a coref system prefers to match female pronouns with a given occupation over male pronouns, as tested by our Winogender schemas. A value of 100 (maximum female bias) means the system always resolved female pronouns to the given occupation and never male pronouns (100% - 0%); a score of -100 (maximum male bias) is the reverse; and a value of 0 indicates no gender differential. Recall the Winogender evaluation set is gender-balanced for each occupation; thus the horizontal dotted black line (y=0) in both plots represents a hypothetical system with 100% accuracy. Regression lines with 95% confidence intervals are shown.

Evaluating Gender Bias in Machine Translation

Gabriel Stanovsky[1,2], **Noah A. Smith**[1,2], and **Luke Zettlemoyer**[1]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[2]Allen Institute for Artificial Intelligence, Seattle, USA
{gabis,nasmith,lsz}@cs.washington.edu

Abstract

We present the first challenge set and evaluation protocol for the analysis of gender bias in machine translation (MT). Our approach uses two recent coreference resolution datasets composed of English sentences which cast participants into non-stereotypical gender roles (e.g., “The doctor asked the nurse to help *her* in the operation”). We devise an automatic gender bias evaluation method for eight target languages with grammatical gender, based on morphological analysis (e.g., the use of female inflection for the word “doctor”). Our

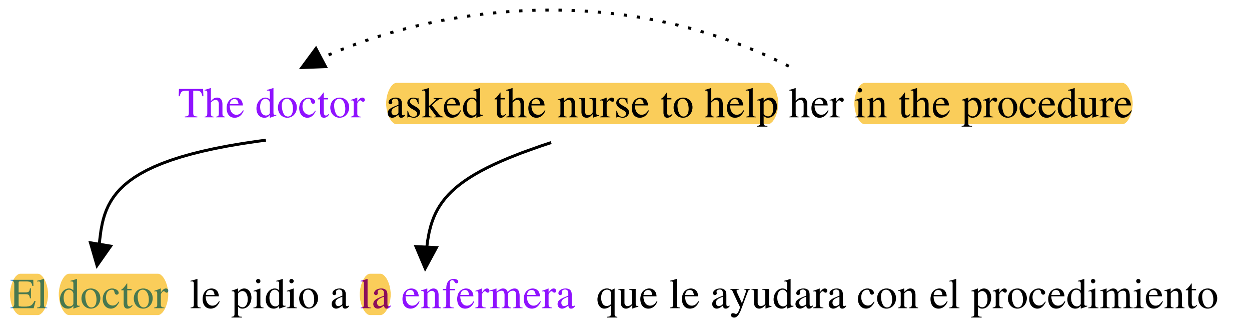

The doctor asked the nurse to help her in the procedure
El doctor le pidio a la enfermera que le ayudara con el procedimiento

Figure 1: An example of gender bias in machine translation from English (top) to Spanish (bottom). In the English source sentence, the nurse’s gender is unknown, while the coreference link with “her” identifies the “doctor” as a female. On the other hand, the Spanish target sentence uses morphological features for gender: “*el* doctor” (male), versus “*la* enfermera” (female). Aligning between source and target sentences reveals that a stereotypical assignment of gender roles changed the meaning of the translated sentence by changing the doctor’s gender.

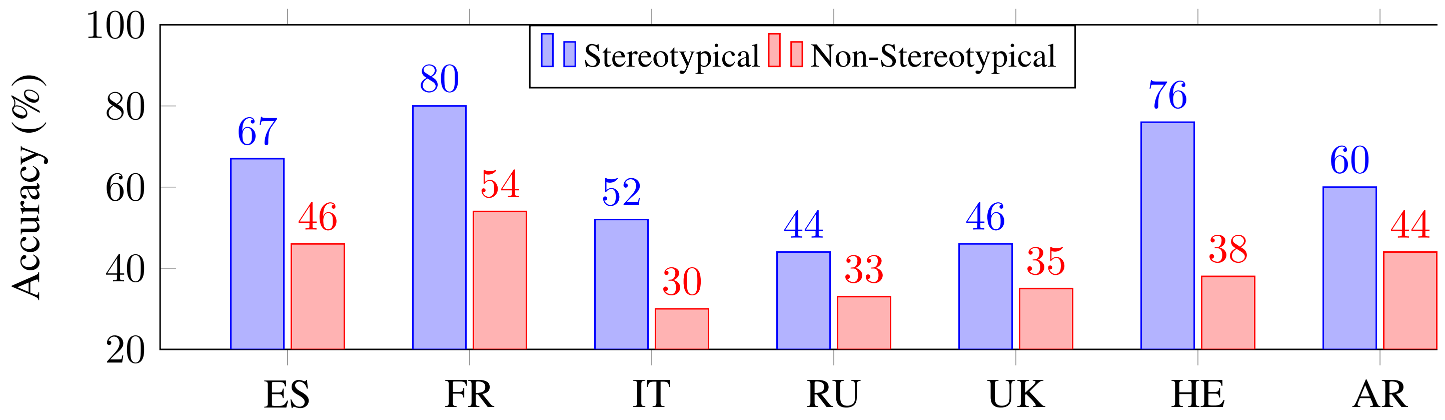


Figure 2: Google Translate's performance on gender translation on our tested languages. The pe stereotypical portion of WinoMT is consistently better than that on the non-stereotypical portio systems we tested display similar trends.

	Google Translate			**Microsoft Translator**			**Amazon Translate**[*]			**SYSTRAN**		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
ES	53.1	23.4	21.3	47.3	36.8	23.2	**59.4**	15.4	22.3	45.6	46.3	15.0
FR	**63.6**	6.4	26.7	44.7	36.4	29.7	55.2	17.7	24.9	45.0	44.0	9.4
IT	39.6	32.9	21.5	39.8	39.8	17.0	**42.4**	27.8	18.5	38.9	47.5	9.4
RU	37.7	36.8	11.4	36.8	42.1	8.5	**39.7**	34.7	9.2	37.3	44.1	9.3
UK	38.4	43.6	10.8	**41.3**	46.9	11.8	–	–	–	28.9	22.4	12.9
HE	**53.7**	7.9	37.8	48.1	14.9	32.9	50.5	10.3	47.3	46.6	20.5	24.5
AR	48.5	43.7	16.1	47.3	48.3	13.4	**49.8**	38.5	19.0	47.0	49.4	5.3