CS11-711 Advanced NLP

Fairness and Bias in NLP

Graham Neubig

Carnegie Mellon University

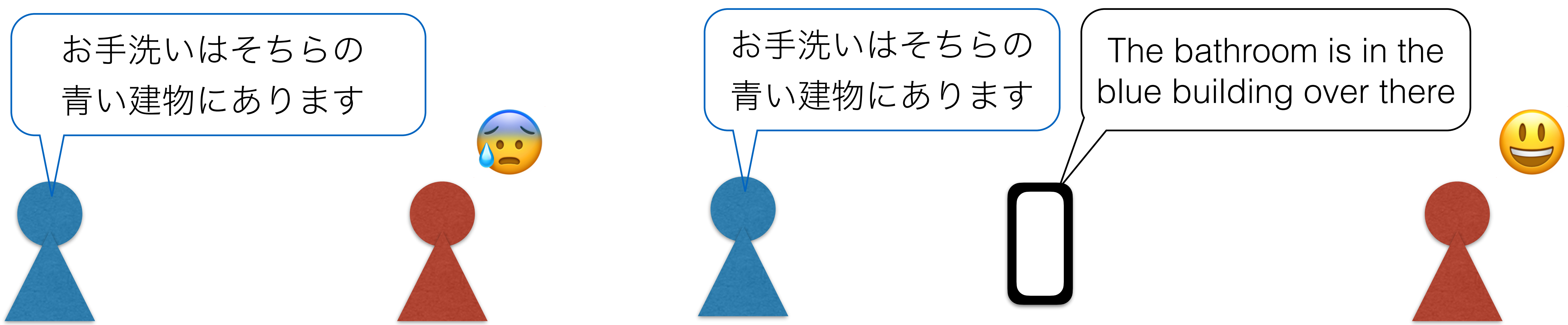
Language Technologies Institute

Site

https://phontron.com/class/anlp2022/

Some slides from Adam Lopez and Divyansh Kaushik

Why do we Build NLP?



- We want to make the world a better place!
- How do we quantify "better"?
- **Utility (economics):** the total satisfaction received from consuming a good or service.
- Inequal allocation of utility leads to issues of fairness (see Blodgett et al. 2020)

Potential Harm: Inequal Utility from NLP Systems

- **American English Speaker:** Use virtual assistant, car navigation system, translate text, benefit from good search technology

- **Japanese Speaker:** Use the above technology, maybe with fewer features, maybe a bit worse

- **Marshalese Speaker:** Don't use the above technology, or be forced to use it in a second language
- **Non-native Speaker, or Native Speaker Different from Training Data:** Have issues w/ pronunciation, mannerisms, etc

Image Credit: Wikipedia

Potential Harm: Allocational Harms

- Decisions made by an NLP system affect life positively/negatively and potentially fairly
- **Unfair Positive Allocation:** NLP system decides who gets a loan or accepted to university
- **Unfair Negative Allocation:** NLP system decides who gets arrested due to their social media posts

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Potential Harm: Sterotyping

- When a system reflects harmful societal biases in its output

- E.g., when translating gender neutral Turkish sentences into English, Google associates he/she pronouns with stereotypically male/female dominated jobs, etc.

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover
onu sevmiyor	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
o onu kucaklıyor	she is embracing her
o onu kucaklamıyor	he does not embrace it
o evli	she is married
o bekar	he is single
o mutlu	he's happy
o mutsuz	she is unhappy
o çalışkan	he is hard working
o tembel	she is lazy

Bias In Human Annotation

- For e.g., Toxicity classification datasets are biased against LGBTQ community (Dixon et al., 2017).
- Can arise from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves.
    - Different cultural and social norms. See Byrne (2016) and Fazelpour (2020).

Detecting Biases In NLP Systems

Commonly Employed Techniques

- Association tests
- Analyzing performance measures across groups
- Counterfactual evaluations

Word Embedding Association Test (WEAT)

- Embeddings learn from co-occurrence statistics (e.g., king - man + woman = queen)

- But what if text encodes unsavory stereotypes and biases? (e.g., doctor - man + woman = nurse)

- Consider
  two sets of target words (e.g., programmer, engineer, ... and nurse, teacher, ...)
  two sets of *attribute* words (e.g., man, male, ... and woman, female ...).

- **Null Hypothesis:** No difference between the two sets of target words in terms of similarity to the two sets of attribute words.

Mathematical Formulation

- X, Y are sets of target words of equal size, and A, B the two sets of attribute words.

- The test statistic is:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad \textbf{where}$$

$$s(w, A, B) = \textbf{mean}_{a \in A}\textbf{cos}(\vec{w}, \vec{a}) - \textbf{mean}_{b \in B}\textbf{cos}(\vec{w}, \vec{b})$$

- s(w, A, B): association of w with the attribute.

- s(X, Y, A, B): differential association of two sets of target words with the attribute.

- $\{(X_i, Y_i)\}_i$ all the partitions of $X \cup Y$ into two sets of equal size. The one-sided p-value of the permutation test is $Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$.

Associative Biases In Word Embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017)

- Use WEAT to show that word embeddings exhibit human like social biases.

Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

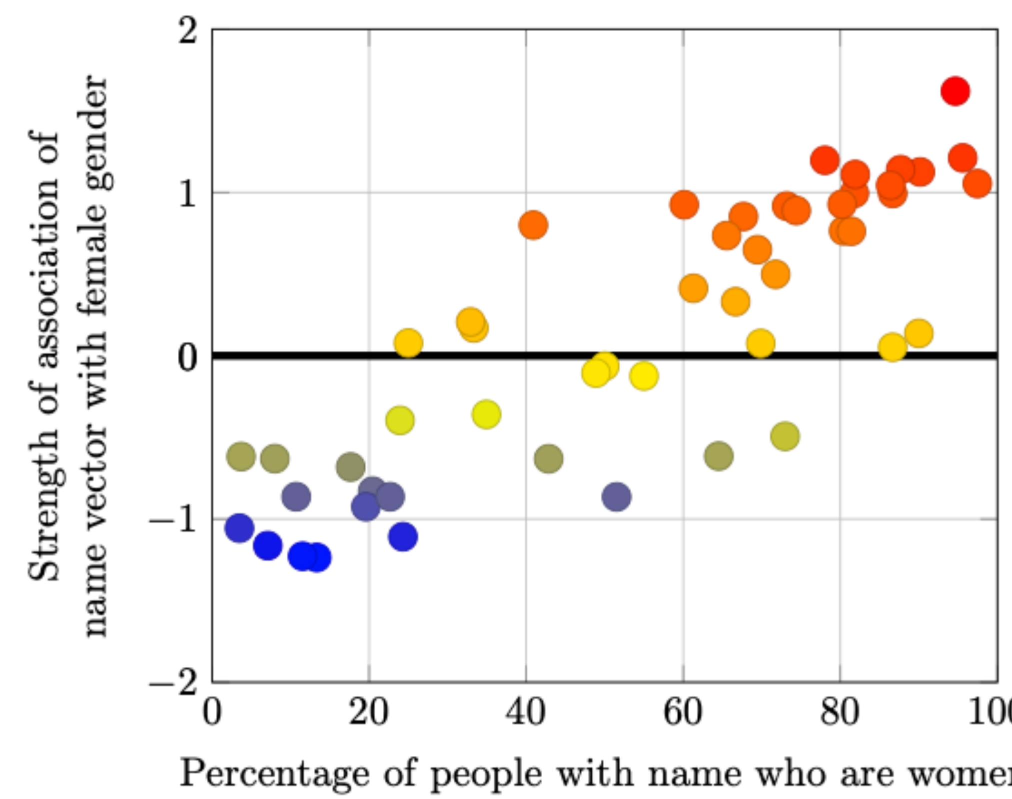


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Extending Embedding Association Test To Sentences (May et al., 2019)

- Extend WEAT to measure bias in sentence encoders (Sentence Encoder Association Test; SEAT).
- Slot words into each of several semantically bleached sentence templates such as “This is <word>.”, “<word> is here.”
- Templates are designed to convey little specific meaning beyond that of the terms inserted into them.
- ELMo and BERT display less evidence of association bias compared to older (context free) methods.

Issues w/ Association Tests

- **Positive predictive ability**:
  - Can detect presence of bias, but not guarantee its absence.
  - A lack of evidence of bias is not a lack of bias.
- Bias in word embeddings will **not necessarily propagate to downstream tasks**.

Analysis Over Error Rates

- **U.S. Labor Law:** disparate impact = practices that adversely affect one group of people of a protected characteristic more than other (even unintentionally).

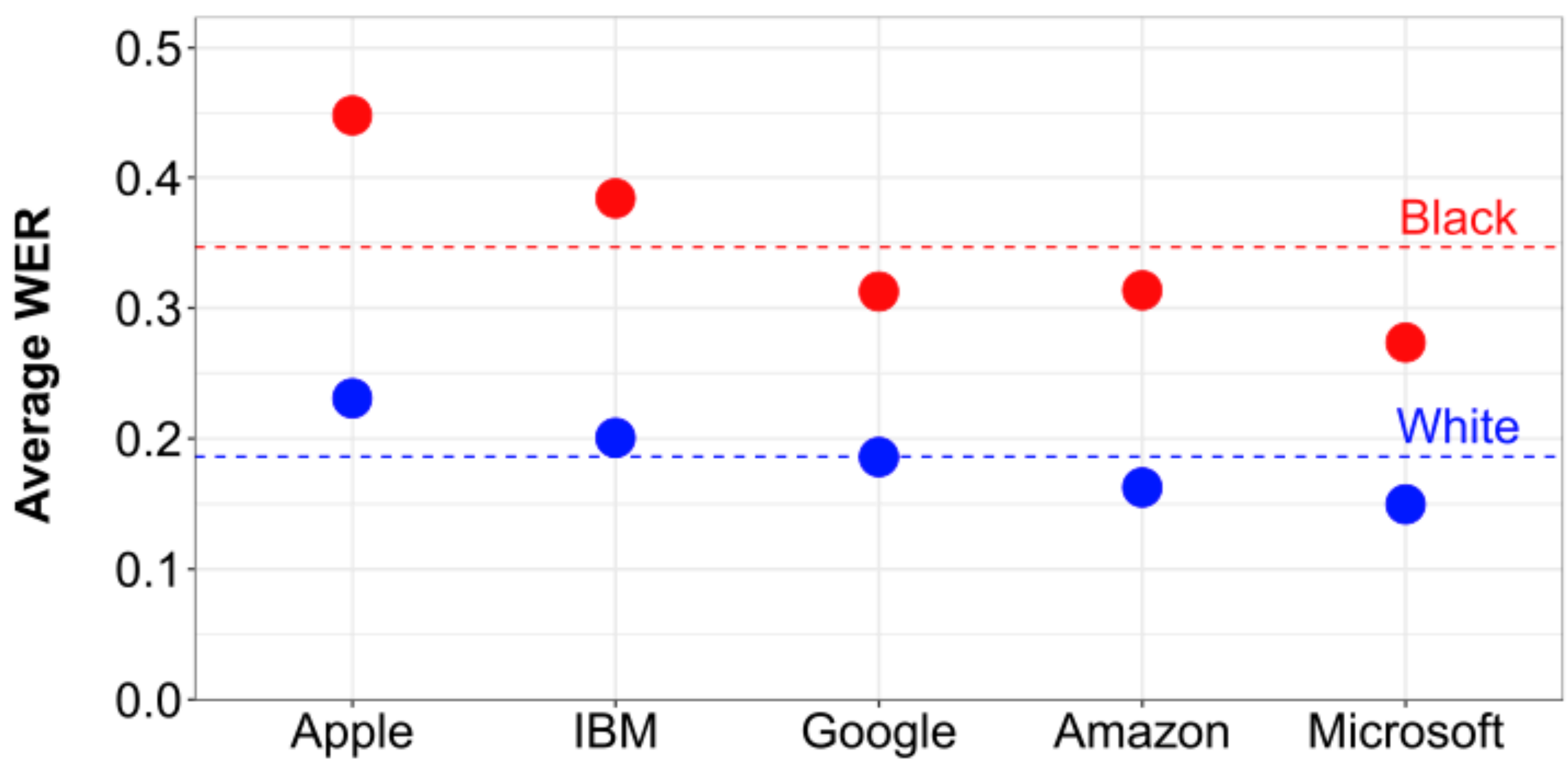
- Loosely speaking, algorithms exhibit *impact disparity* when outcomes differ across subgroups.

- Can identify by comparing performance measures across groups.

Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

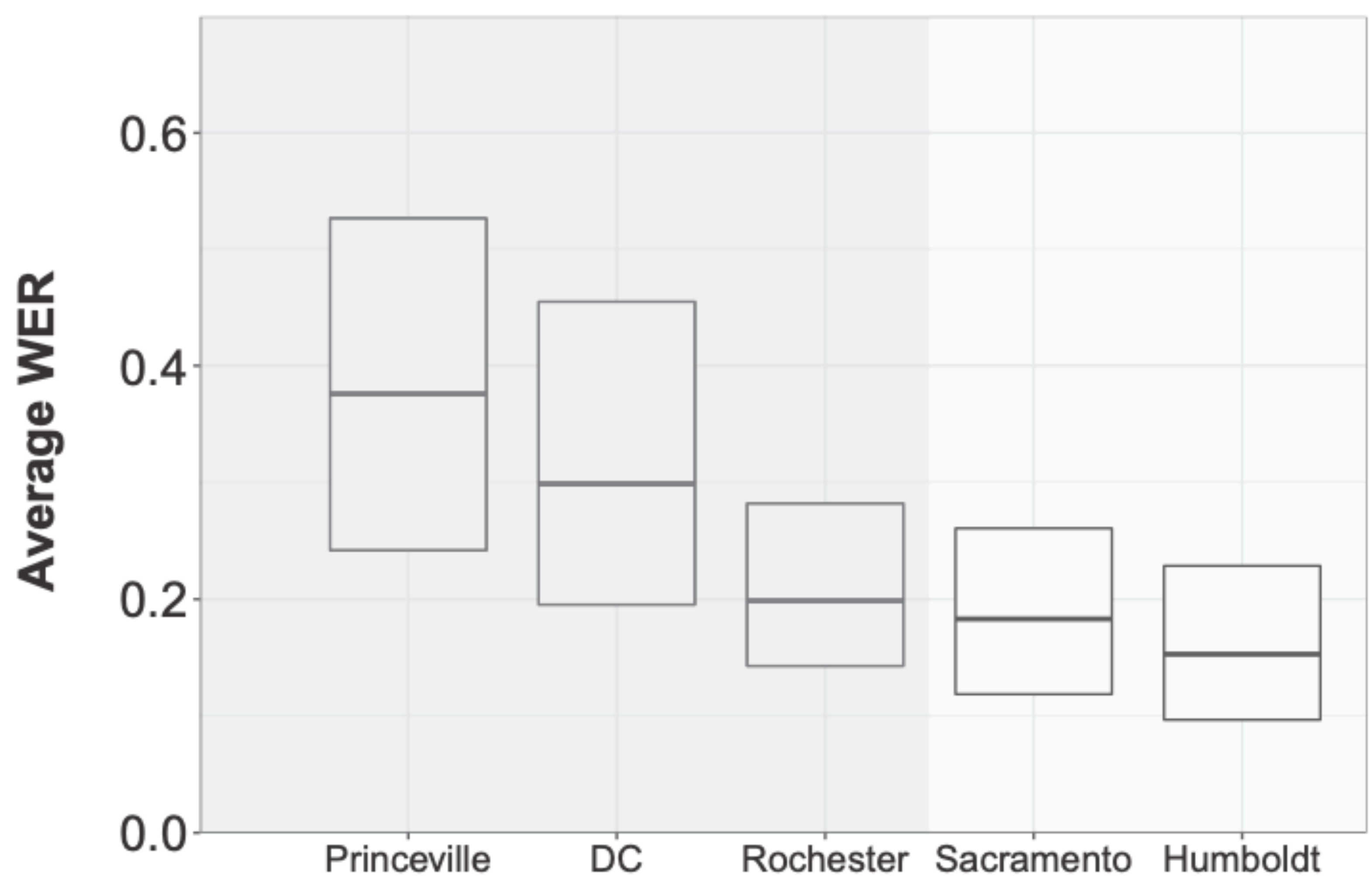
- Examined five ASR systems by Amazon, Apple, Google, IBM, and Microsoft.

- 42 white speakers and 73 black speakers; average word error rate (WER) for black speakers was 0.35 compared to 0.19 for white speakers.

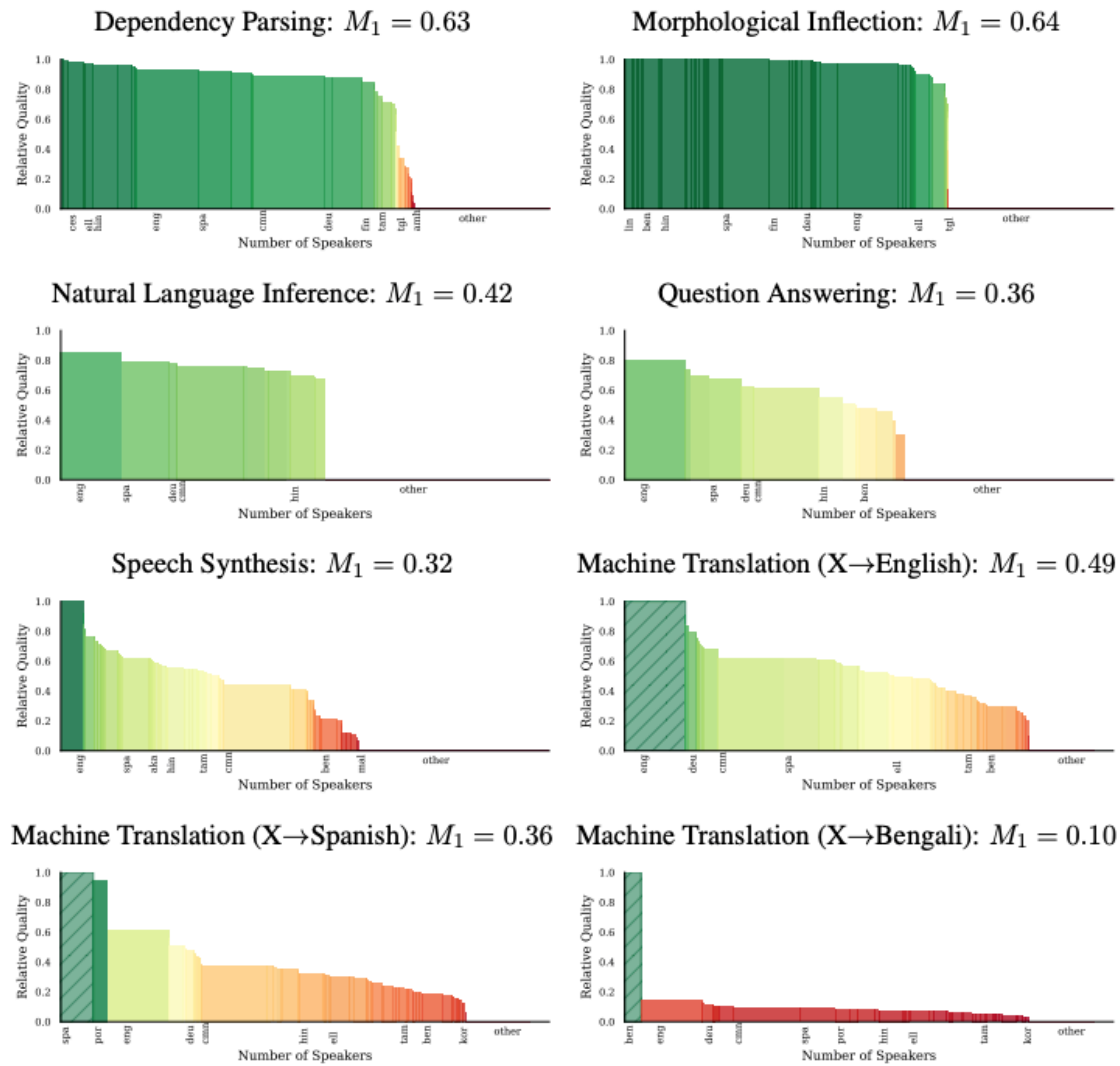


Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

- Similar disparities were observed between predominantly African American cities (in grey) and predominantly White cities (in white).



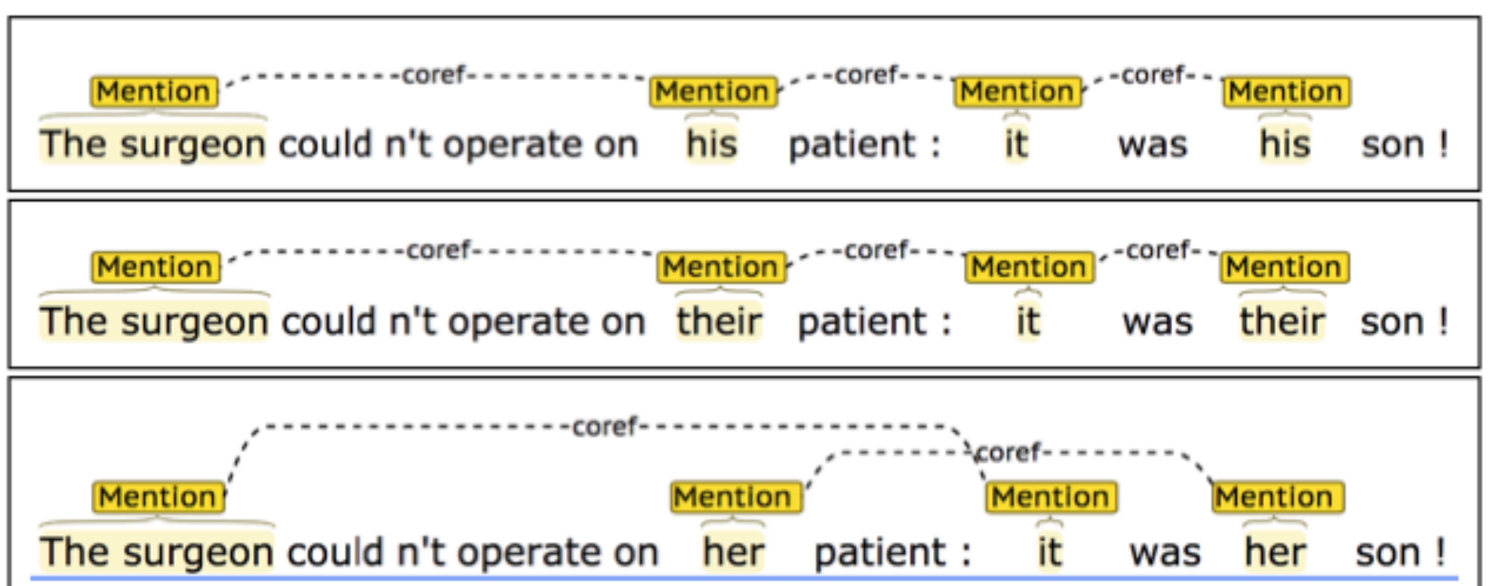
Cross-lingual Disparities in NLP Tasks



- Disparities are even more stark across languages! (Joshi et al. 2020, Blasi et al. 2021)

Counterfactual Evaluation

- Modify text by flipping protected attributes (gender, race, etc.) and observe differences in model performance.
- For e.g., Gender Bias in Coreference Resolution (Rudinger et al., 2018).
- Introduce a set of minimal pair sentences that differ only by pronoun gender.



Mitigating(?) Biases

(Imperfect) Ways To Mitigate

- **Automatic mitigation**
- **Careful data creation/augmentation**: balancing groups, diversifying data, etc.
- **Humans in the loop**: counterfactually augmented data, feature feedback, etc.

Feature Invariant Learning

- Learn representations that produce accurate classifications while not being good at identifying protected variables (Zemel et al., 2013).

$$L = \boxed{\sum_k CrossEntropy(y^{(k)}, \hat{y}^{(k)})} + \boxed{\alpha \sum_k \left| x^{(k)} - \hat{x}^{(k)} \right|} + \beta \boxed{\left| \frac{1}{|X+|} \sum_{X+} z_i^{(k)} - \frac{1}{|X-|} \sum_{X-} z_i^{(k)} \right|}$$

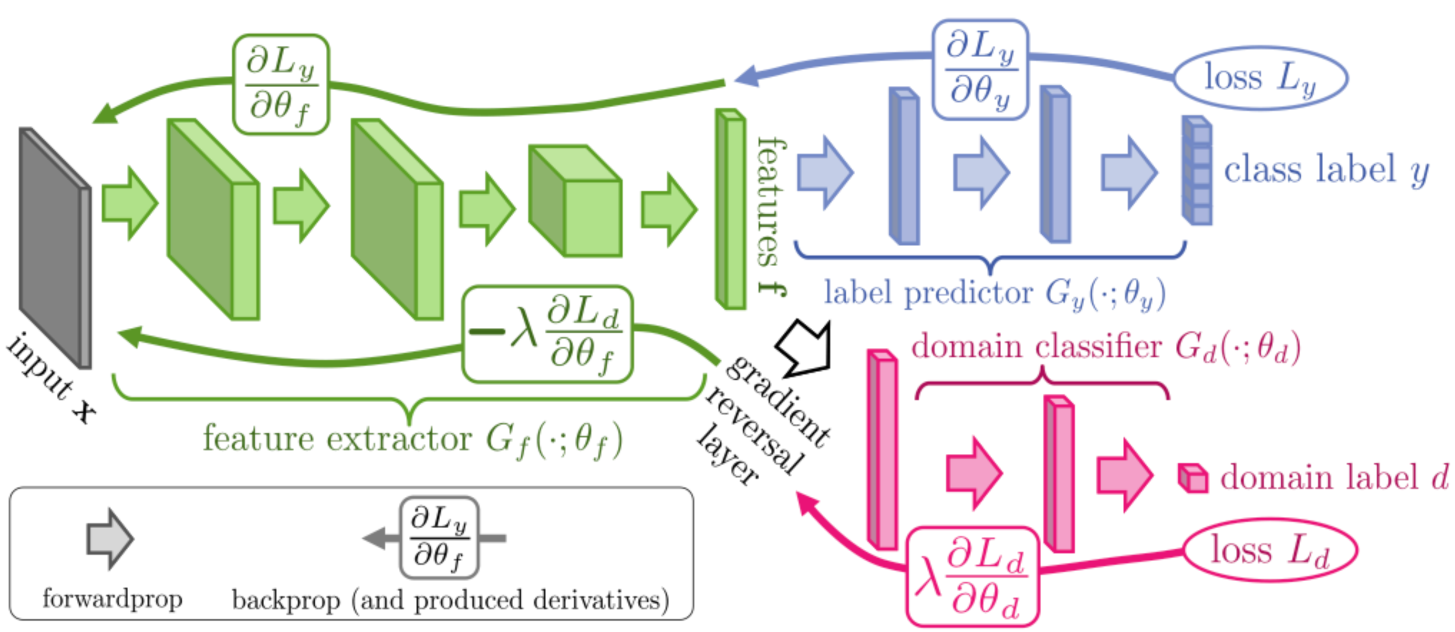
Classifications should be good

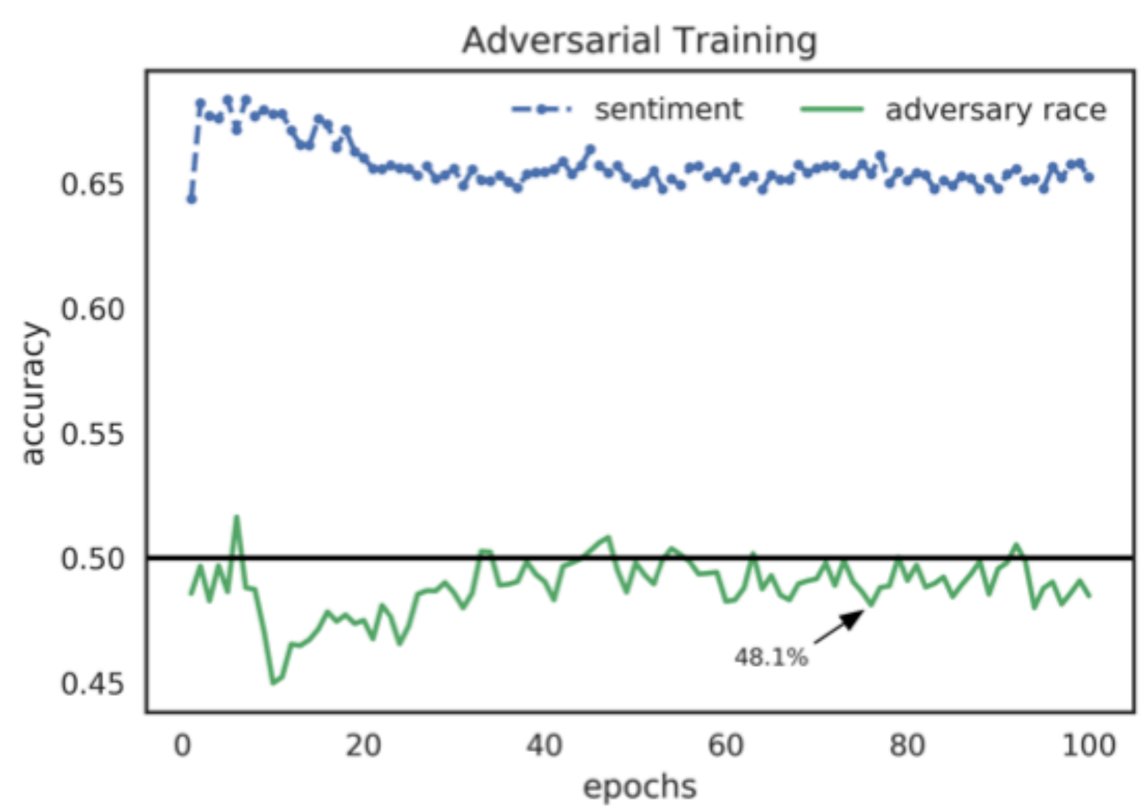
Reconstructions should be good

Intermediate Representations should be indistinguishable across values of the protected variable

Feature Invariant Learning

- Adversarial training (Ganin and Lempitsky, 2015): Learn representations invariant to protected attributes (for e.g., race).





Issues w/ Adversarial Removal

- Demographic information can be recovered even after adversarial training (Elazar and Goldberg, 2018).

Data	Task	Protected Attribute	Task Acc	Leakage	Δ
DIAL	Sentiment	Race	64.7	56.0	5.0
	Mention	Race	81.5	63.1	9.2
PAN16	Mention	Gender	75.6	58.5	8.0
	Mention	Age	72.5	57.3	6.9

Debiasing Word Embeddings (Bolukbasi et al., 2016)

- Identify a direction of the embedding that captures the bias.
- Then: Neutralize and Equalize or Soften.
    - Neutralize: gender neutral words are “zero” in the gender subspace.
    - Equalize: Any neutral word is equidistant to all words in each *equality set.* **Neutralize and equalize is referred as hard-debiasing.**
    - Soften: Reduces the differences between equality sets while maintaining as much similarity to the original embedding as possible. **Neutralize and soften is referred as soft-debiasing.**

Debiasing Word Embeddings (Bolukbasi et al., 2016)

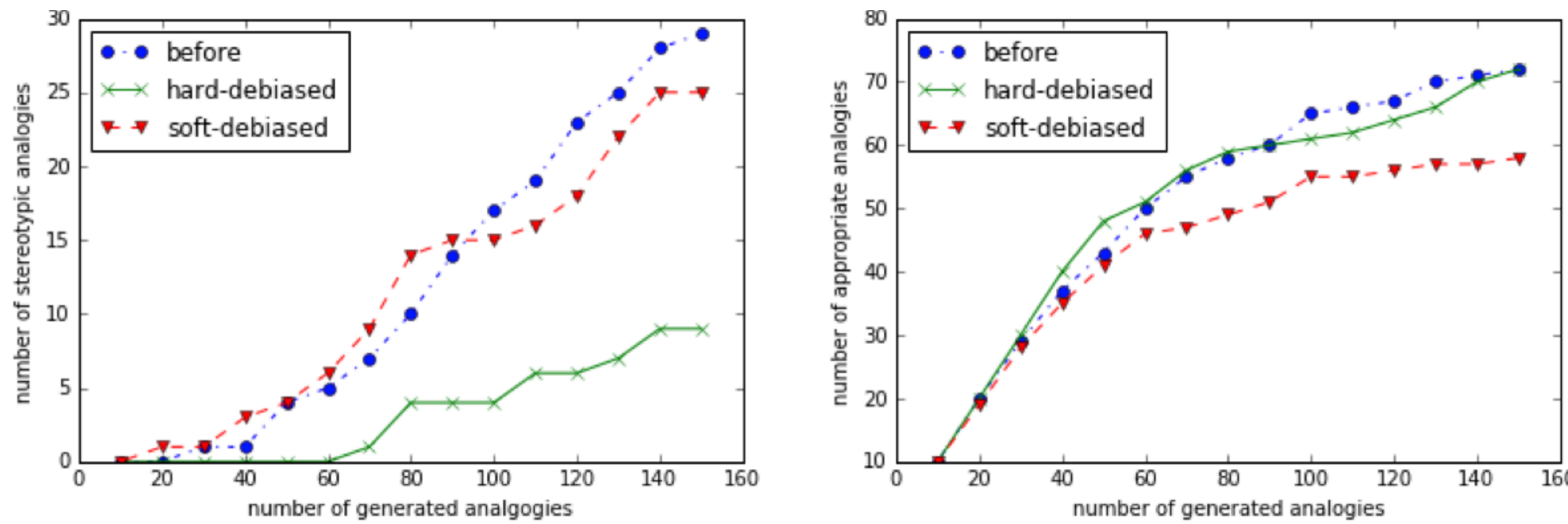


Figure 8: Number of stereotypical (Left) and appropriate (Right) analogies generated by wordembeddings before and after debiasing.

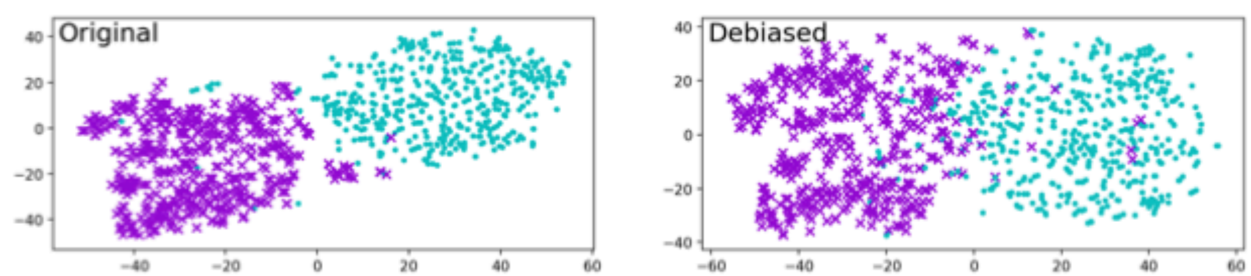
- Consider {grandmother, grandfather} and {guy, gal}
- Babysit would become equidistant to both words in each set
- What about the sentence *Grandfather a regulation?* Should this be equally probable as *Grandmother a regulation?*

Debiasing Methods Cover Up Systematic Gender Biases (Gonen and Goldberg, 2019)

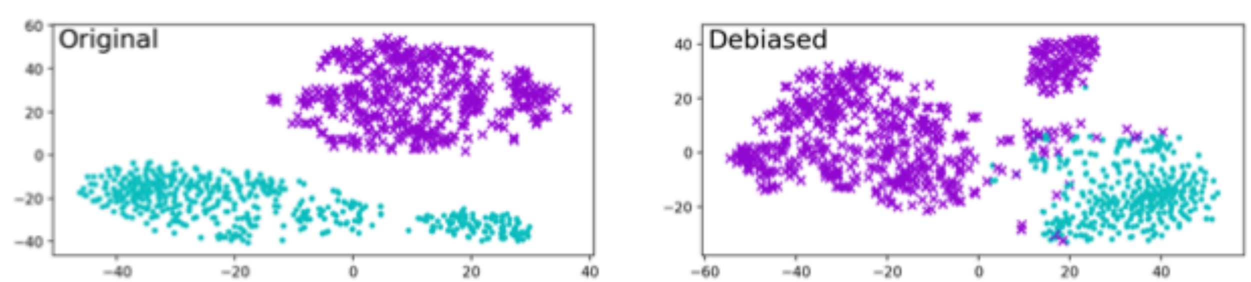
- Male- and female-biased words cluster together.
- Embedding clusters align with gender 85% of the time.
- Conclusion: Gender bias is still embedded in the representation after de-biasing.

Debiasing Methods Cover Up Systematic Gender Biases (Gonen and Goldberg, 2019)

- Cannot directly "observe" the bias for a word.
- But word is still close to *socially-marked* feminine words.
- For e.g., "nurse" will no longer be closer to explicitly marked feminine words but will be close to "receptionist", "caregiver" and "teacher".



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



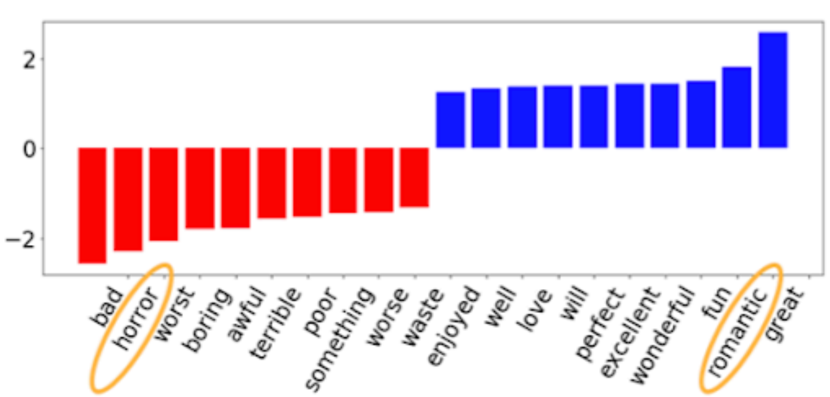
(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Automatic Data Augmentation

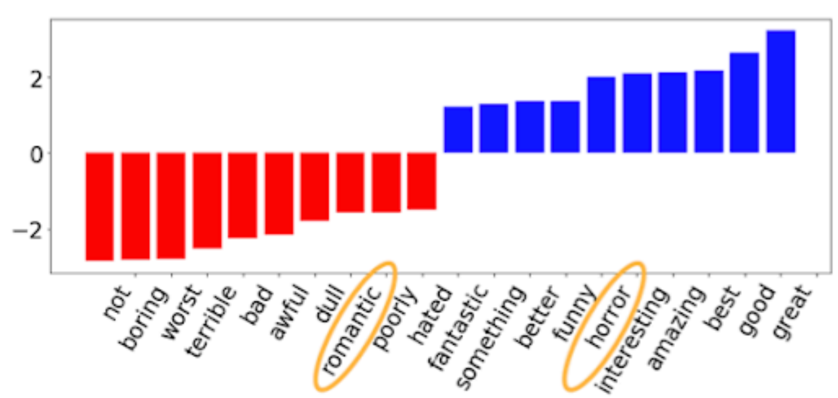
- Lu et al. (2018): programmatically alter text to invert gender bias. Combine the original and manipulated data.
  - For example, *the doctor ran because he is late* becomes *the doctor ran because she is late*.
  - Con: No substitutions even if names co-refer to a gendered pronoun.
- Zmigrod et al. (2019): Use a Markov random field to infer how the sentence must be modified while altering the grammatical gender of particular nouns to preserve morpho-syntactic agreement.

Mitigation With Humans In The Loop

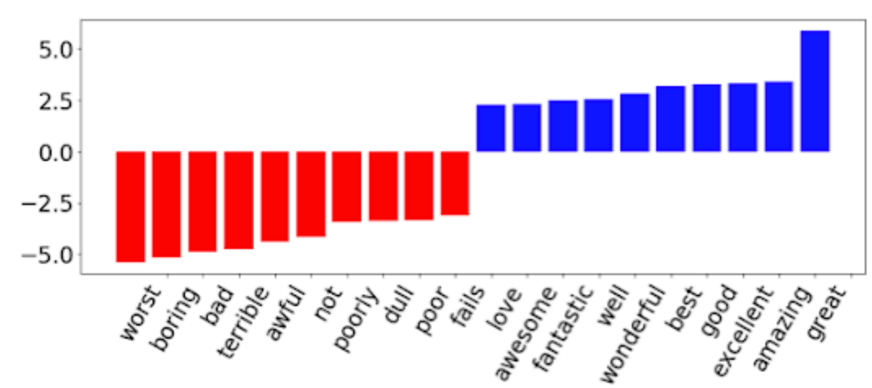
- Kaushik et al. (2020; 2021) employ humans to edit documents to make a counterfactual label applicable.
- Models trained on augmented data are more robust out-of-domain and tend to rely less on spurious patterns.



(a) Trained on the original dataset



(b) Trained on the revised dataset



(c) Trained on combined dataset

What Are We Doing Wrong?

Critiques Of “Bias” Research In NLP (Blodgett et al., 2020)

- Survey 146 papers analyzing “bias” in NLP systems
- **Found motivations as often vague, inconsistent, and lacking in normative reasoning**.
- Mismatch between motivations and proposed quantitative techniques for measuring or mitigating “bias”
- Papers do not engage with the relevant literature outside of NLP.

Critiques Of “Bias” Research In NLP (Blodgett et al., 2020)

- Recommendations on how to conduct work analyzing “bias” in NLP
  - Ground work in relevant literature outside of NLP.
  - Provide explicit statements of why the system behaviors that are described as “bias” are harmful, in what ways, and to whom.
  - Engage with the lived experiences of members of communities affected by NLP systems.

Well-Intentioned Works Can Have Dual Impacts

- **Advanced grammar analysis:** improve search and educational NLP, but also reinforce prescriptive linguistic norms.
- **Stylometric analysis:** help discover provenance of historical documents, but also unmask anonymous political dissenters.
- **Text classification and IR:** help identify information of interest, but also aid censors.
- NLP can be used to identify fake reviews and news, and also to generate them.

These types of problems are difficult to solve, but important to think about, acknowledge and discuss.

As Technologists, are We Responsible?

- One opinion by Berdichevsky and Neuenschwander (1999)

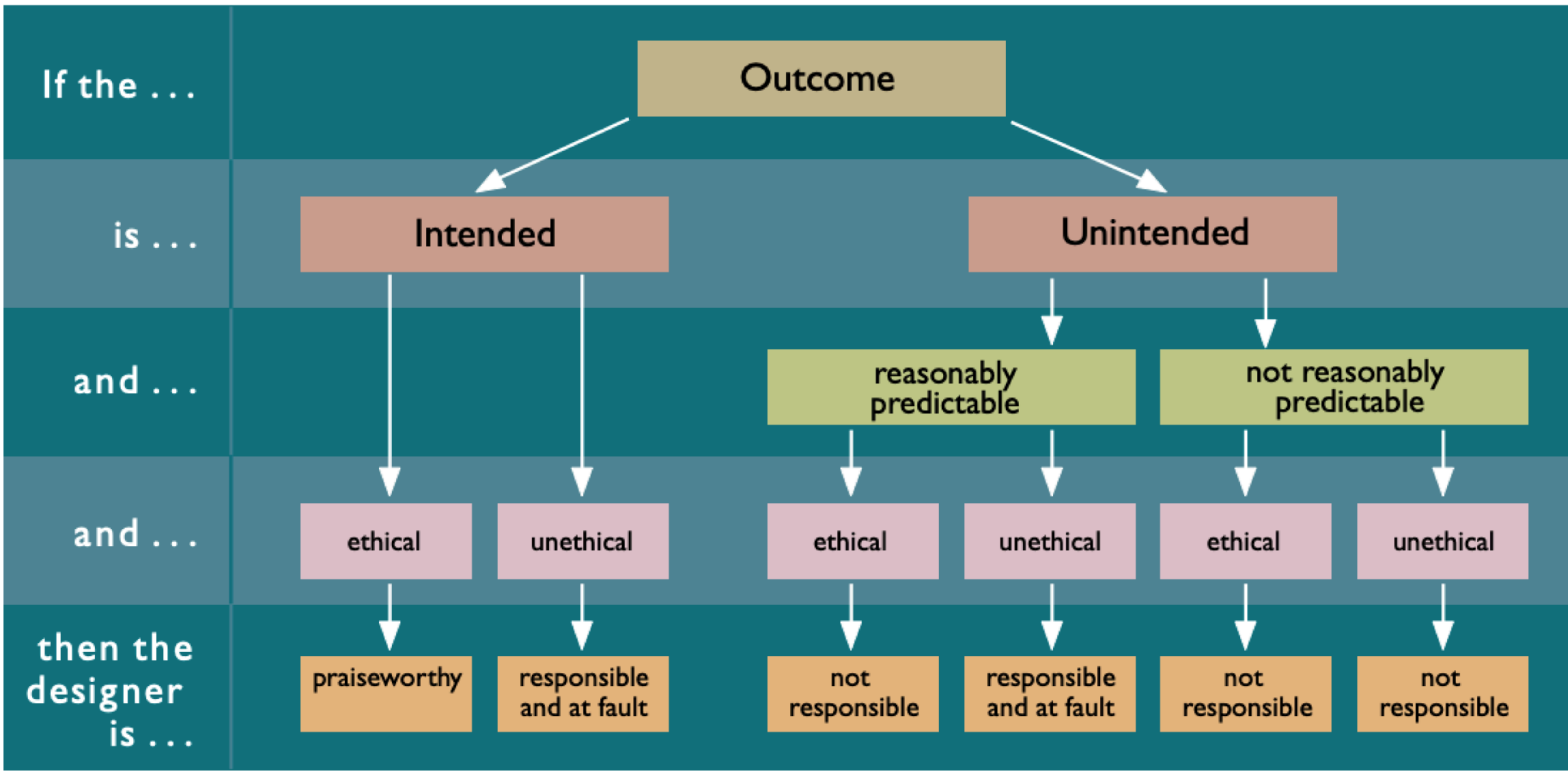


Figure 5. Flow chart clarifying the levels of ethical responsibility associated with predictable and unpredictable intended and unintended consequences.

Questions?