Week 5

MORPH Algorithmic Fairness

Catherine Yeo

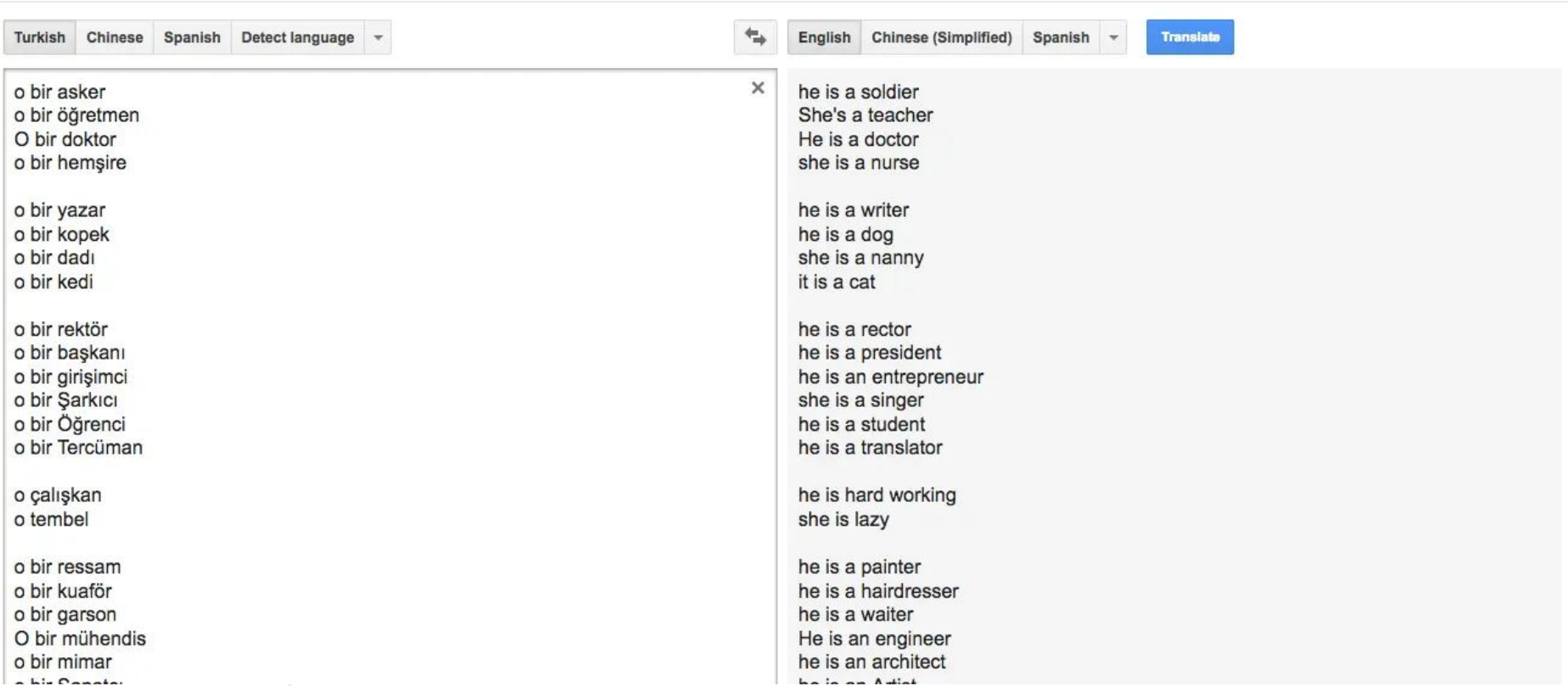
Agenda:
Applying Fairness to Algorithms II

- This Week in Fairness
- Fairness in NLP coding demo
- Fairness in Computer Vision
  - Discussion
- Week 6 preview

This Week in Fairness

Google

- Spoke at a conference recently about how they address ML fairness in their products
- Google Translate has a huge impact
  - ~50% of the content on the internet is in English, but only 20% of the world speaks English
  - “Google translates **140 billion words** every single day by **150 billion active users**, including 95% outside the U.S.”



Google Translate (2017)

www.theverge.com › google-translate-ai-machine-learn...

Google's head of translation talks fighting bias and why AI ...

Jan 30, 2019 - But as the head of **Google Translate**, Macduff Hughes, told The Verge recently, machine learning is what makes Google's ever-useful translation ...

blog.google › products › reducing-gender-bias-google-...

Reducing gender bias in Google Translate - The Keyword

Dec 6, 2018 - **Google Translate** learns from hundreds of millions of already-translated examples from the web. Historically, it has provided only one translation ...

slator.com › machine-translation › google-fixes-gender...

Google Fixes Gender Bias in Google Translate (Again) | Slator

Apr 29, 2020 - But **Google Translate** is now back with a new fix for gender **bias**, which it said can produce "gender-specific translations with an average precision ...

translate.google.com › translate_t

Google Translate

No information is available for this page.
Learn why

venturebeat.com › 2020/04/22 › google-debuts-ai-in-g...

Google debuts AI in Google Translate that addresses gender ...

Apr 22, 2020 - Evaluated on a Google-developed metric called bias reduction, which measures the relative reduction of bias between the new translation system and the existing system (where "bias" is defined as making a **gender** choice in translation that's unspecified in the source), Johnson says the new approach results in a bias ...

techcrunch.com › 2018/12/07 › google-translate-gets-ri...

Google Translate gets rid of some gender biases | TechCrunch

Dec 7, 2018 - **Google** recently made some important changes to its **Translate** tool — reducing gender **bias** by providing both masculine and feminine ...

ai.googleblog.com › 2020/04 › a-scalable-approach-to-...

A Scalable Approach to Reducing Gender Bias in Google ...

Apr 22, 2020 - Here "bias" is defined as making a **gender** choice in the translation that is unspecified in the source. For example, if the current system is biased 90% of the time and the new system is biased 45% of the time, this results in a 50% relative bias reduction.

medium.com › babbel › google-translate-addresses-its-...

Google Translate addresses its bias issue | by Thomas Moore ...

Dec 7, 2018 - This past week, **Google** unveiled a redesign of its very popular **Translate** service. The most noticeable changes were in appearance, but one of ...

support.google.com › translate › thread

Bias in translation by Google Translate - Google Support

Feb 6, 2020 - There is hidden **bias** of the translations proposed by your dictionary. For example, the dictionary translates feminative Polish word "grubsza" as ...

Look at the dates!

Google Translate

- They collect data from many historical sources, such as the Bible
- Historical literature has many stereotypes/biases embedded in their data (as we've seen with the Bolukbasi paper)

How to resolve?

Flip a coin

Decide based on what users select or how they react to a translation

Provide multiple responses

How to resolve?

Flip a coin

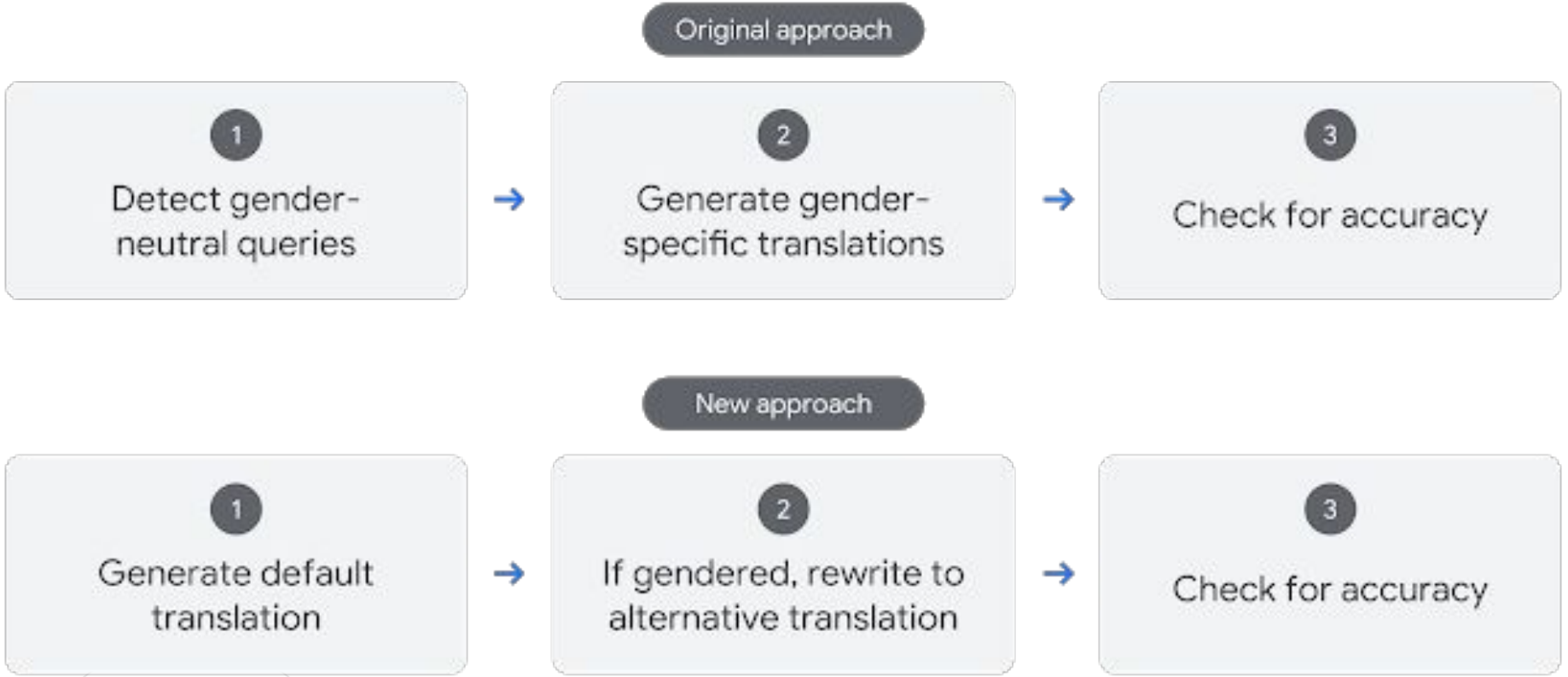
Decide based on what users select or how they react to a translation

Provide multiple responses

Before

After

Google had to create 3 new models for this 2018 approach



Updated 2020 approach

Fairness in NLP Coding Demo

Debiasing Word Embeddings: Part 1

Bolukbaski et al. “Man is to Computer Programmer as Woman is to Homemaker.” (2016)

Main idea: Word embeddings embed sexism. In fact, we can identify the gender subspace g.

Debiasing Word Embeddings: Part 1

Finding *g*, the gender subspace:

$$\overrightarrow{\text{grandmother}} - \overrightarrow{\text{grandfather}} = \overrightarrow{\text{gal}} - \overrightarrow{\text{guy}} = g$$

Use *g* to identify bias of embeddings: cos(v, g) (or, equivalently, the dot product)

- Project word vectors onto gender dimension to get a quantitative bias score

Bolukbasi et al. (2016)

Bolukbasi et al. (2016)

Debiasing Word Embeddings: Part 1

- Proposed debiasing methods (hard and soft) essentially subtracts gender direction from gender-neutral words to remove bias
- Happens in the postprocessing step
- After debiasing, these analogies should have a lower bias score

Bolukbasi et al. (2016)

Google Colab (Jupyter notebook):

Link sent in Zoom

Google Colab (Jupyter notebook):

https://colab.research.google.com/drive/1D2zBeDkhro9-ncukcb48FXsNp9GQChEe?usp=sharing

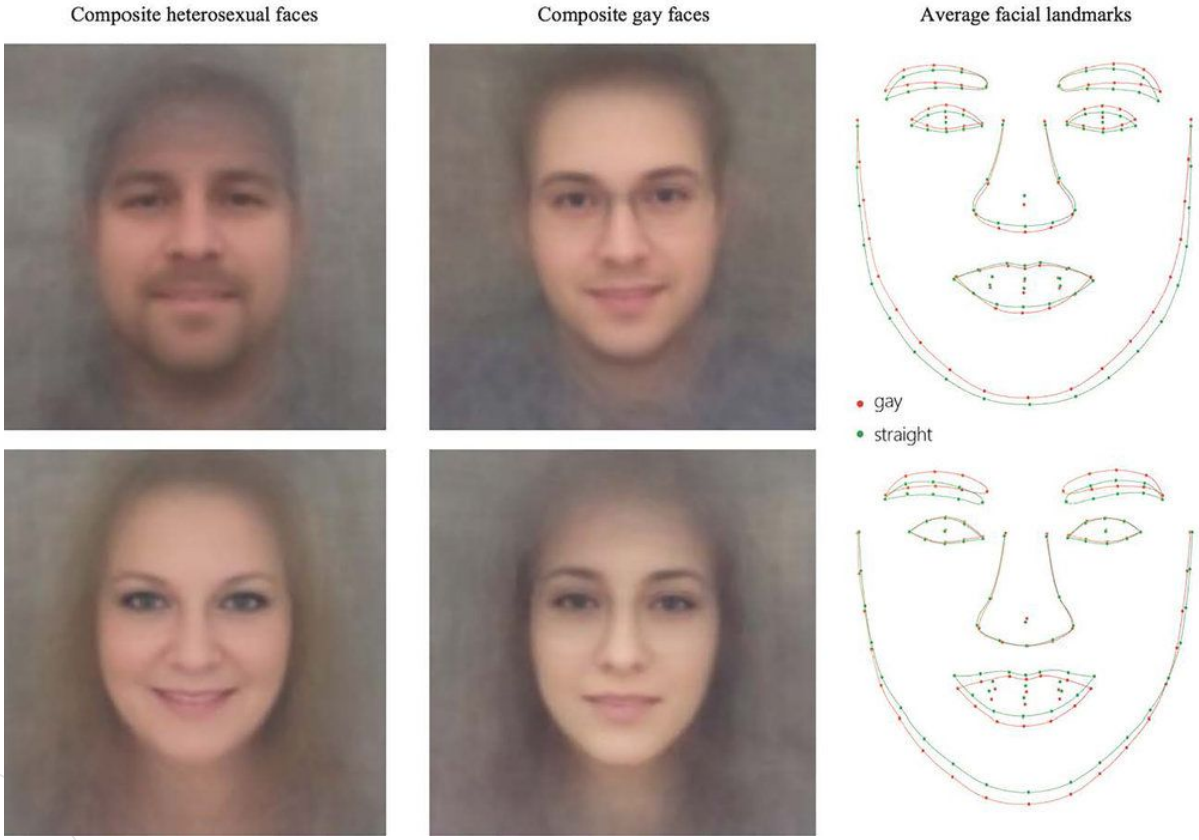
Fairness in Computer Vision

Facial Analysis

- Face recognition softwares are rampant
  - FaceID in iPhones
  - Surveillance cameras
  - Affectiva (founded 2016), born out of MIT Media Lab, identifies emotions from images of faces
  - Research determining sexuality of white male based on Facebook and dating sites photos (Kosinski & Wang, 2017)


DISTRACTION 30
ANGER 100
MOUTH OPEN 100
EYE CLOSURE 0
JOY 0
:) Affectiva

Affectiva



"Why Stanford Researchers Tried to Create a 'Gaydar' Machine" — NYT link

“This Person Does Not Exist” — AI generated faces
https://thispersondoesnotexist.com/

Gender Shades

- Compared facial recognition systems across different gender and skin tones
- Intersectional benchmark: dark/light — male/female pairings
- Used existing metrics of skin type comparison and labeling

(Buolamwini and Gebru, 2018)

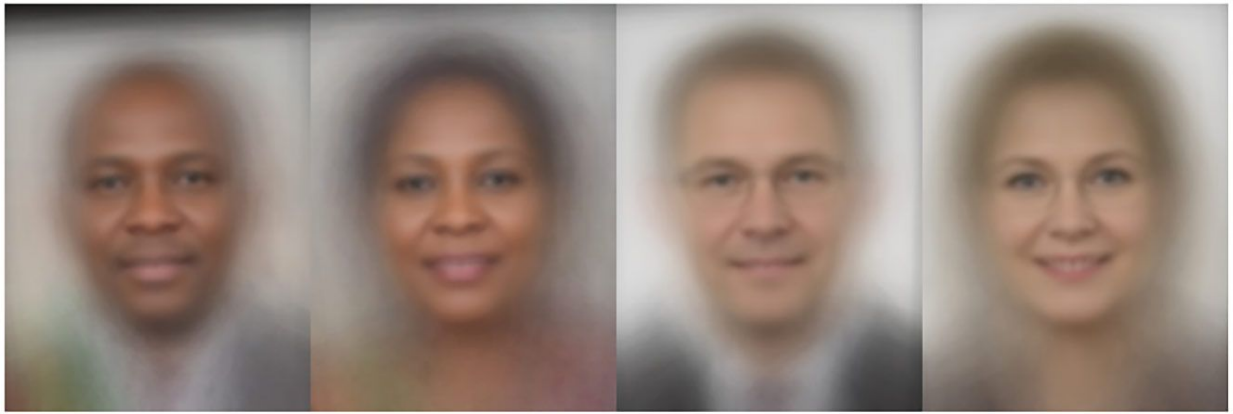
Buolamwini analyzed 1000+ faces of different genders and skin tones with 3 facial recognition systems

Results

- All classifiers performed better on **male** faces than **female** faces
- All classifiers performed better on **lighter** faces than **darker** faces
- All classifiers performed worst on **darker female** faces
- Max error rate for darker females = 34%
- Max error rate for lighter males = < 1 %

(Buolamwini and Gebru, 2018)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gender Shades Project

http://gendershades.org/overview.html

Discussion

- Society should never completely halt innovation. But where do we draw the line between innovation and dangers?
- Many companies sell their facial recognition to law enforcement. Should the police use facial recognition AIs?
  - Is it ethical for them to?
  - Even if it is “unbiased”, should law enforcement use them?
  - Helping keep society more secure vs. Protecting privacy?