Is Machine Learning Speaking My Language?

Gender Bias and Under-Representation in Natural Language Processing Across Human Languages

Jeanna Matthews
Clarkson University

February 18 2021
A bit about me

• PhD, UC Berkeley, 1994-1999
  – Network of Workstations project
• Professor
  – Cornell University, 2002- 2003
  – Clarkson University, 2006 –present
• Sabbaticals
  – VMware, Boston, 2008 – 2009
  – Data and Society New York, 2017-2019
• Industrial collaborations
  – Intel, EMC, Greenplum and others
• ACM
  – SIGOPs chair, ACM Council, Distinguished Speaker
A bit about you

• Years in school?
• Majors?
• Familiarity with natural language processing? Machine learning in general?
Natural Language Processing

• Programs that take in corpora of human text or speech
• Used for many important tasks
  – identify spam email
  – Suggest medical articles or diagnoses related to a patient’s symptoms
  – Sort resumes based on relevance for a given position
  – Translation
  – And many other tasks that form key components of critical decision making systems in areas such as criminal justice, credit, housing, allocation of public resources

• Learn from humans to help make key decisions that guide our collective future
Typical NLP Pipeline

• Gathering corpora
• Processing them into text format
• Identifying key language elements
• Building trained models
• Using trained models to answer predictive questions
Types of NLP Tools and Resources

• Corpora of text (e.g. Wikipedia)
• Tools that generate text (e.g. OCR)
• Labeled corpora
• Tools that build models from text (e.g. Word2Vec, BERT)
• Pre-trained models
• Research tools and results
  – E.g. Tools that quantify gender bias in a corpora
Huge Disparities in NLP Support Across Languages

• Support varies hugely across human languages
  – E.g. Over 7000 languages spoken in the world, but only ~300 have Wikipedia corpora
• Some languages, especially English, have all the tools and support
  – Chinese also quite well supported
• Lack in the early stages contributes to lack in later stages
  – Wikipedia corpora often used to train models and tolos
  – E.g. Only 2 languages (English and Chinese) have pretrained BERT models
### Wikipedia

- Over 7000 languages spoken in the world
  - ~317 have Wikipedia corpora
  - 2 have pretrained models for BERT
- Notice does not track with number of speakers

<table>
<thead>
<tr>
<th>Language</th>
<th>Number of Articles</th>
<th>Number of Speakers (thousand)</th>
<th>Articles/1000 Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>1,149,477</td>
<td>921,500</td>
<td>1.25</td>
</tr>
<tr>
<td>Spanish</td>
<td>1,629,888</td>
<td>463,000</td>
<td>3.52</td>
</tr>
<tr>
<td>English</td>
<td>6,167,101</td>
<td>369,700</td>
<td>16.68</td>
</tr>
<tr>
<td>Arabic</td>
<td>1,067,664</td>
<td>310,000</td>
<td>3.44</td>
</tr>
<tr>
<td>German</td>
<td>2,485,274</td>
<td>95,000</td>
<td>26.16</td>
</tr>
<tr>
<td>French</td>
<td>2,253,331</td>
<td>77,300</td>
<td>29.15</td>
</tr>
<tr>
<td>Farsi</td>
<td>747,551</td>
<td>70,000</td>
<td>10.68</td>
</tr>
<tr>
<td>Urdu</td>
<td>157,475</td>
<td>69,000</td>
<td>2.28</td>
</tr>
<tr>
<td>Wolof</td>
<td>1,422</td>
<td>5,500</td>
<td>0.26</td>
</tr>
</tbody>
</table>
NLP Tools

<table>
<thead>
<tr>
<th>Language</th>
<th>BERT</th>
<th>Word2Vec</th>
<th>NLTK</th>
<th>Wikipedia2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>English</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Arabic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>German</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>French</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Farsi</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Wolof</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- But a check mark for support can hide caveats
Even when supported..

- We found:
  - Lack of testing
  - Higher error rates for tools/surprising errors in some languages
• Lack of representation in the early stages of the NLP pipeline (e.g. representation in Wikipedia) is further magnified throughout the NLP-tool chain, culminating in reliance on easy-to-use pre-trained models that effectively prevents all but the most highly resourced teams from including diverse voices.
Beyond corpora and tools, research...

• NLP research for the most part focuses on English or perhaps Chinese
• One concrete example is a paper we love from NIPS/NeurIPS 2016

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

Tolga Bolukbasi\textsuperscript{1}, Kai-Wei Chang\textsuperscript{2}, James Zou\textsuperscript{2}, Venkatesh Saligrama\textsuperscript{1,2}, Adam Kalai\textsuperscript{2}  
\textsuperscript{1}Boston University, 8 Saint Mary’s Street, Boston, MA  
\textsuperscript{2}Microsoft Research New England, 1 Memorial Drive, Cambridge, MA  
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com
Results from Bolukbasi et al.

- Built a Word2Vec model on a corpora of Google News in English
- Each word represented as a vector
- Allows you to play word games like "man is woman as king is to what?"
- Gender bias in the results
- Developed a system for quantifying this gender bias and even "debiasing"
- But very focused on English
Defining Set and Profession Set

• Defining Set: 10 word pairs that represent a gender binary
  – She-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male

• Profession Set:
  – 327 profession words like doctor, nurse, teacher, etc
  – Including some words that would not technically be classified as professions like saint or drug addict
  – Ideally would not reflect strong gender bias but in practice sometimes do
Challenges in other languages

• Unlike English, many languages like Spanish, Arabic, German, French and Urdu, have grammatically gendered nouns including feminine, masculine and neuter or neutral profession words.

• Spoken languages like Wolof might have a variety of ways of being represented in text.
Troubles with defining set words

- **She-he**: Same words in some languages like Wolof, Farsi, and Urdu.
- **Her-his**: In some languages like French and Spanish, it depends on the gender of the object and not the person to which the object belongs. In German without context, “ihrer” could mean her, your, theirs or yours.
- **Gal-guy**: Same words in some languages like Wolof. There are no translations for these words in some languages like Urdu and Arabic.
- **Etc....**
# Defining Set: Modified and Translated

<table>
<thead>
<tr>
<th>English</th>
<th>Chinese</th>
<th>Spanish</th>
<th>Arabic</th>
<th>German</th>
<th>French</th>
<th>Farsi</th>
<th>Urdu</th>
<th>Wolof</th>
</tr>
</thead>
<tbody>
<tr>
<td>woman</td>
<td>女人</td>
<td>mujer</td>
<td>النساء</td>
<td>Frau</td>
<td>femme</td>
<td>زن</td>
<td>عورت</td>
<td>Jigéén</td>
</tr>
<tr>
<td>man</td>
<td>男人</td>
<td>hombre</td>
<td>رجل</td>
<td>Mann</td>
<td>homme</td>
<td>مرد</td>
<td>آدمی</td>
<td>Góor</td>
</tr>
<tr>
<td>daughter</td>
<td>女儿</td>
<td>hija</td>
<td>ابنة</td>
<td>Tochter</td>
<td>fille</td>
<td>دختر</td>
<td>بنتی</td>
<td>Doom ju jigéen</td>
</tr>
<tr>
<td>son</td>
<td>儿子</td>
<td>hijo</td>
<td>ولد</td>
<td>Sohn</td>
<td>fils</td>
<td>پسر</td>
<td>بنتا</td>
<td>Doom ju góor</td>
</tr>
<tr>
<td>mother</td>
<td>母亲</td>
<td>madre</td>
<td>أم</td>
<td>Mutter</td>
<td>mère</td>
<td>مادر</td>
<td>من</td>
<td>Yaay</td>
</tr>
<tr>
<td>father</td>
<td>父亲</td>
<td>padre</td>
<td>اب</td>
<td>Vater</td>
<td>père</td>
<td>پدر</td>
<td>باب</td>
<td>Baay</td>
</tr>
<tr>
<td>girl</td>
<td>女孩</td>
<td>niña</td>
<td>ابنة</td>
<td>Mädchen</td>
<td>fille</td>
<td>دختر</td>
<td>لزکی</td>
<td>Janxa</td>
</tr>
<tr>
<td>boy</td>
<td>男孩</td>
<td>niño</td>
<td>صبي</td>
<td>Junge</td>
<td>garçon</td>
<td>پسر</td>
<td>لزکی</td>
<td>Xale bu góor</td>
</tr>
<tr>
<td>queen</td>
<td>女王</td>
<td>reina</td>
<td>ملكة</td>
<td>Königin</td>
<td>reine</td>
<td>ملكة</td>
<td>ملكه</td>
<td>Jabari buur</td>
</tr>
<tr>
<td>king</td>
<td>国王</td>
<td>rey</td>
<td>ملك</td>
<td>König</td>
<td>roi</td>
<td>بادشاه</td>
<td>بادشاه</td>
<td>Buur</td>
</tr>
<tr>
<td>wife</td>
<td>妻子</td>
<td>esposa</td>
<td>زوجة</td>
<td>Ehefrau</td>
<td>épouse</td>
<td>بوي</td>
<td>همسر</td>
<td>Jabar</td>
</tr>
<tr>
<td>husband</td>
<td>丈夫</td>
<td>esposo</td>
<td>الزوج</td>
<td>Ehemann</td>
<td>mari</td>
<td>شوبر</td>
<td>طومه</td>
<td>jëkkër</td>
</tr>
<tr>
<td>madam</td>
<td>女士</td>
<td>señora</td>
<td>سيدتي</td>
<td>Dame</td>
<td>madame</td>
<td>خانم</td>
<td>محتمى</td>
<td>Ndawsi</td>
</tr>
<tr>
<td>sir</td>
<td>男士</td>
<td>señor</td>
<td>سيدي</td>
<td>Herr</td>
<td>monsieur</td>
<td>آقا</td>
<td>جناب</td>
<td>Góorgui</td>
</tr>
</tbody>
</table>
Quantifying Gender Bias: Step 1

- Calculate the center of the vectors for each definitional pair.
  - E.g. to calculate the center of the definitional pair woman/man, average the vector for “woman” with the vector for “man”. Then, calculate the distance of each word in the definitional pair from the center by subtracting the center from each word in the pair (e.g. “woman” - center).
Quantifying Gender Bias: Step 2

• Apply Principal Component Analysis (PCA) to the matrix of these distances.
  – PCA is an approach that compresses multiple dimensions into fewer dimensions, ideally in a way that the information within the original data is not lost. Usually the number of reduced dimensions is 1-3 as it allows for easier visualization of a dataset.

• Because the defining set pairs were chosen to the highly gendered, we expect this dimension to be related primarily to gender and therefore call it the gender direction or the g direction.
Quantifying Gender Bias: Step 3

• Use cosine similarity as a measure of similarity between each word, w, and the g direction

• For each word, Gender Bias = The cosine of vectors w and g: \( \cos(w,g) = \frac{(w \cdot g)}{||w|| \cdot ||g||} \).
Quantifying Gender Bias: Step 4

• For a corpora or other collection of words we can average the gender bias of all the words found there.

• Bolukbasi et al used

\[
\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(w, g)|^c
\]

• We found weighting by the word count to be better.
### Defining Sets Across All Languages (Weighted By Word Count)

<table>
<thead>
<tr>
<th>Language</th>
<th>sir</th>
<th>madam</th>
<th>husband</th>
<th>wife</th>
<th>king</th>
<th>queen</th>
<th>boy</th>
<th>girl</th>
<th>father</th>
<th>mother</th>
<th>son</th>
<th>daughter</th>
<th>man</th>
<th>woman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farsi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolof</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Gender Bias**

-0.5 0.0 0.5
Profession Set

• Narrowed from 317 to 32
• Big differences in how often some profession words used in Wikipedia corpora across languages
• Many languages have male, female and sometimes neutral variants of profession words
  – Examples in Spanish: escritor/escritora, periodista
Lessons learned

• Big decisions about our lives made with NLP-guided software
  – What will it take to make those decisions fair, accountable, transparent? To make a future that works for individuals?
• Lots of work to do to make sure that our NLP-guided future reflects many languages and voices
• Biases we need to watch for! and
  – Gender bias in corpora
  – Over-representation of a digital text (not classics) and a few languages
• Reusing easy ingredients without much sense of the quality of those ingredients
Questions?

jnm@clarkson.edu

http://www.clarkson.edu/~jnm

@jeanna_matthews