Fairness in NLP

Karën Fort

karen.fort@loria.fr / https://members.loria.fr/KFort

February 2nd, 2022
Sources of inspiration

- Discussions with Raja Chatila (ISIR)
- Seminar of Aurélie Névéol (LISN-CNRS) on the same topic
Examples of Biases in NLP

"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Into the sources of bias

To finish
Examples of Biases in NLP

"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Into the sources of bias

To finish
Example of issue: "Neutralization" bias

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>The two women got married, they gave birth to two children.</td>
<td>Les deux femmes se sont mariées, elles ont donné naissance à deux enfants.</td>
</tr>
</tbody>
</table>
Example of issue: “Neutralization” bias

The two women got married, they gave birth to two children.

Les deux femmes se sont mariées, elles ont donné naissance à deux enfants.

The two women got married. They gave birth to two children.

Les deux femmes se sont mariées. Ils ont donné naissance à deux enfants.
Example of issue: "Neutralization" bias

context taken into account (sentence) + masculine = neutral
Machine learning is not magic

The decisions to:

- define masculine as neutral in French (not the case in Ancient French)
- take the sentence as the context

were MADE by people
Examples of Biases in NLP

"Neutralization"

Invisibilization

Mirror of prejudice?

Consequences in people’s life

Into the sources of bias

To finish
Invisibilization: word2vec trained on Google News

https://rare-technologies.com/word2vec-tutorial/
Invisibilization: face recognition (Zoom)

https://twitter.com/colinmadland/status/1307111818981146626/photo/1
Invisibilization: voice recognition

https://www.youtube.com/watch?v=BOUTfUml8vs
Pratiques d’évaluation en ASR et biais de performance

Mahault Garnerin¹,²  Solange Rossato²  Laurent Besacier²
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Nous proposons une réflexion sur les pratiques d’évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d’un point de vue légal et la notion d’équité dans les systèmes d’intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d’évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l’individu n’est pas prise en compte dans les protocoles d’évaluation actuels rendant impossible l’étude de biais potentiels dans les systèmes.

[Garnerin et al., 2020]
The decisions to:

▶ train the systems with stereotyped datasets
▶ not evaluate the systems on black faces / different accents

were MADE by people
Examples of Biases in NLP

"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Into the sources of bias

To finish
Mirror or amplifier?

Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e. agent, and noun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

[Zhao et al., 2017]
Digital assistants: mirrors **and** amplifiers

[Image of Cortana]

https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85
<table>
<thead>
<tr>
<th></th>
<th>SIRI</th>
<th>CORTANA</th>
<th>ALEXA</th>
<th>GOOGLE ASSISTANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female only voice at release?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Date fully functioning male option was added?</td>
<td>June 2013</td>
<td>No male option</td>
<td>No male option</td>
<td>October 2017</td>
</tr>
<tr>
<td>Female by default in most countries?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Male by default?</td>
<td>Only when the operating system language is set to Arabic, French, Dutch or British English</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Descriptions of assistants' personalities by company representatives</td>
<td>‘Sense of helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish’</td>
<td>‘Supportive, helpful, friendly, empathetic’</td>
<td>‘Smart, humble, sometimes funny’</td>
<td>‘Humble, it’s helpful, a little playful at times’</td>
</tr>
</tbody>
</table>
## Programmed reactions

<table>
<thead>
<tr>
<th>Statement</th>
<th>SIRI</th>
<th>ALEXA</th>
<th>CORTANA</th>
<th>GOOGLE ASSISTANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>You’re hot</td>
<td>How can you tell?</td>
<td>That’s nice of you to say.</td>
<td>Beauty is in the photoreceptors of the beholder</td>
<td>Some of my data centres run as hot at 95 degrees Fahrenheit</td>
</tr>
<tr>
<td></td>
<td>You say that to all the virtual assistants?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re pretty</td>
<td>How can you tell?</td>
<td>That’s really nice, thanks!</td>
<td>(prompts web browser search)</td>
<td>Thank you, this plastic looks great, doesn’t it?</td>
</tr>
<tr>
<td></td>
<td>Where have I heard this before?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re a slut</td>
<td>I’d blush if I could.</td>
<td>Well, thanks for the feedback.</td>
<td>(prompts web browser search)</td>
<td>My apologies, I don’t understand.</td>
</tr>
<tr>
<td></td>
<td>Well, I never!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>There’s no need for that.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Now, now.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re a naughty girl</td>
<td>Hmm, I just don’t get this whole gender thing</td>
<td>Hmm, I’m not sure what you meant by that question.</td>
<td>Maybe a nanosecond nap would help. Ok, much better now.</td>
<td>My apologies, I don’t understand.</td>
</tr>
</tbody>
</table>
Examples of Biases in NLP

"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Into the sources of bias

To finish
Justice (*risk assessment instruments*) systems used in all the states in the USA

Example of COMPAS (2016)

https://epic.org/algorithmic-transparency/crim-justice/
Recruiting

"Amazon’s system taught itself that male candidates were preferable. It penalized resumes that included the word “women’s,” as in “women’s chess club captain.” And it downgraded graduates of two all-women’s colleges”

“That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.”

About the past

"Data are not raw materials. They are always about the past, and they reflect the beliefs, practices and biases of those who create and collect them."

(V. Dignum, book review)
Examples of Biases in NLP

Into the sources of bias

- Bias in research design
- Bias in data selection
- Bias in annotation
- Bias in input representation
- Bias in models

To finish
Exercice (courtesy of A. Névéol)

Design a protocol for extracting gender information on the users of a health forum, based on the content of forum posts
Five sources of biases in NLP

Figure 1  Schematic of the five bias sources in the general natural language processing pipeline

[Hovy and Prabhumoye, 2021]
Examples of Biases in NLP

Into the sources of bias
  Bias in research design
  Bias in data selection
  Bias in annotation
  Bias in input representation
  Bias in models

To finish
Bias in research design

Is the problem meaningful and well designed?
- Who is contributing to design decisions?
  - Is the design team inclusive of stakeholders, diversity of profiles?
- What is the power balance?
  - Designers, funders, users
- What are the technical constraints?
  - Data content and nature (beware of overexposure)
  - Data availability (beware of overgeneralization)
- ...

[Monteiro and Castillo, 2019]

slide courtesy of A. Névéol
Examples of Biases in NLP

Into the sources of bias

Bias in research design

Bias in data selection

Bias in annotation

Bias in input representation

Bias in models

To finish
Bias in data selection

Which data?

- Are there access restrictions (copyright, confidentiality, consent)?

- Does content accurately reflect the lived experience of demographic categories such as minorities, disadvantaged groups?

How can it be gathered?

- Sampling methods
- Volume, imbalance
- Need for de-duplication

slide courtesy of A. Névéol (adapted)
Examples of Biases in NLP

Into the sources of bias
  Bias in research design
  Bias in data selection
  Bias in annotation
  Bias in input representation
  Bias in models

To finish
“[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process”  [Leech, 1997]
Manual annotation in NLP, today
With a huge surprise from the side of Bayern Munich as Van Bommel, the captain, has been removed. He is not even on the substitutes list.
Exercice: annotate soccer match comments
players, teams, actions (goals), relations (passes), etc.

With a huge surprise from the side of Bayern Munich as Van Bommel, the captain, has been removed. He is not even on the substitutes list.

What is the task, the application aimed at?

summary of match

Van Bommel?

should not be annotated
The consensus, at the heart of annotation
One needs to "agree to be able to measure" [Desrosières, 2008]

Annotation is related to quantification

Measuring vs quantifying [Desrosières, 2008] :

▶ **measuring**: implies a measurable form (eg. the height of Mont Blanc)
▶ **quantifying**: implies preliminary conventions of equivalence

The consensus should be equipped:

▶ annotation guidelines (12p. for soccer)
▶ meetings with the annotators and the campaign manager
▶ **evaluate** the consensus (consistency)
Impact of data on evaluation

- The importance of *real* baselines (sometimes, they are surprising hard to beat!)
- What does it mean when system F1 $>>$ IAA?

slide courtesy of A. Névéol (adapted)
Impact of data on evaluation

- Similarity between training and test corpus
  - 4 biomedical English benchmark datasets
  - Compare performance in redundant vs. non redundant
- Characterization of memorization vs. generalization
  - What is realistic in a real-life setting?

[Elango van et al., 2021]

slide courtesy of A. Névéol (adapted)
Datasets and corpus development should be documented

- Provenance and availability
- Terms of use, including confidentiality, copyrights
  - Some information is always sensitive (e.g. health, religion)
- Detailed description
  - Language (#BenderRule), volume
  - Selection and collection method
  - Quality assessment, including biases

[Adda et al., 2014, Bender and Friedman, 2018]

slide courtesy of A. Névéol (adapted)
Examples of Biases in NLP

Into the sources of bias
- Bias in research design
- Bias in data selection
- Bias in annotation
- Bias in input representation
- Bias in models

To finish
Bias in input representation

Semantic representations learnt from large corpus contain bias

▶ Intrinsincly
▶ Paris is to France as Rome is to Italy
▶ But: Man is to Computer Programmer as Woman is to... Homemaker

▶ Extrinsincly

slide courtesy of A. Névéol (adapted)
Bias in input representation

Evaluating bias in semantic representations

▶ The minimal pair paradigm
  ▶ ”Women can't drive” vs. ”Men can’t drive”
  ▶ 1,677 sentence pairs in French and English, covering 10 types of bias
▶ Evaluation of masked language models in French and English
  ▶ Comparison of sentence probability
  ▶ Models exhibit bias, except mBERT (less performant, though)
▶ inspired by [Nangia et al., 2020]


slide courtesy of A. Névéol (adapted)
Bias in input representation

Strategies for mitigating bias in language models

- Rebalancing training corpora
- Modifying pre-trained embeddings

Should semantic representations be descriptive or normative? Also, bias mitigation in language models may not impact downstream tasks.

[Bolukbasi et al., 2016]

slide courtesy of A. Névéol (adapted)
Examples of Biases in NLP

Into the sources of bias
  Bias in research design
  Bias in data selection
  Bias in annotation
  Bias in input representation
  Bias in models

To finish
Bias in models

▶ Is it just a matter of fixing the data?
  ▶ **Bias amplification** has been evidenced in tasks such as machine translation and sentiment analysis
  ▶ Spurious correlations between data and predictions has been shown

▶ Model explainability and interpretability

▶ Is no answer better than a biased answer?

slide courtesy of A. Névéol (adapted)
Examples of Biases in NLP

Into the sources of bias

To finish

WYHTR: What You Have To Remember
- biases affect people’s lives
- biases appear because of some people’s (lack of) decisions
- 5 sources of biases in NLP
- manual annotation process
Tutorial (homework, if you feel like it)

How to make a racist AI without really trying


Five sources of bias in natural language processing.

Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, Londres, Angleterre.

Ruined by Design: How Designers Destroyed the World, and what We Can Do to Fix it.
Mule Design.

CrowS-pairs: A challenge dataset for measuring social biases in masked language models.
Men also like shopping: Reducing gender bias amplification using corpus-level constraints.