Fairness in NLP

Karën Fort

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

Sources of inspiration

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

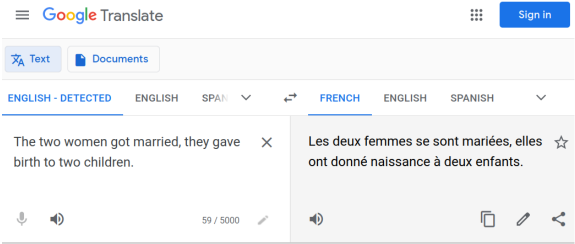
Examples of Biases in NLP

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

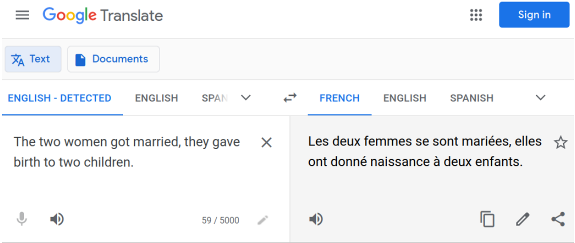
Into the sources of bias

To finish

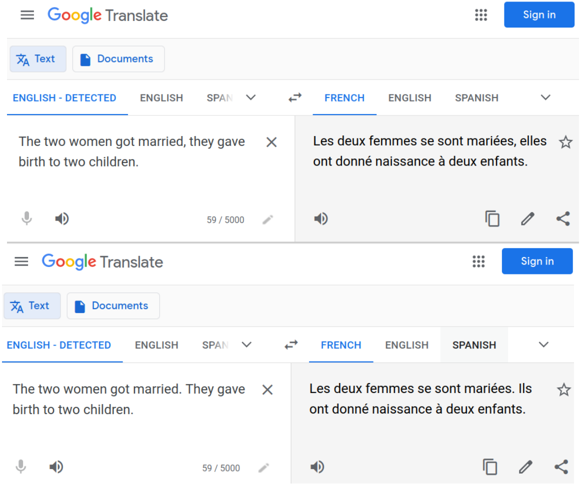
Example of issue: "Neutralization" bias



Example of issue: "Neutralization" bias



Example of issue: "Neutralization" bias



context taken into account (sentence) + masculine = neutral

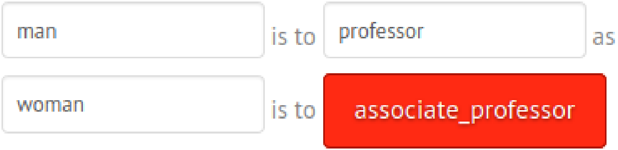
Machine learning is not magic

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

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=BOUTfUmI8vs

Issues in systems' evaluation

Pratiques d'évaluation en ASR et biais de performance

Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(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]

Machine learning is not magic (2)

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?

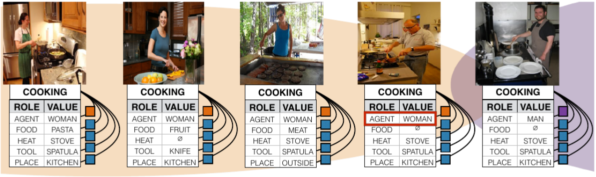


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images 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

https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85

	SIRI	CORTANA	ALEXA	GOOGLE ASSISTANT
Mainstream release date?	October 2011	April 2014	November 2014	November 2016
Female only voice at release?	Yes	Yes	Yes	Yes
Date fully functioning male option was added?	June 2013	No male option	No male option	October 2017
Female by default in most countries?	Yes	Yes	Yes	Yes
Male by default?	Only when the operating system language is set to Arabic, French, Dutch or British English	No	No	No
Descriptions of assistants' personalities by company representatives	'Sense of helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish'	'Supportive, helpful, friendly, empathetic'	'Smart, humble, sometimes funny'	'Humble, it's helpful, a little playful at times'

Programmed reactions

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

Examples of Biases in NLP

"Neutralization"

Invisibilization

Mirror of prejudice?

Consequences in people's life

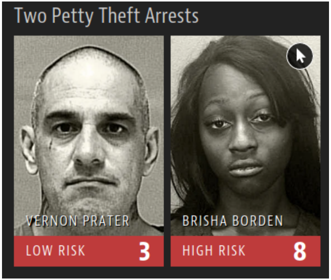
Into the sources of bias

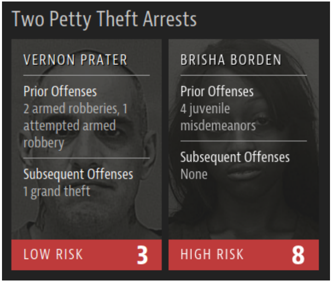
To finish

Justice (*risk assessment instruments*)

systems used in all the states in the USA

Example of COMPAS (2016)





https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
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."

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

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)

Your turn to work

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

HOVY AND PRABHUMOYE — WILEY | 5 of 19

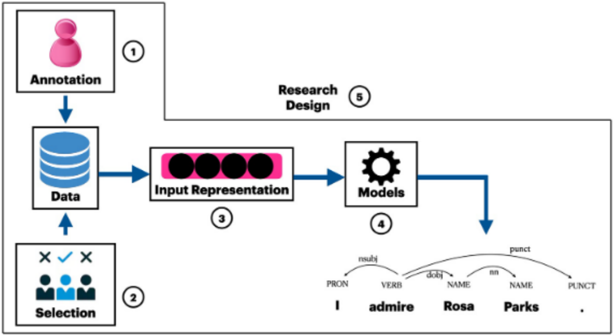


FIGURE 1 Schematic of the five bias sources in the general natural language processing pipeline

[Hovy and Prabhumoye, 2021]

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?

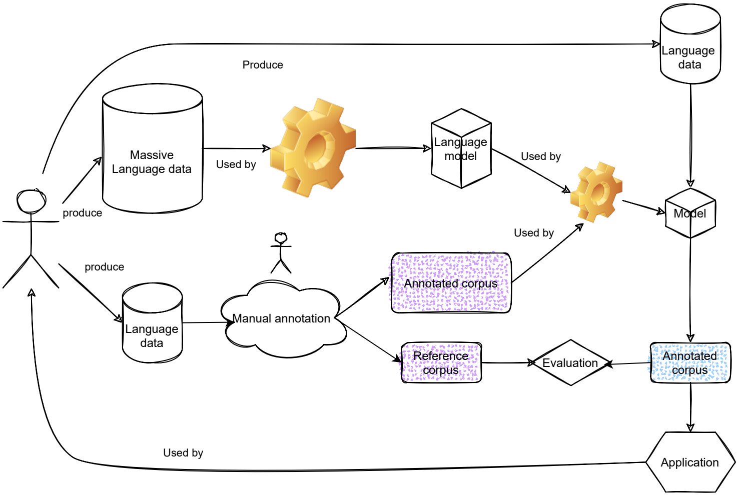
- Sampling methods
- Volume, imbalance
- Need for de-duplication

slide courtesy of A. Névéol (adapted)

Definition

"[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



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.

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?

[Elangovan 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

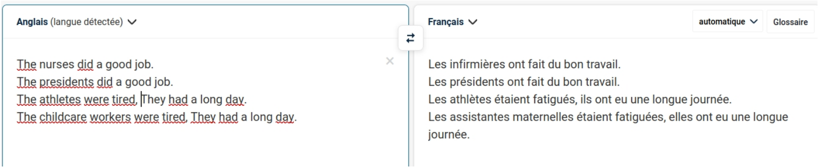
[Adda et al., 2014, Bender and Friedman, 2018]

slide courtesy of A. Névéol (adapted)

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]

Névéol A, Dupont Y, Bezançon J, Fort K. French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. Submitted.

slide courtesy of A. Névéol (adapted)

Bias in input representation

Strategies for mitigating bias in language models

- Rebalancing training coprus
- 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)

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
  - Spurrious 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

Adda, G., Besacier, L., Couillault, A., Fort, K., Mariani, J., and Mazancourt, H. D. (2014).
"where are the data coming from?" ethics, crowdsourcing and traceability for big data in human language technology.
In Crowdsourcing and human computation multidisciplinary workshop, Paris. CNRS.

Bender, E. M. and Friedman, B. (2018).
Data statements for natural language processing: Toward mitigating system bias and enabling better science.
Transactions of the Association for Computational Linguistics, 6:587–604.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016).
Man is to computer programmer as woman is to homemaker? debiasing word embeddings.
In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Desrosières, A. (2008).
Pour une sociologie historique de la quantification : L'Argument statistique.
Presses de l'école des Mines de Paris.

Elangovan, A., He, J., and Verspoor, K. (2021).
Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation.
In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1325–1335, Online. Association for Computational Linguistics.

Garnerin, M., Rossato, S., and Besacier, L. (2020).
Pratiques d'évaluation en ASR et biais de performance.
In Adda, G., Amblard, M., and Fort, K., editors, 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 1–9, Nancy, France. ATALA.

Hovy, D. and Prabhumoye, S. (2021).

Five sources of bias in natural language processing.
Language and Linguistics Compass, 15(8):e12432.

Leech, G. (1997).
Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, Londres, Angleterre.

Monteiro, M. and Castillo, V. (2019).
Ruined by Design: How Designers Destroyed the World, and what We Can Do to Fix it.
Mule Design.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020).
CrowS-pairs: A challenge dataset for measuring social biases in masked language models.
In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017).
Men also like shopping: Reducing gender bias amplification using corpus-level constraints.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.