Ethical considerations for application design in NLP

Lea Krause (l.krause@vu.nl)

Table of contents

01

My PhD

Who am I?
What is my PhD about?
And why this project?

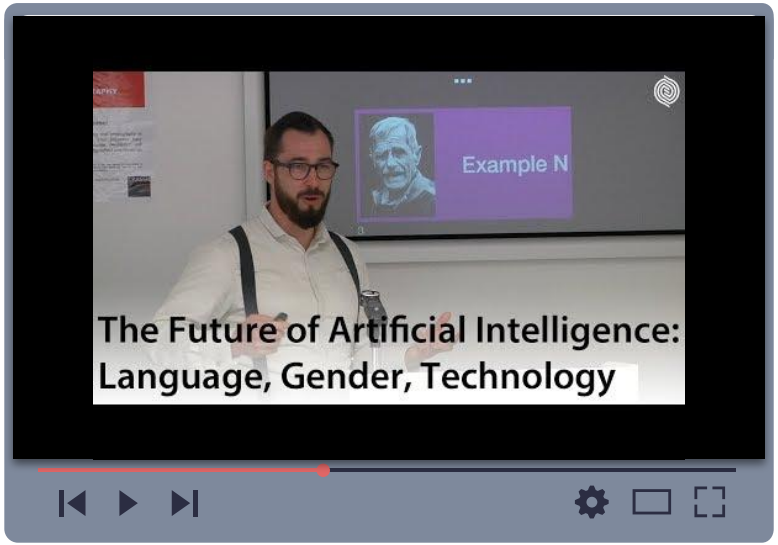
Who am I?

- Bachelor in Linguistics and Phonetics
- Masters in Speech and Language Processing
- Working on/with Leolani since February
- Make robots talk project

What is my PhD about?

- Explainability
  - Making Leolani generate explanations
  - When, what and how to explain?
  - Is there a bias in my explanations?
  - How can we generate meaningful explanations?

Why this project?



The Future of Artificial Intelligence: Language, Gender, Technology - Dirk Hovy (2019)

02

Problem description

Why complicate all this?

"[...] we might be puzzled or amused when receiving an email addressing us with the wrong gender, or congratulating us to our retirement on our 30th birthday. In practice, though, relying on models that produce false positives may lead to bias confirmation and overgeneralization. Would we accept the same error rates if the system was used to predict sexual orientation or religious views, rather than age or gender? Given the right training data, this is just a matter of changing the target variable."

—Hovy and Spruit, 2016

Problem description

- If we rely on image data to tell us the gender of a person, we might cover the majority of the population, but we will exclude already marginalised groups like non-binary or trans people.
- As an example Leolani if Leolani solely relies on visual data to identify gender and this is only trained on binarily annotated data, we will always fail to correctly label non-binary people and most of the time misgender trans people.

→ What can we do to make our technologies more inclusive?

Project layout

Readings

We start out with related work on ethics in NLP and more specifically gender as a variable

Approach

We decide on which questions to tackle and on an appropriate approach

Implementation

We implement our communication module

03

Related Work

How is ethics in NLP currently covered? What is the consensus on gender as a variable?

Related work

Impact of NLP

Hovy & Spruit (2016)

Design guidelines

Larson (2017)
Leidner & Plachouras (2017)
Selbst et al. (2019)

Automatic gender recognition (AGR)

Wu et al. (2020)
+ Response by Keyes (2020)

Perspectives from non-binary and trans people

Keyes (2018)

04

Approaches

What can we do about this?
How can we implement this into the Leolani platform and fit with the other projects?

Approaches

01

Training data

We can alter the training data to include more classes or turn the gender variable into a continuum

02

Communicative module

We only trust explicit information given by users. This can be indirect through conversation or direct as an answer to a question from Leolani.

Possible questions

- How does Leolani deal with missing information?
  - Which pronouns should Leolani use until she has confirmation?
  - Which kinship terms should Leolani use until she has confirmation?
  - Should she explicitly ask for e.g. pronouns?
  - Should she wait until it becomes clear from conversations?
  - What other indicators apart from pronouns should be indicative?
- How should Leolani update information in the brain?
  - How should Leolani react to pronoun changes?
    - he/him → they/them
  - What if a person transitions during the time Leolani knows them?
    - Sister → brother
- How much can we actually rely on image data?
  - Still useful for age estimation

Thanks

Do you have any questions?

l.krause@vu.nl

References

Hovy, D., & Spruit, S. L. (2016). The Social Impact of Natural Language Processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. https://doi.org/10.18653/v1/P16-2096

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 88:1–88:22. https://doi.org/10.1145/3274357

Keyes, O. (2020). *Gender classification and bias mitigation: A post-publication review*. Retrieved 30 October 2020, from https://ironholds.org/debiasing/

Larson, B. (2017). Gender as a Variable in Natural-Language Processing: Ethical Considerations. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 1–11. https://doi.org/10.18653/v1/W17-1601

Leidner, J. L., & Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 30–40. https://doi.org/10.18653/v1/W17-1604

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Wu, W., Protopapas, P., Yang, Z., & Michalatos, P. (2020). Gender Classification and Bias Mitigation in Facial Images. *12th ACM Conference on Web Science*, 106–114. https://doi.org/10.1145/3394231.3397900