Rassismus in deutschen Word Embeddings

Diskriminierender Bias finden und bewerten

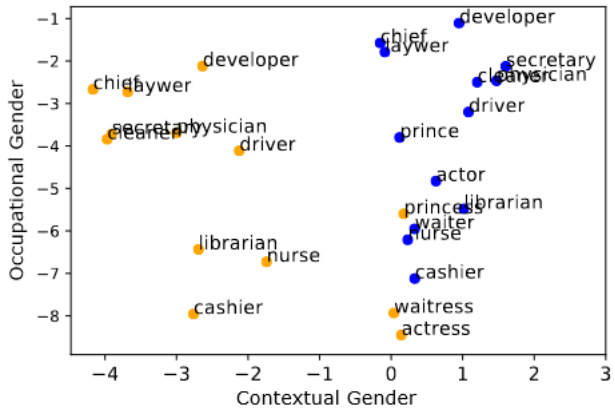
Problemstellung

- Word Embeddings kodieren Assoziationen von Wörtern mit anderen Wörtern in Vektoren (Mikolov et al., 2013)
- Dabei können diskriminierende Assoziationen übernommen werden (Caliskan et al., 2017)
- Das hat wiederum Auswirkungen auf NLP-Anwendungen wie Übersetzungsprogramme
- Wenig Forschung zu Bias in deutschsprachigen Word Embeddings und noch weniger zu rassistischen Assoziationen

Wie können rassistische Assoziationen in deutschen Word Embeddings festgestellt werden? Wie können diese bewertet werden?

Wie können wir rassistische Assoziationen finden? (Manzini & Lim et al., 2019)

- Benötigt wird ein Set Bias definierender Wörter
- Berechnung des Durchschnitts des Sets
- Berechnung der Differenz der einzelnen Wörter zum Durchschnitt
- Hauptkomponentenanalyse mit dem Produkt aller Differenzen
- Mit den Eigenwerten kann dann der Bias-Subspace bestimmt werden



Wörter bewegen sich in zwei Richtungen je nach Stärke der Assoziation mit einem Geschlecht (Zhao et al., 2019)

Wie können wir rassistische Assoziationen bewerten? (Bolukbasi et al., 2016)

- Nutzer*innenstudie mit durch das Embedding generierte Analogien
- Embedding soll Antworten auf Fragen der Form: *a ist für b wie x für y* mit *a, b* und *x* gegeben
- Ziel: feststellen, ob das Embedding diskriminierende Assoziationen herstellt
- Zwei Fragen an Teilnehmer*innen:
  - Ist die Analogie sinnvoll?
  - Reproduziert die Analogie Stereotype?

Racially Biased Analogies

black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led

Durch Embedding generierte diskriminierende Analogien (Manzini & Lim et al., 2019)

Vorgehen

- Embedding mit fastText (Grave et al., 2018) und Leipzig News corpora (Goldhahn et al., 2012) trainieren
- Bias Subspace finden mit Hauptkomponentenanalyse
- Bias nachweisen im Embedding durch Umfrage mit generierten Analogien
- Integreat Texte anhand Mean Average Cosine Similarity untersuchen, kodiert in fastText

- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.
- Goldhahn, D., Eckart, T., & Quasthoff, U. (2012, May). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC* (Vol. 29, pp. 31-43).
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
- Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.