How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns

Stephanie Brandl, Ruixiang Cui, Anders Søgaard

University of Copenhagen, Denmark

eMail: {brandl, rc, soegaard}@di.ku.dk

Introduction

Gender-neutral pronouns serve to

a) include non-binary people

b) as a generic singular

Recent results from psycholinguistics suggest that gender-neutral pronouns (in Swedish) are *not* associated with human processing difficulties. This, we show, is in sharp contrast with automated processing.

TL;DR

70-åringen dammsög golvet i vardagsrummet. Han/Hen skulle få besök på kvällen.
The 70-year-old vacuumed the living room floor. He/They would have visitors in the evening.

- we translate stimuli from a Swedish eye-tracking study [1] into English and Danish
- we compare model perplexity across gendered and gender-neutral pronouns for all three languages
- we investigate performance differences across pronouns for natural language inference (NLI) and coreference resolution
- we find that NLP models, unlike humans, are challenged by gender-neutral pronouns, incurring significantly higher losses when gendered pronouns are replaced with their gender-neutral alternatives

Results

	en			da			sv		
	she/he	they	xe	hun/han	de	høn	hon/han	hen	
perplexity	1	1.49	2.37	1	1.21	3.35	1	1.8	
correlation	0.12	0.26	0.32	-0.14	0.03	-0.1	0.19	0.09	
	0.28	0.33	0.49	0.13	0.17	0.21	0.65	0.72	
	0.28	0.33	0.49	0.13	0.17	0.22	0.65	0.72	

Table 1: Perplexity scores across pronouns and languages for the eye-tracking stimuli. Perplexity values for the datasets with gendered pronouns are set to 1 and we show relative increase for gender-neutral pronouns within a language. Correlation between attention flow and perplexity are listed row-wise for layers 1, 6 and 12.

- perplexity scores for sentences with gender-neutral pronouns are significantly higher
- for Swedish and English we can see a clear increase in correlation (attention flow [2] vs. perplexity) across layers which is even clearer for *hen*
- for Danish we see a much weaker correlation

To investigate those effects across layers further, we look at word embeddings for all Swedish pronouns from all 12 layers in BERT and compute pair-wise cosine similarity including the Swedish word for book (*bok*) as a baseline where we expect no specific relation to pronouns.

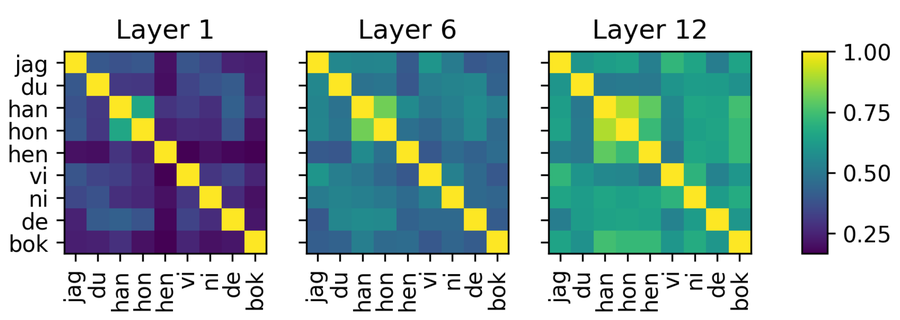

Figure 1: Pair-wise cosine similarity between word representations of all pronouns and the Swedish word *bok* (book) as a baseline for different layers of BERT. We see that gender-neutral *hen* grows from being an outsider (similar to *bok*) in the 1st layer into the cluster of gendered 3rd person pronouns *hon/han* across layers.

- we see less similarity between *hen* and the other pronouns in layer 1
- for layer 6 and 12 word representations seem to be more similar and the three 3rd person pronouns *hen, han, hon* get closer to each other

Results

Natural Language Inference

	en			da			sv		
	orig.	they	xe	orig.	de	høn	orig.	de	hen
mBERT	**83.33**	83.23	81.82	71.15	**71.24**	69.72	**71.91**	71.14	71.06
XLM-R	**95.13**	94.81	94.05	**80.19**	79.18	75.48	**78.79**	78.5	78.58

Table 2: Accuracy [in %] on NLI for English, Danish and Swedish for both models mBERT and XLM-R. Accuracies are calculated on the subset of sentences that contain relevant pronouns (924 for en and 2339 for da/sv). The first column for each language shows the accuracy on the original data, second and third columns show accuracies for respective gender-neutral pronouns.

- we overall see a very small drop in performance for the datasets with gender neutral pronouns compared to the original sentences
- we see the biggest difference for the Danish pronoun *høn* in comparison to the original dataset

Coreference resolution

	she	he	they	xe
acc in %	42.92	**43.75**	27.92	0

Table 3: Results for the pronoun resolution task on the English Winogender dataset.

	orig.	de	høn
F1-score	**0.64**	0.63	0.62
Prec.	**0.70**	0.69	0.69
Recall	**0.59**	0.57	0.56

Table 4: Results for the Danish coreference resolution task. Pronouns in the original dataset (orig.) have been exchanged for singular *de* and gender-neutral *høn*.

- **English:** we see a clear drop in performance from gendered pronouns (*she, he*) to both gender-neutral pronouns (*they, xe*)
- for *xe*, the model was not able to perform coreference resolution at all
- **Danish:** a more extensive coreference resolution task
- small drops in performance for singular *de* and *høn*

Conclusion

We provide a first study on how well language can handle gender-neutral pronouns in Danish, English and Swedish for various tasks.
We find that

- the increase in perplexity implies that language models indeed struggle with the use of gender-neutral pronouns, even with singular *they*
- this is most likely because gender-neutral pronouns are not represented as much as gendered pronouns in the training data
- word representations of all Swedish 3rd person pronouns growing closer in middle/top layers suggests that relevant information is learned for *hen*
- classification in NLI probably does not heavily rely on individual pronouns in most cases hence the small performance differences
- the clear drop in pronoun resolution for singular *they* is surprising
- the smaller drop in performance for Danish coreference resolution might be because this dataset does not solely focus on pronoun resolution

We strongly argue that more needs to be done to adapt language models to a more gender inclusive language, initiatives like the rewriting task as proposed by [3] need to be implemented and extended.
Our code is available at
`github.com/stephaniebrandl/gender-neutral-pronouns`

References

[1] Vergoossen, H. P., Pärnamets, P., Renström, E. A., and Gustafsson Sendén, M. (2020). Are new gender-neutral pronouns difficult to process in reading? The case of Hen in SWEDISH. Frontiers in psychology, 11, 2967.

[2] Abnar, S., and Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4190-4197).

[3] Sun, T., Webster, K., Shah, A., Wang, W. Y., and Johnson, M. (2021). They, them, theirs: Rewriting with gender-neutral english. arXiv preprint arXiv:2102.06788.