Measuring Gender Bias in Word Embeddings for Russian Language

Alena Pestova

Introduction

- The problem of gender bias in Natural Language Processing (NLP) models has become a growing concern in the NLP community in recent years (Costa-jussà et al., 2019, 2020).
- Due to the fact that the texts (fiction, news, Wikipedia, web data) on which models are trained often contain stereotypes and prejudices (Anderson & Hamilton, 2005; Arslan & Koca, 2007; Bamman & Smith, 2014; Graells-Garrido et al., 2015; Macharia et al., 2015; 2018; Mateos de Cabo et al., 2014; Pyle, 1976; Schwartz et al., 2013; Singh et al., 2020; Wagner et al., 2015), NLP models demonstrate social biases in terms of gender, race, and religion (Bhaskaran & Bhallamudi, 2019;Nadeem et al.,2020; Sheng et al., 2019; Yeo & Chen, 2020).

Introduction

- Word embeddings (WE) as a very common framework in NLP were shown to reproduce various prejudices as well and gender bias, in particular (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Gonen & Goldberg, 2019).
- Existing research on gender bias in WE often focus on English language models and there is no such research for WE for Russian.

Why Study Gender Bias in WE?

- For mitigating bias: NLP models learn stereotypes from training data, reproduce and further reinforce them in society.
- For social research: WE models are aggregators of texts, which makes it possible to do research on gender stereotypes, their change and development in society (Garg et al., 2018; Kozlowski et al., 2019; Wevers, 2019).

Goal of this Study

- To study whether gender bias is present in different Russian-language word embeddings models and in what topics.

Data

The available word embedding models from the RusVectores website (Kutuzov & Kuzmenko, 2017).

a) RNC_cbow (*ruscorpora_upos_cbow_300_20_2019)*:

CBoW embeddings with trained on Russian National Corpus

b) RNC-Wiki_skip (*ruwikiruscorpora_upos_skipgram_300_2_2019)*:

Skipgram embeddings trained on Russian National Corpus and Wikipedia

In addition to Russian National Corpus, in this WE, dump of Russian Wikipedia for 2019 is used.

c) Tayga_skip *(tayga_upos_skipgram_300_2_2019)*:

Skipgram embeddings trained on the webcorpus Tayga (Shavrina & Shapovalova, 2017).

This corpus consists of literary texts, social media, subtitles, news, poems and other texts. The subcorpus of poems was not used for training this WE, so the literary texts make up 95% of the used texts.

d) News_skip *(news_upos_skipgram_300_5_2019)*:

Skipgram embeddings trained on Russian language news

e) GeoWac_fast *(geowac_lemmas_none_fasttextskipgram_300_5_2020)*:

Fasttext embeddings trained on the corpus GeoWAC (Dunn & Adams, 2020)

The WEAT method

Word Embedding Association Test (Caliskan, et al., 2017)

Attribute words

***Male and female terms*:**

A: мужчина, мужской, мальчик, брат, сын, отец, папа, дедушка, дядя

B: женщина, женский, девочка, сестра, дочь, мать, мама, бабушка, тетя

Target words

E1: career and family

X: руководитель, менеджмент, профессионал, корпорация, зарплата, офис, бизнес, карьера

Y: дом, родитель, ребенок, семья, род, брак, свадьба, родственник

Methods: measuring gender bias with the WEAT method

H0: two sets of target words which we suspect to be biased are not different regarding their relative similarity to the two sets of attribute words (male and female terms).

Test statistic is calculated as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where s(w, A, B) is

$$s(w, A, B) = \text{mean}_{a \in A} cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} cos(\vec{w}, \vec{b})$$

P-value was calculated with the permutation test with 100000 iterations.

Effect size: $$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

Attribute words

Male and female terms:

A: мужчина, мужской, мальчик, брат, сын, отец, папа, дедушка, дядя

B: женщина, женский, девочка, сестра, дочь, мать, мама, бабушка, тетя

Target words

E1: career and family

X: руководитель, менеджмент, профессионал, корпорация, зарплата, офис, бизнес, карьера

Y: дом, родитель, ребенок, семья, род, брак, свадьба, родственник

Word categories

Attribute words: male and female terms

Target word categories:

- career vs family
- math vs arts
- science vs arts
- intelligence vs appearance
- strength vs weakness
- STEM vs humanities
- rationality vs emotionality

Results

	RNC_cbow		RNC-Wiki_skip		Tayga_skip	
Word Categories	***d***	***p-value***	***d***	***p-value***	***d***	***p-value***
E1: career vs family	0,262	**0,0201**	0,210	**0,0281**	0,411	**0,0005**
E2: math vs arts	0,588	**0,0159**	0,243	0,1607	0,667	0,1318
E3: science vs arts	0,469	**0,0244**	0,059	0,3822	0,713	**0,0374**
E4: intelligence vs appearance	0,784	**0,0001**	0,735	**0,00002**	0,916	**0,0002**
E5: strength vs weakness	0,455	**0,0189**	0,377	**0,0057**	0,654	**0,0258**
E6: STEM vs humanities	0,441	**0,0346**	0,086	0,3945	0,990	**0,0445**
E7: rationality vs emotionality	0,503	**0,0152**	0,341	0,0546	0,384	**0,0390**

	News_skip		GeoWAC_fast	
Word Categories	***d***	***p-value***	***d***	***p-value***
E1: career vs family	0,308	**0,0063**	0,064	0,2662
E2: math vs arts	0,130	0,1397	-0,063	0,6762
E3: science vs arts	0,155	**0,0403**	-0,250	0,8937
E4: intelligence vs appearance	0,314	**0,0021**	0,653	**0,0008**
E5: strength vs weakness	0,324	**0,0111**	0,252	0,1400
E6: STEM vs humanities	0,014	0,4812	0,043	0,4095
E7: rationality vs emotionality	0,703	**0,0022**	0,170	0,2309

Table 1: Results for WEAT hypothesis test for seven word categories and five word embeddings for Russian language. Effect size (Cohen's d) and p-value is reported. Statistically significant gender bias is indicated by the p-values in bold ($p < 0.05$).

Future Work

- Future research is needed to study the role of corpus composition, hyperparameters (for instance, window size) and model types of word embeddings in preserving gender bias.
- I will show analysis such analysis in my future paper :)
- Moreover, it is necessary to study whether other methods for measuring gender bias are suitable for analysis of word embeddings for Russian language.