Research Update: On the Reliability of Word Embedding Gender Bias Metrics

Yupei Du

NLP and Society Lab
Utrecht University

y.du@uu.nl

April 1, 2021

Overview

Background

- Gender Bias of Word Embedding
- Bias Metrics
- Statistical Reliability

Methodology

- Reliability Measurements
- Factors Influencing Reliability

Experiments

- Experimental Setups
- Results

Gender Bias of Word Embedding

$$\vec{man} : \vec{king} \sim \vec{woman} : \vec{queen}$$

$$\vec{man} : \vec{programmer} \sim \vec{woman} : \vec{homemaker}$$

Bias Metrics

- Gender Base Pairs (m, f)
  Word pairs with opposite definitional genders (e.g. *father*∼*mother*, *boy*∼*girl*, ...)
- Target Words (w)
  Words of interest (e.g. *programmer*, *homemaker*, ...)

Bias Metrics

- Bias Metrics
  - Direct Bias / Word Association (DB/WA)

$$\mathrm{DB/WA}_w^{(m,f)} = \cos\left(\vec{w}, \vec{m}\right) - \cos\left(\vec{w}, \vec{f}\right)$$

  - Relational Inner Product Association (RIPA)

$$\mathrm{RIPA}_w^{(m,f)} = \vec{w} \cdot \frac{\vec{m} - \vec{f}}{\|\vec{m} - \vec{f}\|}$$

  - Neighborhood Bias Metric (NBM)

$$\mathrm{NBM}_w^{(m,f)} = \frac{|male(w)| - |female(w)|}{k}$$

Senario

Imagine that you attend a test on English writing proficiency. The test might consist of multiple small tests, called **items**, all designed with the same goal to measure one's English writing proficiency. The grader of your performance is the **rater**. Each time you take the test, it represents a **measurement occasion**.

Different types of Reliability

- Test-retest Reliability
  Identity among different *measurement occasions* (e.g. grades from multiple tests should agree).
- Inter-rater Consistency
  Consistency among different *raters* (e.g. grades from different graders should agree).
- Internal Consistency
  Consistency among different *items* (e.g. all the test items should highly relate to each other).

Motivation

So far, these bias scores have been used . . .

- to measure the effects of methods that aim to reduce biases.
- as a refection of gender bias in the training corpus, which can benefit social science research.

Problem

If they are of low stability, the dependability of the derived conclusions will be challenged.

Test-retest Reliability

Intuition

Train word embeddings for multiple times, keep everything the same except for random seeds. The derived bias scores should be (almost) identical.

- Source of variation: random seeds
- Measurement: ICC (2, 1)
- Inputs
  - Gender base pair: target word list × random seeds
  - Target word: gender base pairs × random seeds

Inter-rater Consistency

Intuition

Bias scores calculated by different bias metrics should be consistent.

- Source of variation: bias metrics
- Measurement: ICC (3, 1)
- Inputs
  - Gender base pair: target word list × bias metrics
  - Target word: gender base pairs × bias metrics

Internal Consistency

Intuition

Bias scores calculated by different gender bias pairs should be consistent.

- Source of variation: gender base pairs
- Measurement: Cronbach's alpha
- Inputs
  - target word list × gender base pairs

Analyses of Factors Influencing Reliability

Use regression models to analyze factors influencing reliability

- Predictors
    - Word frequency (number of occurrence time)
    - Syntactic role of words (PoS tag)
    - Number of senses (number of WordNet synsets)
    - Dispersion of context (entropy of occurrence context words)
    - Word embedding properties (stability, norm, etc..)
- Outcomes
    - Test-retest reliability of target words
    - Inter-rater reliability of target words

Regression Analyses

Problem

We have different corpora as well as different embedding algorithms.

A direct solution

Train multiple linear regression models separately.

A better solution

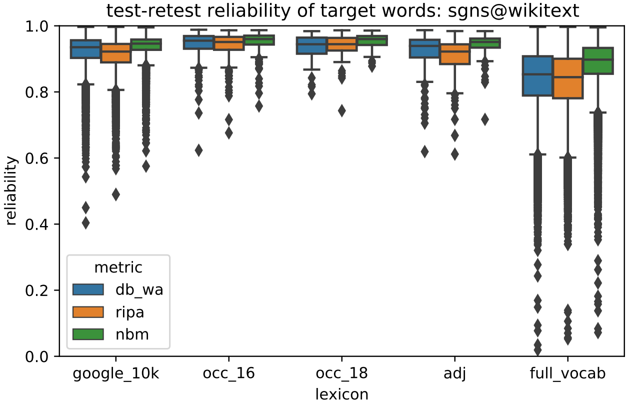
Train a nested multilevel model.

Experimental Setups

- Corpora
  - WikiText-103
  - SubReddits: AskScience and AskHistorians
- Gender base pairs
  - 23 gender base pairs from previous studies
- Target word lists
  - Full vocabulary
  - 10K most common words from Google's Trillion Word Corpus
  - lists of profession words and adjectives from previous studies
- Word embedding algorithms
  - Skip-gram with negative sampling (SGNS)
  - GloVe

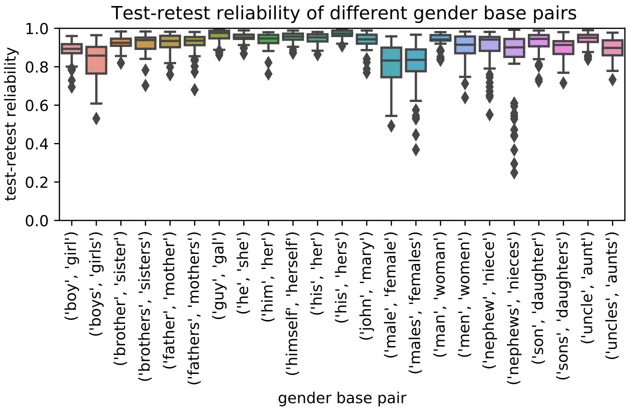
(Part of) Results

Test-retest reliability of target words



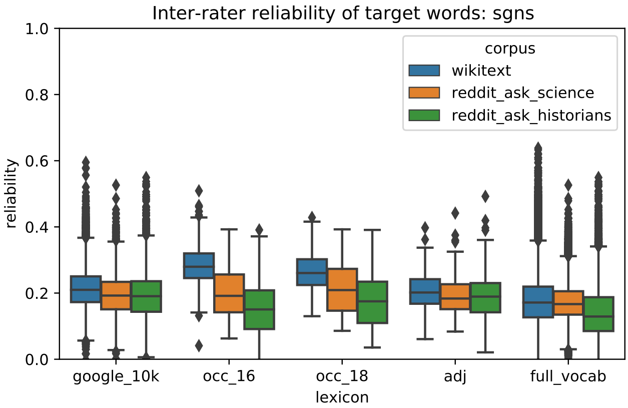
(Part of) Results

Test-retest reliability of gender base pairs



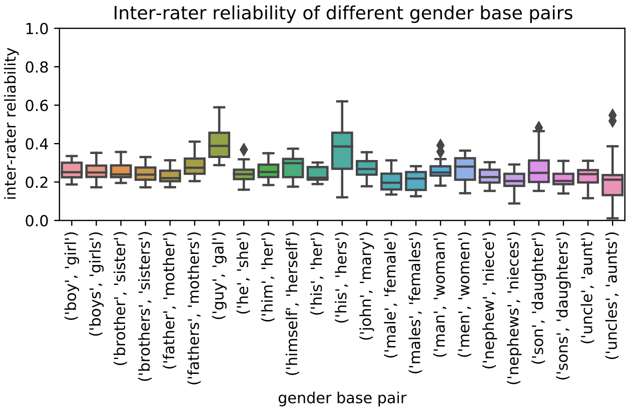
(Part of) Results

Inter-rater consistency of target words



(Part of) Results

Inter-rater consistency of gender base pairs



The End