Measurement and validity

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

14 Sep 2019

Announcements, logistics

- Early exam:
  - Some example (way outdated) exams posted to Canvas
  - Make-up exam for Hopper attendees: Tue 8 Oct, 5p-6:15p, IRB-2137
    - If this doesn't work, let me know NOW!
- Hal's OHs shifted this week to 1p-2p (Thr 26 Sep)
- HW2:
  - Please develop locally and only submit to JupyterHub when you're done
  - There may be a lot of demand on the server around the deadline
    - You'll be less stressed out if you submit early!
- Hal needs to rush out after class today

Last time

- Where do you get your data from?
- How do you annotate it?
- How do you measure annotator agreement?
- How can you make sure you've produced high quality data

Today

- What does it even mean to agree
- What if there's no "gold standard"
- What are we even annotating in the first place?

Goal: translate validity from education → NLP

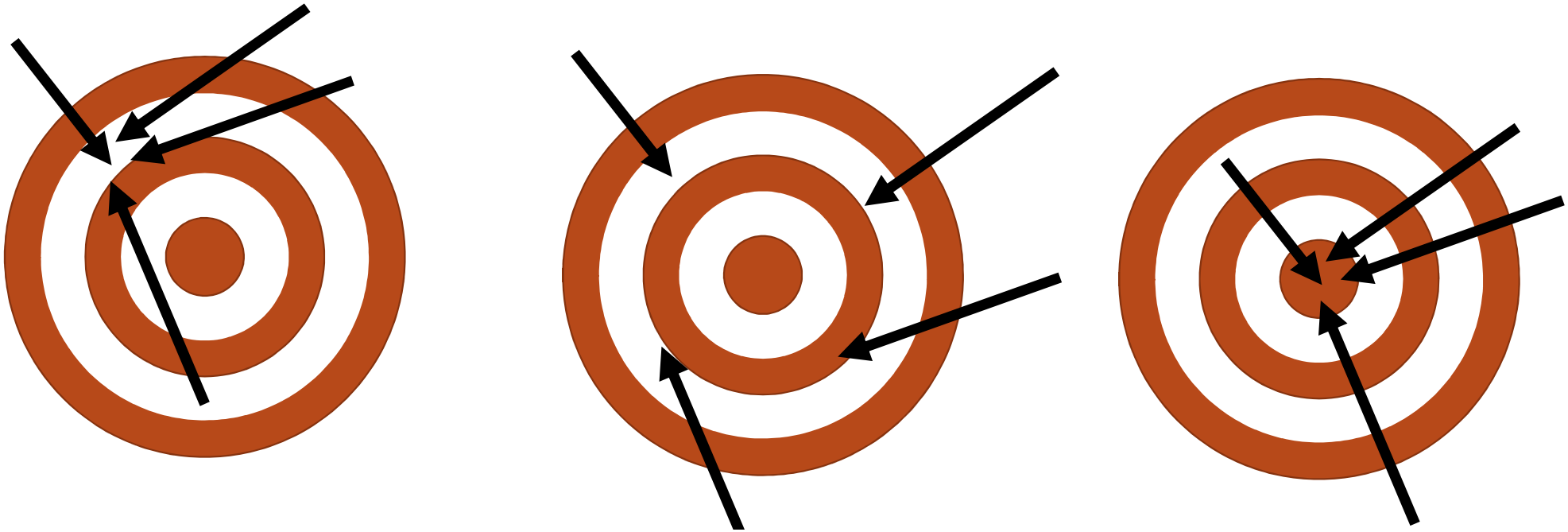
- Education focuses on whether *educational tests* actually measure *learning*
- Analogy 1:
  - Theoretical construct: Hal's weight
  - Possible measurements: ...
- Analogy 2:
  - Theoretical construct: Socio-economic status
  - Possible measurements: …
- Analogy 3:
  - Theoretical construct: Gender bias in word embeddings
  - Possible measurements: …
- Analogy 4:
  - Theoretical construct: Language toxicity
  - Possible measurements: …

What is your measurement good for?

- First, evaluate the construct.
- Is it essentially contested?
- How is it (multiply) defined and what are the sources of disagreement?
- What theory will you use?

Properties that measurements should have…

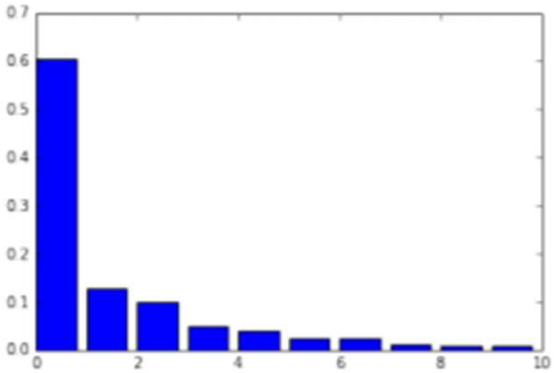
- Validity
- Reliability



Running example: gender bias in embeddings

"man:woman :: computer programmer:homemaker"*

- Construct: gender stereotypes
- Measurement device: analysis of subspaces
  - Define a set of *paired equality sets* (e.g., he/she, man/woman, king/queen)
  - Compute principle direction(s) of variation across those pairs



  - Define the resulting subspace to be the "gender" subspace
  - Define the amount of bias to be how much *neural* words vary in this subspace

**note to readers: this paper is quite trans-exclusionary*

Reliability

- If you repeated this process, how different would the results be?
- Where would sources of variation come from?

Different types of validity

- Construct validity: does it measure what it claims to be measuring?
  - Face validity: does it pass the sniff test?
  - Exclusivity: is it redundant?
  - Discriminant validity: is it uncorrelated with things it should be uncorrelated with?
  - Predictive validity: can it be used to predict things that it should? (also: concurrent validity)
  - Consequential validity: what are the potential risks if scores are invalid/improperly interpreted?
  - Hypothesis validity: do the results match what theory suggests?
- Content validity: does it wholly operationalize the substantive content of the construct?
- Convergent/criterion validity: does it correlate with previously validated tests?

Face validity: does it pass the sniff test?

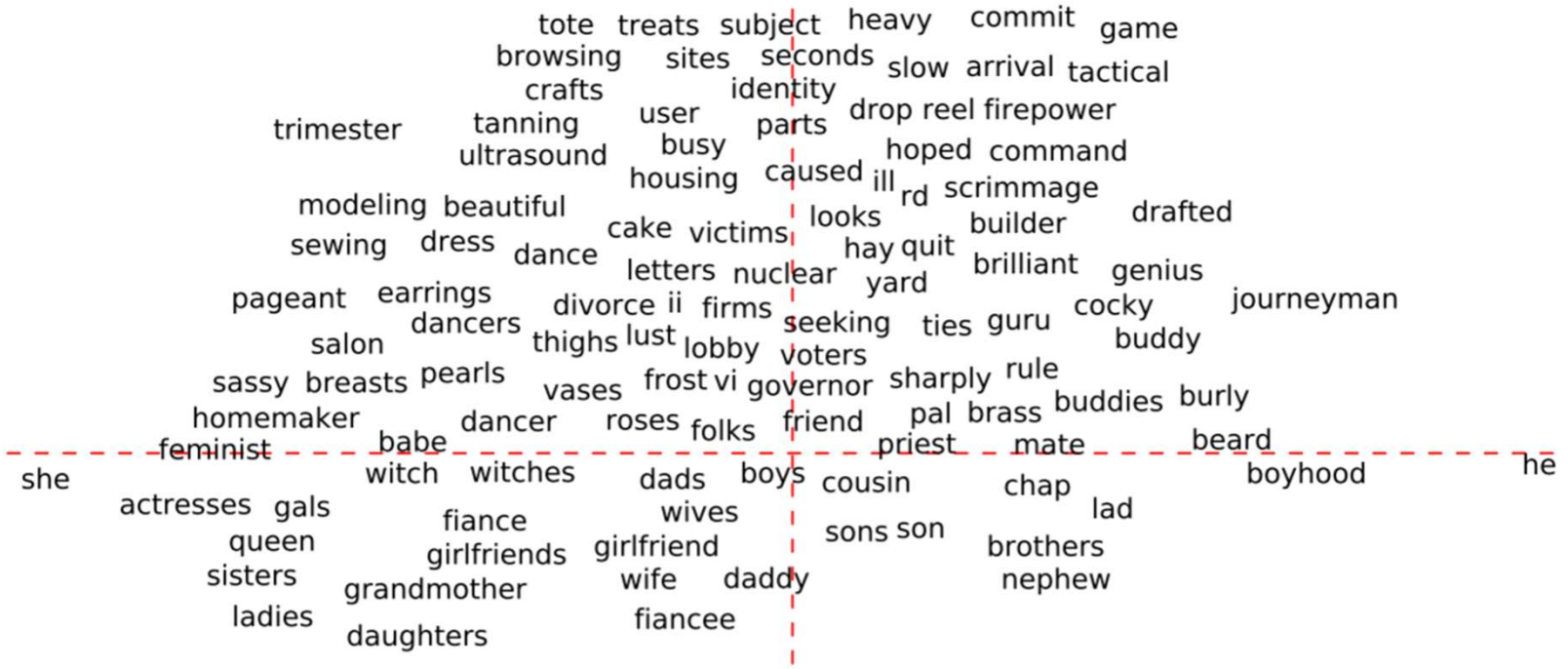


Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

Exclusivity: is it redundant?

- There's only a single measurement when using k=1 with SVD
- How else might we test redundancy?

Discriminant validity: is it uncorrelated with things it should be uncorrelated with?

- Use of “known neutral words” explicitly captures this, for that set of words
- Could also test for…

Predictive validity: can it be used to predict things that it should? (also: concurrent validity)

- What prediction tasks could we set up using this representation?

Consequential validity: what are the social consequences of using it for this purpose?

Hypothesis validity: do the results match what theory suggests?

Different types of validity

- Construct validity: does it measure what it claims to be measuring?
  - Face validity: does it pass the sniff test?
  - Exclusivity: is it redundant?
  - Discriminant validity: is it uncorrelated with things it should be uncorrelated with?
  - Predictive validity: can it be used to predict things that it should? (also: concurrent validity)
  - Consequential validity: what are the potential risks if scores are invalid/improperly interpreted?
  - Hypothesis validity: do the results match what theory suggests?
- Content validity: does it wholly operationalize the substantive content of the construct?
- Convergent/criterion validity: does it correlate with previously validated tests?

Content validity: does it wholly operationalize the substantive content of the construct?

- What does sociology/gender studies/queer theory tell us about gender?
- What does sociology/anthropology tell us about stereotypes?
- What limitations are there in looking at embeddings?

Convergent/criterion validity: does it correlate with previously validated tests?

- N/A because (as far as I know) this was the first test

- There were later tests proposed that converge in some ways but not others

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them
Hila Gonen and Yoav Goldberg
NAACL, 2019

Going back to *language toxicity*
How would you measure this?

- Construct validity: does it measure what it claims to be measuring?
  - Face validity: does it pass the sniff test?
  - Exclusivity: is it redundant?
  - Discriminant validity: is it uncorrelated with things it should be uncorrelated with?
  - Predictive validity: can it be used to predict things that it should? (also: concurrent validity)
  - Consequential validity: what are the potential risks if scores are invalid/improperly interpreted?
  - Hypothesis validity: do the results match what theory suggests?
- Content validity: does it wholly operationalize the substantive content of the construct?
- Convergent/criterion validity: does it correlate with previously validated tests?

Today

- Measurement arises in many aspects of NLP applications:
  - How you define your features
  - How you define your outcome
  - How you use language to measure sociological constructs

- Thinking specifically about measurement matters especially *when the construct in question is essentially contested*

- Different validity measures capture different things; top three categories:
  - Construct validity
  - Content validity
  - Convergent/Criterion validity

- Like many things:
  - many different categorizations exist
  - use what fits, don't use what doesn't