Evaluating Words against Human Assessment?

The whole idea of evaluating word vectors by relating to human judgements is risky.

- Human-produced datasets are subjective.
- Similarity vs. relatedness.
  - Relatedness: *teacher* $\approx$ *student*, *coffee* $\approx$ *cup*
  - Similarity: *teacher* $\approx$ *professor*, *car* $\approx$ *train*
  - Hill et al. (2017) observed a soft tendency:
    - Monolingual models reflect non-specific relatedness,
    - NMT models reflect conceptual similarity.
    - We saw that too for English-Czech (Abdou et al., 2017).
  - Even if we distinguish them, which should be reflected in embeddings?

Details: Faruqui et al. (2016); Survey of eval. methods: Bakarov (2018)