Caveat on Evaluation (1/2)

Consider word2vec "comprehensive" test set (Mikolov et al., 2013):

- 8.8k "semantic" and 10.6k "syntactic" questions,
- w2v "accuracy is quite good" (eyeballing)
  - The authors do mention that exact-match is "only about 60%").

Kocmi and Bojar (2016) carefully examined the test set:

- "Semantic" questions cover only 3 question types:
  - country→city, country→currency, masculine family member→ feminine
  - Vylomova et al. (2016) test many other relations, e.g. walk-run, dog-puppy, bark-dog, cook-eat.
- "Syntactic" questions constructed by combinations:
  - starting from only 313 distinct word pairs,
  - (leading to only 35 different pairs per question on average),
  - And of the 313 pairs, 286 are formed regularly.