A Critique of a Critique of Word Similarity Datasets: 
Sanity Check or Unnecessary Confusion?

Minh Le 
m.n.le@vu.nl 
July 14, 2017

Abstract

Critical evaluation of word similarity datasets is very important for computational lexical semantics. This short report concerns the sanity check proposed in Batchkarov et al. (2016) to evaluate several popular datasets such as MC, RG and MEN – the first two reportedly failed. I argue that this test is unstable, offers no added insight, and needs major revision in order to fulfill its purported goal.

1 Evaluating Evaluation Datasets

Computational linguists rely on datasets to guide our exploration, judge our methods, and, practically speaking, get our papers published. But before we allow them to substitute our linguistic universe, it is only wise to give them a close examination.

It does not take a genius to know that small datasets are unreliable. Some popular datasets in lexical semantics certainly fall into this category: Rubenstein and Goodenough (1965) and Miller and Charles (1991) which contain only 65 and 30 examples respectively. It is well-known that any conclusions drawn from those datasets must be taken with a grain of salt but Batchkarov et al. make a stronger claim:

“RG and MC do not sufficiently capture the degradation of vector quality as noise is added because $\rho$ may increase with $n$. The variance of the measurements is also very high. Both datasets therefore fail the sanity check.”

Here, the authors do not talk about uncertainty anymore but make a categorical decision. This can be likened to the $p$-value approach to hypothesis testing. In both cases, researchers force yes/no decision by checking some property: whether $p$ is less than a certain value or whether a plot is monotonic.

2 Randomness and Instability

One crucial difference between Batchkarov et al.’s sanity check and significance tests is that the computation of the former involves randomness while the latter does not. To test sensitivity to noise, the authors sample random vectors and add them to the vectors predicted by word2vec, expecting that, as the magnitude of the random vectors increase, they will override the information inside word2vec vectors and drive Spearman’s $\rho$ down.

Figure 1 demonstrates this intuition using synthetic data.

I use synthetic datasets of the same size as those reported in Batchkarov et al.¹ and follow their procedure by adding noise of increasing magnitude to synthetic word vectors. The number of

¹Here, men is a synthetic dataset of the same size as men (Bruni and Baroni, 2014), etc. Notice that we obtain a graph very similar to Figure 3 in Batchkarov et al.
samples $k$ (i.e. the times noise vectors are sampled and performance is measured) is not reported in their paper so I start with a small number $k = 5$. It is immediately clear from the figure that similarity scores tend to go down as error increases, though with occasional surges.

Batchkarov et al. go further, and I think it is here that they get it wrong, by claiming that correlation should always go down:

“Model performance, as measured by the evaluation method, should be a monotonically decreasing function of the amount of noise added.”

In other words, they consider the surges of scores as not an artifact of the randomness in their methodology but an unwanted property of the dataset and those that demonstrate such surges are considered to “fail” the test.

It is not hard to show the contrary though. By increasing $k$ from 5 to 500 (Figure 2), one can observe that the lines of $mc$ and $rg$ smooth out and get closer and closer to that of $men$. At $k = 500$, depending on the particular run, we can get a perfectly monotonically decreasing line for $mc$ (Figure 2d) or $rg$ (Figure 2e). In other words, sometimes $mc$ fails the sanity check and sometimes it passes; the same goes for $rg$ – the test is instable.

3 Conclusions

Batchkarov et al. (2016) proposed a sanity check of word similarity datasets, among other things. In this report, I argue and demonstrate by synthetic data that their test leads to inconsistent conclusions and therefore should not be used. The root cause appears to be confusion between tendency and monotonicity and wrong attribution of abnormality. Moreover, the behavior of the test appears to reflect solely the size of the datasets it is run on. To go beyond an illustration of the obviously high variance of measurement on small datasets and become a viable quality assurance tool, the test needs to be revised to respond to factors such as the word frequency, similarity range, the relationship between included words, etc.

The source code of the experiments is available at: https://git.io/vQMAS.

References

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. A critique of word similarity as a method for evaluating distributional semantic models. Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 7–12, 2016.
Elia Bruni and Marco Baroni. Multimodal Distributional Semantics. 49:1–47, 2014.

George A Miller and Walter G Charles. Language and Cognitive Processes Contextual correlates of semantic similarity Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, 6(1):1–28, 1991. ISSN 0169-0965. doi: 10.1080/01690969108406936.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633, 1965. ISSN 00010782. doi: 10.1145/365628.365657.