Evaluating Verb Similarity Performance of Computational Models

Ellis Cain

Indiana University

Abstract

Abstract: This paper aims to reproduce the evaluation of distributional semantics models and lexical databases using the human similarity rankings found in Gerz et al.'s SimVerb-3500 (2016) gold-standard. Similar to the SimLex-999 (Hill et al., 2015) and SimVerb-3500, the counter fitted paragram model that included context and other parameters to adjust word embeddings performed well in correlation and confusion matrices analysis. However, WordNet did not perform as well as Hill et al. claimed in SimLex-999. Overall, these results support the idea that current computational methods for evaluating similarity are still far from human performance and suggest that distributional semantics models with context or retrofitting could lead to better performance.

Evaluating Verb Similarity Performance of Computational Models

Language learning, or verb learning specifically, is an amazing feat given the complexity and difficulty of the task; it takes years of experience to learn and master, where each learner experiences a multitude of different ambiguous instances. However different the learning situations, we all converge to near-identical representations and mappings between word relationships. In their *Seeking Meaning* paper, Zhang et al. explored the statistical learning of verbs and the ambiguous learning situations using the Human Simulation Paradigm (2020). With the Human Simulation Paradigm, adult participants are asked to "simulate" being child learners by watching egocentric child-view videos, where each short video is muted and a beep is played when a parent utters a target verb. Participants provide a label when a parent utters a verb, which is aggregated together to get a distribution of guesses for each video. Interestingly, there was a whole section in this paper dedicated to using computational methods to evaluate the peoples guesses. Their previous version focused on nouns, so they were able to use a binary correct or incorrect to evaluate the guesses. However, for verbs, "twist" and "turn" could be used to describe the same motion, yet are two unique words. How should they be evaluated? This question led Zhang et al. to use similarity quantification methods like GloVe and WordNet, and eventually to collect their own human similarity judgements. Their conclusion, for that section, was that these computational methods are far from human performance on similarity judgements.

There are plenty of gold-standards for evaluating similarity performance of distributional semantics models, such as WordSim-353 or MEN, which contemporary models have either reached or surpassed human performance (Hill et al., 2014), however, as in the *Seeking Meaning* paper mentioned above, the distributional semantics models (DSM) do not perform accurately.

SimLex-999 (Hill et al., 2014), and the follow-up paper SimVerb-3500 (Gerz et al., 2016), both aim at evaluating previous standards and establishing a new gold-standard that can be used to guide research. There are two main innovations with these papers; they contain adjective, verb, and noun concept pairs that vary for concreteness, and both use specific instructions to tease out similarity (car and bike) rather than association (car and gasoline), as previous standards used the terms interchangeably. The SimLex-999 paper showed that all of the models performed poorly on verbs, which this difference in part of speech evaluations most likely arises from the fact that verbs (and adjectives) are relational concepts (Markman and Wisniewski, 1997), so models that have specific parameters to include this context information should be implemented to increase performance.

Methods

One of the distributional semantics models that will be evaluated is Mrkšić et al. (2016) counter fitted paragram word embeddings. These embeddings are trained on the Paraphrase database (Ganitkevitch et al., 2013) to learn word vectors (d=300) with a focus on paraphrasability. Once complete, the paragram vectors are counter fitted by applying linguistic constraints from the original paraphrase database to improve their quality. Similarity for this model is measured using cosine similarity distance rather than traditional Euclidean distances between the paragram vectors. In SimLex-999, Hill et al. claimed WordNet with Wu-Palmer path similarity performed very closely to human performance on similarity tasks. Thus, the other model is WordNet, a lexical database of English constructed by linguists, where words are organized into synsets (cognitive synonyms or groups of words). Here, similarity is calculated using the Wu-Palmer path similarity function which is based on the number of jumps between synsets.

For comparison, the two models will be compared against the human similarity rankings gathered by Gerz et al. in their SimVerb-3500 paper. The 3,500 verb pairs (827 distinct verbs) were selected from the University of South Florida free association norms dataset and the VerbNet verb lexicon that covered a range of similarities. Human participants were recruited using the online crowdsourcing platform Prolific Academic, which is similar to Amazon Turk. There was a total of 843 participants who generated 65,000 ratings (45% US based, 53% UK based, 2% Ireland based; 54% female, 46% male). Each participant was given a survey consisting of a subset of 70 verb pairs (50 participant-unique pairs, 20 consistency pairs) from the original 3500 verb pairs, with nine checkpoint questions throughout the survey. Participants were given specific instructions to differentiate between similarity and association, which included examples of very similar pairs that are synonymous, similar pairs that are nearly synonymous, and associated pairs that are not similar at all. For each pair, participants were asked to rank their similarity, not association, on a scale from 0-6 (converted to 0-10 to match other rankings). The survey was broken into pages of 7-8 verb pairs, with a duplicate pair from the previous page to ensure consistent ratings. There were also three checkpoint questions throughout the survey, where participants were asked to choose the most similar pair out of three choices. In order to ensure quality ratings, they discarded survey responses if the participants failed any of the checkpoint questions.

Procedure

The human similarity judgements from SimVerb-3500 downloaded from their host website. For the two models, the pretrained counter fitted word vectors from Mrkšić et al. that reached state-of-the-art performance on SimLex-999 were obtained from their GitHub page and

the WordNet python implementation was used. I wrote a python script to read in the SimVerb-3500 human similarity judgements txt file. Next, the SciPy implementation of cosine similarity distance was used to calculate the similarity between the pairs from SimVerb-3500 with the counter fitted paragrams, and the Wu-Palmer path similarity function was used for WordNet. There were 2 verbs in the SimVerb-3500 dataset that were not included in the pretrained counter fitted paragram vectors, so the related 8 verb pairs were not included in the analysis. The NetworkX python implementation was used with the spring layout to create semantic network visualizations of the similarity ranking distributions. I wrote R code to calculate correlations between the human rankings and each model. Before the rankings were split into four bins (<0.25, 0.25-0.5, 0.5-0.75, 0.75-1.0) for confusion matrices, the SimVerb-3500 rankings were divided by 10 to match the same range as the other similarity rankings. The vectors, python scripts, and R analysis code are all located on my GitHub, which the link can be found in the appendix.

Results

For the following visualizations and analyses, the similarity values for the counter fitted paragram vectors were changed to *1 - cosine similarity distance* such that all scales had the highest value (1 or 10) as most similar and the lowest value as least similar. Figure 1 below shows that as expected, based on the construction and selection of the verb pairs, the SimVerb-3500 human rankings had a generally even distribution across the similarity rankings. As seen in Figure 2, the counter fitted paragram similarity scores were left-skewed, but still covered the range of similarity rankings. The distribution of similarity scores for WordNet Wu-Palmer path

similarity can be seen in Figure 3, which show that there is a gap between high and low similarity rankings.

Figure 1: Distribution of human similarity rankings from SimVerb-3500

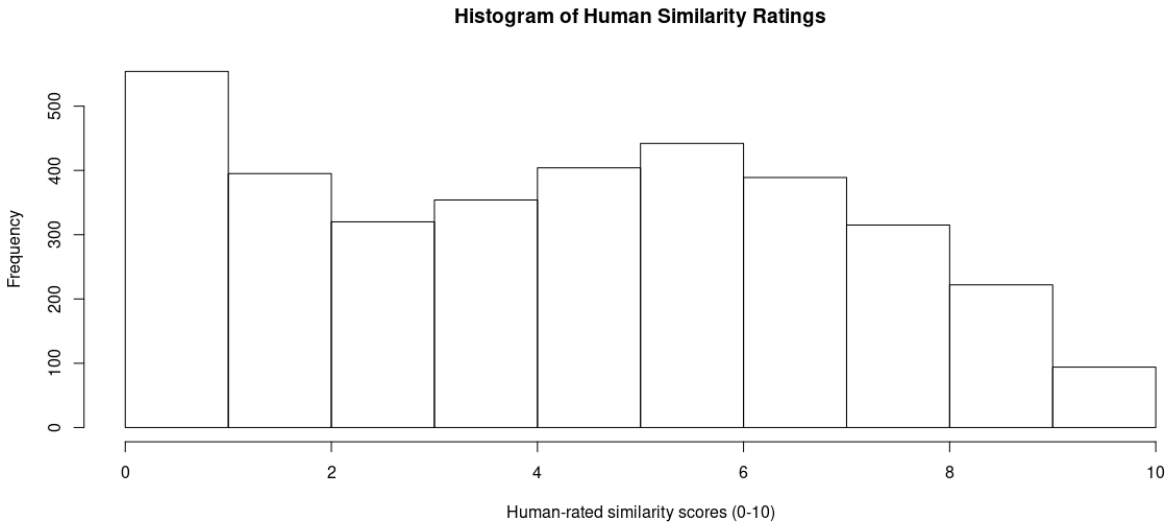


Figure 2: Distribution of similarity rankings from counter fitted paragram vector cosine distance

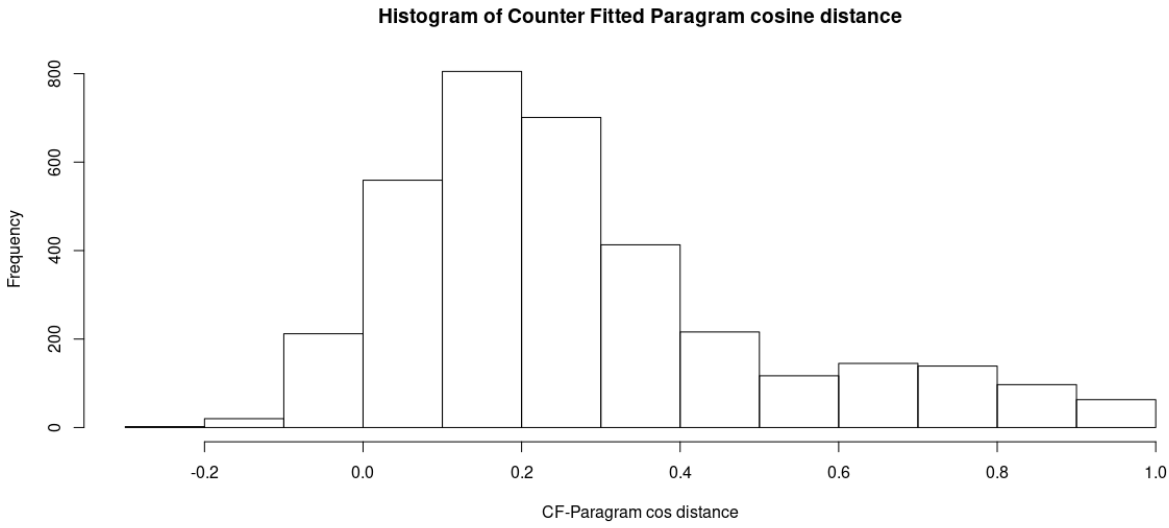
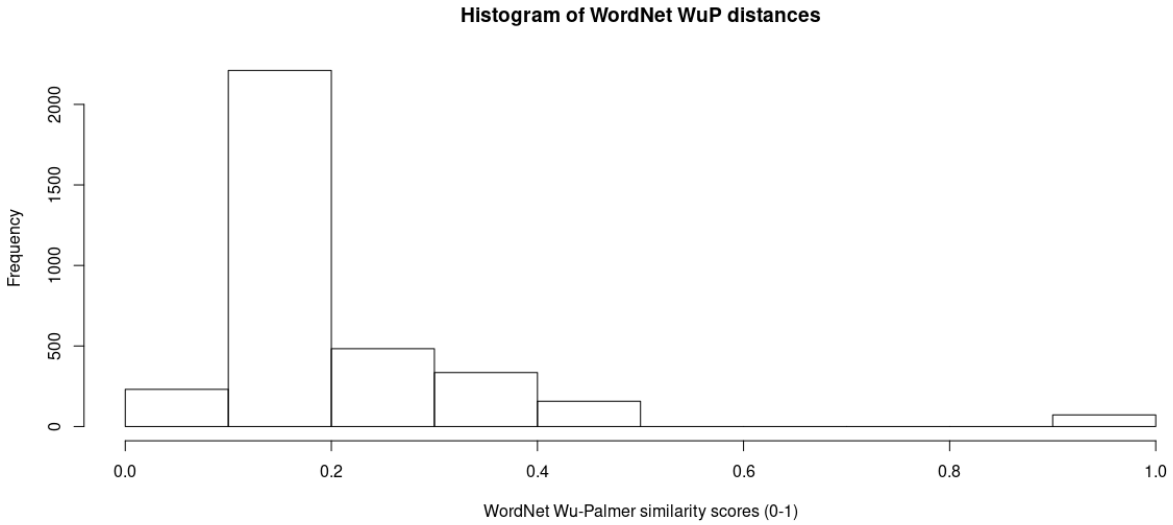


Figure 3: Distribution of similarity rankings from WordNet Wu-Palmer path similarity scores



For alternate visualizations of similarity scores, Figures 4-6 below show the semantic networks for each ranking method. The average clustering coefficients for each graph was 0.216.

Figure 4: Semantic network for human ratings in SimVerb-3500

Figure 5: Semantic network based on similarity ratings from the counter fitted paragram vectors

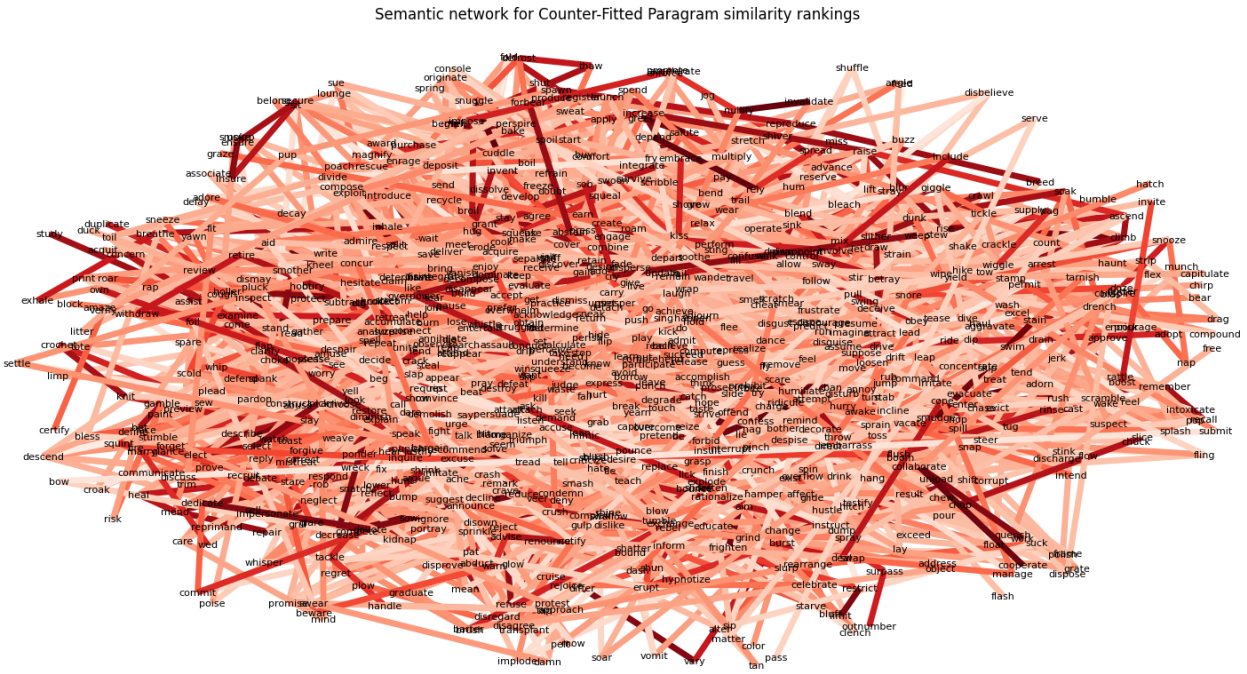
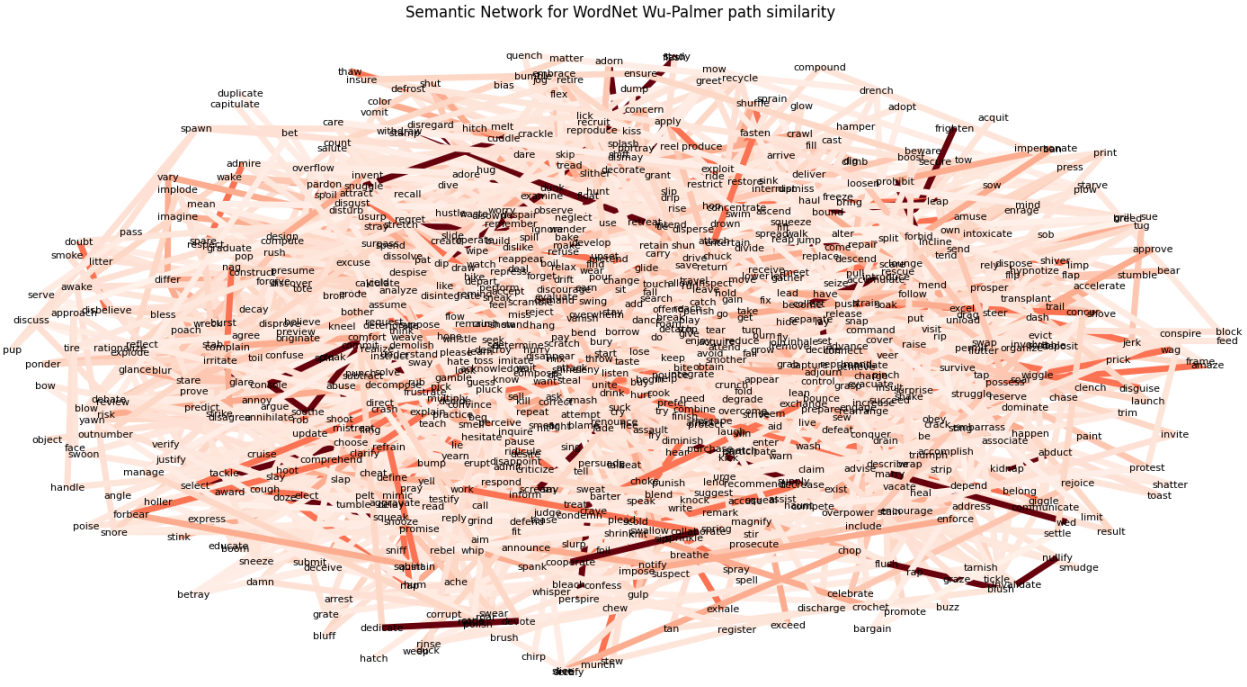


Figure 6: Semantic network based on similarity ratings from WordNet Wu-Palmer path similarity



As seen in Figures 7 and 8, the counter fitted paragrams in general correlated better with the human rankings from SimVerb-3500 when compared to the WordNet similarity scores.

There was a significant correlation of 0.616 between the human similarity ratings and the counter fitted paragram model, $t = 46.187$, $df = 3487$, $p < 2.2e\text{-}16$. There also was a significant correlation of 0.274 between the human similarity ratings and the WordNet Wu-Palmer path similarity ratings, $t = 16.846$, $df = 3487$, $p < 2.2e\text{-}16$. The regression line for SimVerb-3500 and counter fitted paragrams correlation had an intercept of 0.03720 and a slope of 0.05527. The regression line for SimVerb-3500 and WordNet Wu-Palmer correlation had an intercept of 0.15195 and a slope of 0.01515. The values for the linear regression model can be seen in Table 1 below.

Figure 7: Correlation scattergram for SimVerb-3500 rankings and the counter fitted paragram similarity

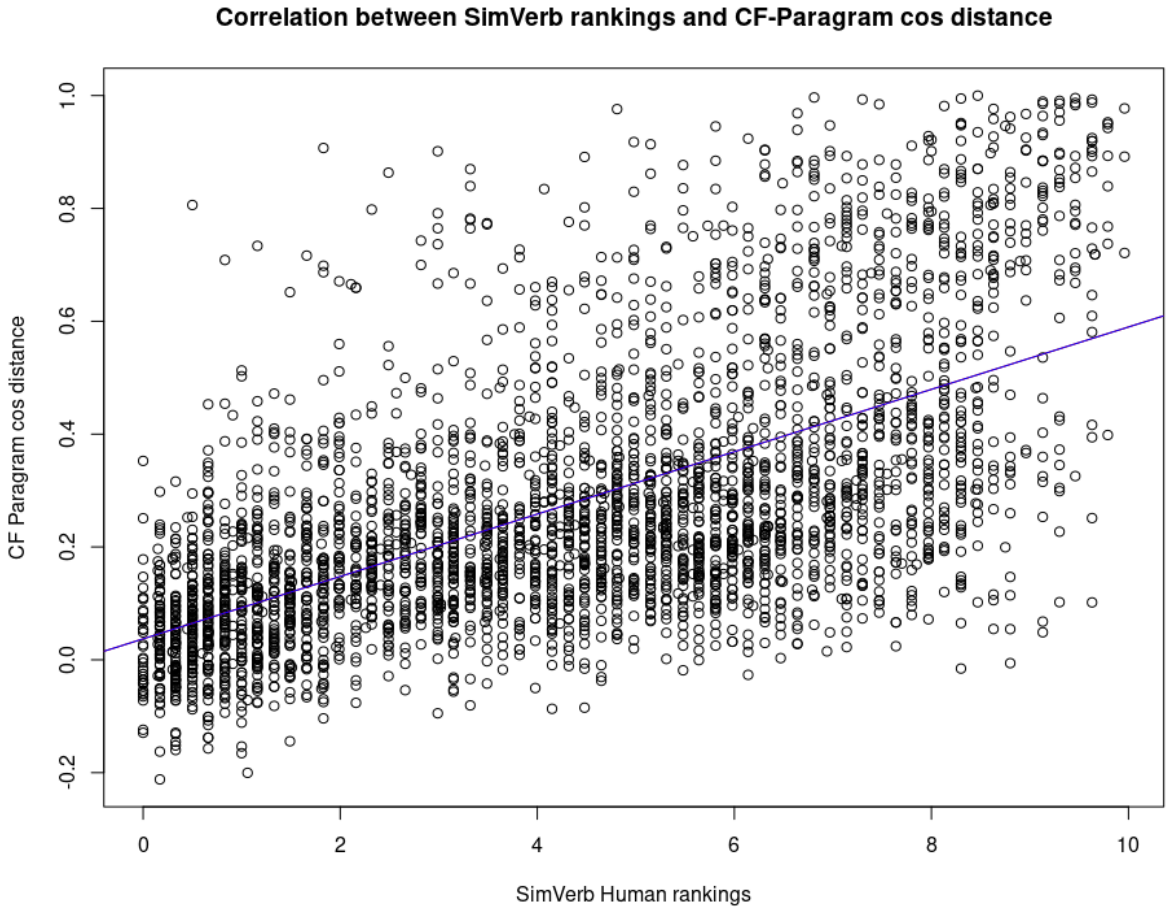


Figure 8: Correlation scattergram for SimVerb-3500 rankings and WordNet Wu-Palmer similarity

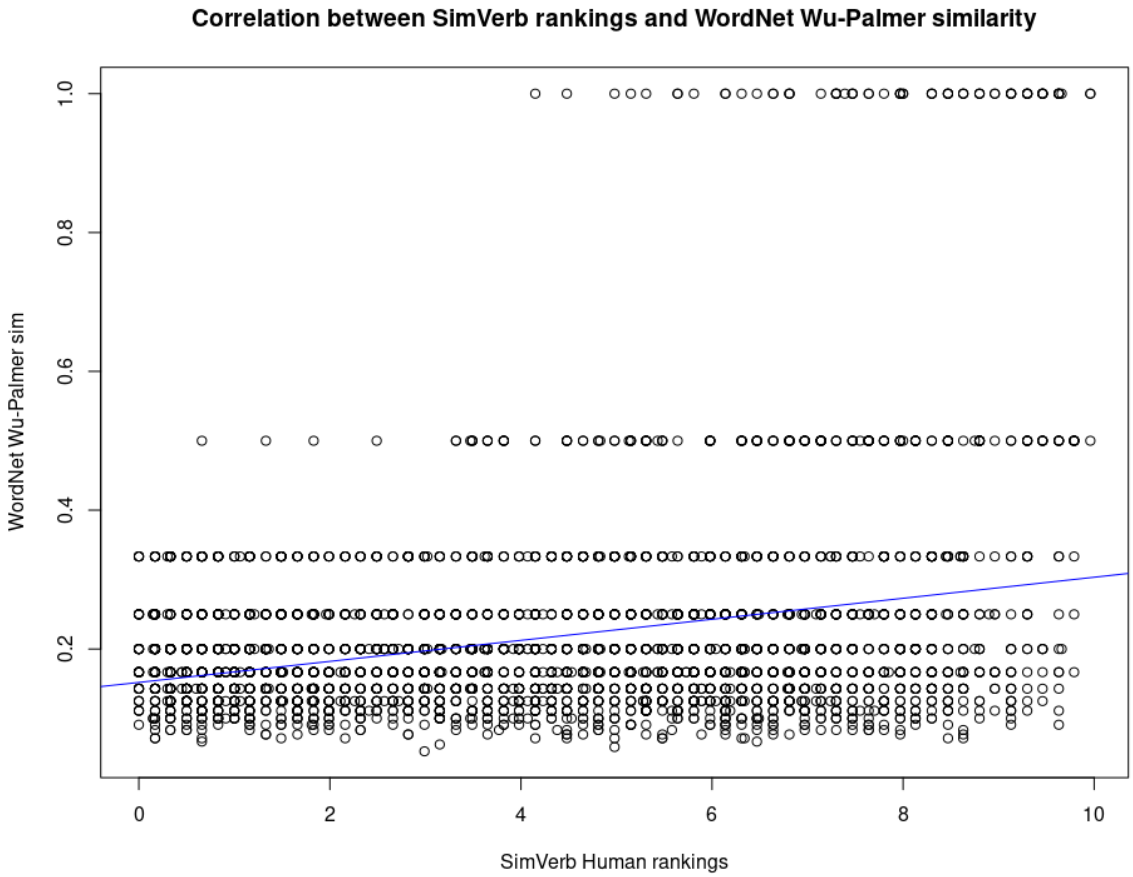


Table 1: Linear intercept and slope of correlations for the two models

Model	Intercept	Slope
Counter fitted paragrams	0.03720	0.05527
WordNet Wu-Palmer	0.15195	0.01515

The last analysis is a confusion matrix; since each of the two computational models output a continuous similarity measure, they were put into four bins (<0.25, 0.25-0.5, 0.5-0.75, 0.75-1.0). Based on the confusion matrix for the counter fitted paragrams, the model had an accuracy of 0.4382, $p < 2.2e-16$, while the WordNet confusion matrix had an accuracy of 0.3325, $p < 0.02$. With only four bins, the counter fitted paragrams did much better than chance, while the WordNet performance was only slightly better than chance. While the accuracies for the two

models were relatively close, Cohen's Kappa was much higher for the counter fitted paragrams (0.212) than WordNet (0.040). The values for the confusion matrix can be seen below in Table 2.

Table 2: Confusion matrix statistics

Model	Accuracy	95% Confidence interval	Kappa
Counter fitted paragrams	0.4382	(0.4217, 0.4549)	0.212
WordNet	0.3325	(0.3168, 0.3484)	0.040

Discussion

Both correlation test and confusion matrix showed that the counter fitted paragram model performed more closely to human performance on the similarity rating task than WordNet's Wu-Palmer path similarity function. The distribution of counter fitted paragram similarity scores also closely match the SimVerb similarity distributions, both generally left-skewed. I was originally hoping to use the semantic networks as a way to analyze and evaluate the performance, but since I used a weighted graph, all of the nodes would be connected, leading to the identical clustering coefficients of 0.216. The counter fitted paragram model also performed much better on the SimLex-999 evaluation dataset, achieving state of the art performance of 0.74 (Mrkšić et al., 2016). In general, the addition of counter fitting the paragrams seemed to do a good job of improving performance when compared with other DSMs like GloVe and other computational methods.

When looking at the biggest difference between model and human performance, there are two representative examples; (jerk, prick) and (draw, attract). For the former pair, the SimVerb-3500 human raters gave it a similarity of 0.5/10, while the counter fitted paragrams gave it a

similarity of 0.865/1 (10 and 1 as most similar respectively). Based on our own intuitions, it can be seen that the SimVerb-3500 raters are more accurate when the two are interpreted as verbs. One can say "I gave the door handle a *jerk*" and "The balloon popped after the needle *pricked* it," while the meaning would change greatly if one were to say "I gave the door handle a *prick*" and "The balloon popped after the needle *jerked* it." The two words are not synonymous at all and cannot be used interchangeably, supporting the SimVerb-3500 rating that they are not similar. However, when the two words are interpreted as nouns, they become synonymous; in the sentence "That _____ turned me into a newt", both *jerk* and *prick* can be used to achieve the same meaning. In the noun sense of the words, the counter fitted paragrams more accurately assessed their similarity. For the latter pair, the SimVerb-3500 human raters gave the pair a similarity of 3.32/10, while the counter fitted paragrams gave it a similarity of 0.839/1. This pair highlights one potential issue with the SimLex or SimVerb survey paradigm; context and multiple meanings. *Draw* in the sense of "I will *draw* two African swallows carrying a coconut by the husk" is not similar to *attract*, while *draw* in the sense of "I put out some scones and tea to *draw* in the lumberjack" is very similar to *attract*. While the SimLex and SimVerb made major improvements with instructions that are able to lead people to isolate similarity from association when ranking pairs, there still is the issue of context, or lack thereof, when the pairs are presented. In this case, each individual subject has to decide which sense of the word *draw* they will use when rating the pairs. The only way for subjects to consistently interpret verb meanings would be by providing context with each verb pair.

The sub-par performance of WordNet Wu-Palmer path similarity scores on correlation and confusion matrices compared to the human ratings is surprising, especially since Hill et al. referred to it as performing very closely to human ratings. One factor that could have influenced

this is the implementation of synsets; in WordNet, words are organized into synsets. For example, with a word like *hold*, it could link to multiple synsets that refer to different senses and definitions, ranging from zero to over twenty synsets. Due to this, there is no simple way to easily match a word, out of context, to a synset in WordNet. This problem is further exacerbated by the lack of context in the original SimLex or SimVerb ratings. Using the previous example of the pair (draw, attract), since there is no context or method to ensure the human raters interpret the same meaning, on the back-end there is also no way to match the WordNet synsets to their interpreted meaning. WordNet may have the potential to perform close to human performance, but the lack of clear interpretation for each verb pair greatly hampers its performance.

Accurate and precise word and verb similarity quantification methods are not only important for the Zhang et al. study mentioned earlier, but also for tasks such as machine translation and automatic ontology generation. WordNet in particular would be potentially useful for cross-linguistic evaluation, since the Open Multilingual WordNet connects various languages by aligning them to the original English WordNet. SimLex-999, SimVerb-3500, and other gold-standard could not only be improved by further developing their instructions used to tease out similarity from association, but also adding some information or context for the verb pairs to ensure that the subjects correctly interpret *draw* as causing an object to move continuously towards a target instead of as the artistic act of using pencils. As these gold-standards are being further developed, word vector models like Mrkšić et al.'s counter fitted paragram can also benefit from developing more complete retrofitting parameters. One potential parameter would be to construct verb-specific (or other parts of speech specific) retrofitting parameters to improve the model's performance on pairs such as (jerk, prick), where both words can have different similarities depending on their interpretations.

Lastly, models with state-of-the-art performance similar to the counter fitted paragram vectors can also greatly benefit other studies of linguistic performance. Ryskin et al. (2020) looked at linguistic prediction and reevaluated the evidence that relate to the role that executive resources play in linguistic prediction. Their section on computational modelling suggested varying training data to model different populations that have different linguistic experiences, such as L2 learners, older generations, or young learners. With counter fitted word vectors, training data could be manipulated to model these demographics, and moreover, the counter fitting parameters themselves could also be adjusted to reflect the given group's linguistic competency or mapping.

Appendix

Code can be found at: https://github.com/ellissc/evaluating_verb_DSM

The human similarity judgements from SimVerb-3500 are available at their website https://www.repository.cam.ac.uk/handle/1810/264124.

The pretrained counter fitted paragram vectors from Mrkšić et al. are available at their GitHub https://github.com/nmrksic/counter-fitting.

References

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. arXiv preprint arXiv:1608.00869.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013, June). PPDB: The paraphrase database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 758-764).

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4), 665-695.

Markman, A., Wisniewski, E. (1997). Similar and different: The differentiation of basic-level categories.Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1).

Mrkšić, N., Séaghdha, D. O., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P. H., ... & Young, S. (2016). Counter-fitting word vectors to linguistic constraints. arXiv preprint arXiv:1603.00892.

Ryskin, R., Levy, R. P., & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, *136*, 107258.

Zhang, Y., Amatuni, A., Cain, E., Yu, C. (2020). Seeking Meaning: Examining a Cross-situational Solution to Learn Action Verbs Using Human Simulation Paradigm. In S. Denison., M. Mack, Y. Xu, & B.C. Armstrong (Eds.), Proceedings of the 42nd Annual Conference of the Cognitive Science Society (pp. 2854-2860). Cognitive Science Society