Evaluating Verb Similarity Performance of Computational Models

ELLIS CAIN

https://github.com/ellissc/evaluating_verb_DSM

Verb Learning

- Using the Human Simulation Paradigm to study statistical learning of verbs and ambiguous learning situations (Zhang et al., 2020)
  - Adult learners "simulate" being child learners through 1st person videos
  - Collect the labels provided by the participants to get a distribution of guesses
- Evaluation of participant guesses
  - Noun version used binary correct/incorrect marker
  - Verb version used similarity quantification methods like GloVe and WordNet, along with human similarity judgements

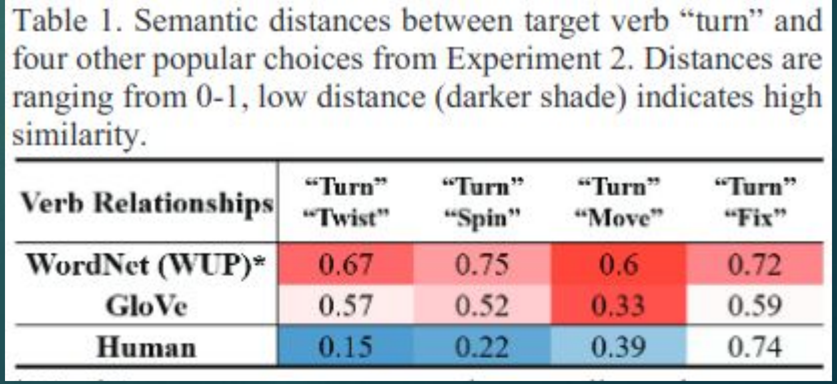


Table 1. Semantic distances between target verb "turn" and four other popular choices from Experiment 2. Distances are ranging from 0-1, low distance (darker shade) indicates high similarity.

Verb Relationships	"Turn" "Twist"	"Turn" "Spin"	"Turn" "Move"	"Turn" "Fix"
WordNet (WUP)*	0.67	0.75	0.6	0.72
GloVe	0.57	0.52	0.33	0.59
Human	0.15	0.22	0.39	0.74

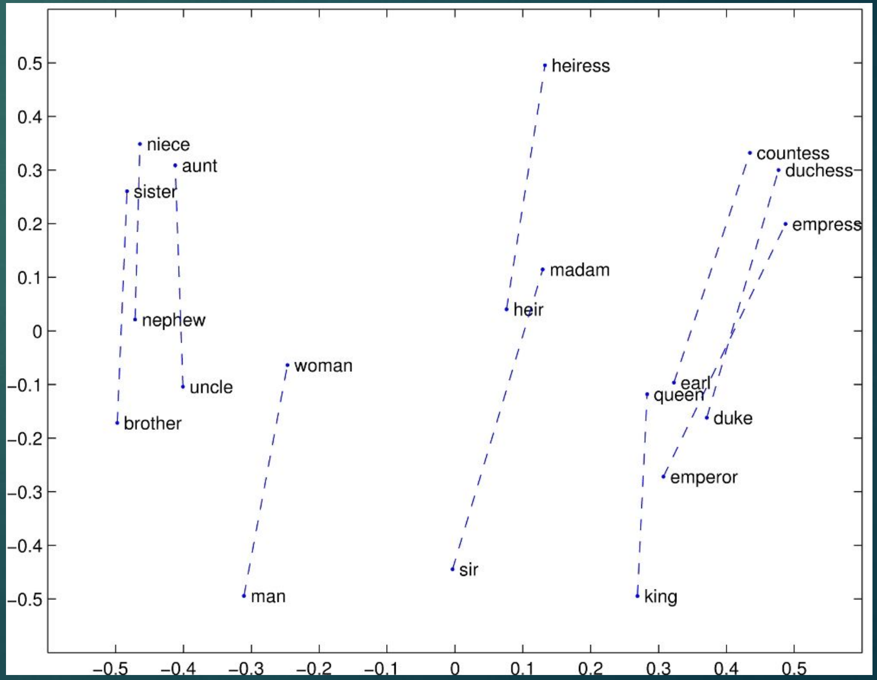
Word similarity evaluation literature

- SimLex-999 (Hill et al., 2014) & SimVerb-3500 (Gerz et al., 2016)
  - Aims to serve as a new gold-standard for evaluating distributional semantic models' performance
  - Contains adjective, verb, and noun concept pairs that vary for concreteness
  - Uses specific instructions to tease out similarity rather than association

- Verbs (and adjectives) are relational concepts (Markman and Wisniewski, 1997), so specific parameters must be implemented to accurately assess similarity

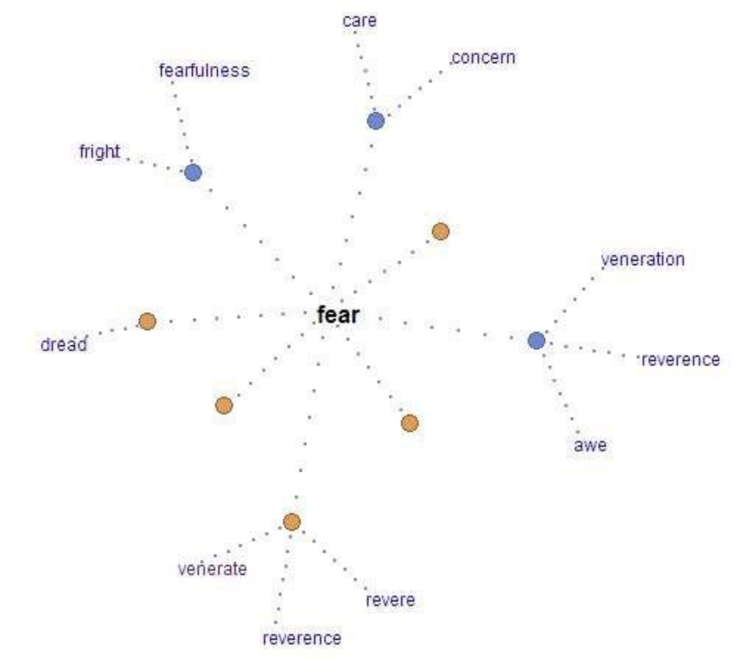
Similarity Models:

- Distributional Semantics model: Counter Fitted Paragram word embeddings (Mrkšić et al., 2016)
  - Uses the Paraphrase database (Ganitkevitch et al., 2013) to learn word vectors (d=300)
  - Paragram vectors are counter-fitted by applying linguistic constraints from the database to improve their quality
  - Similarity measured with cosine similarity distance



Similarity Models

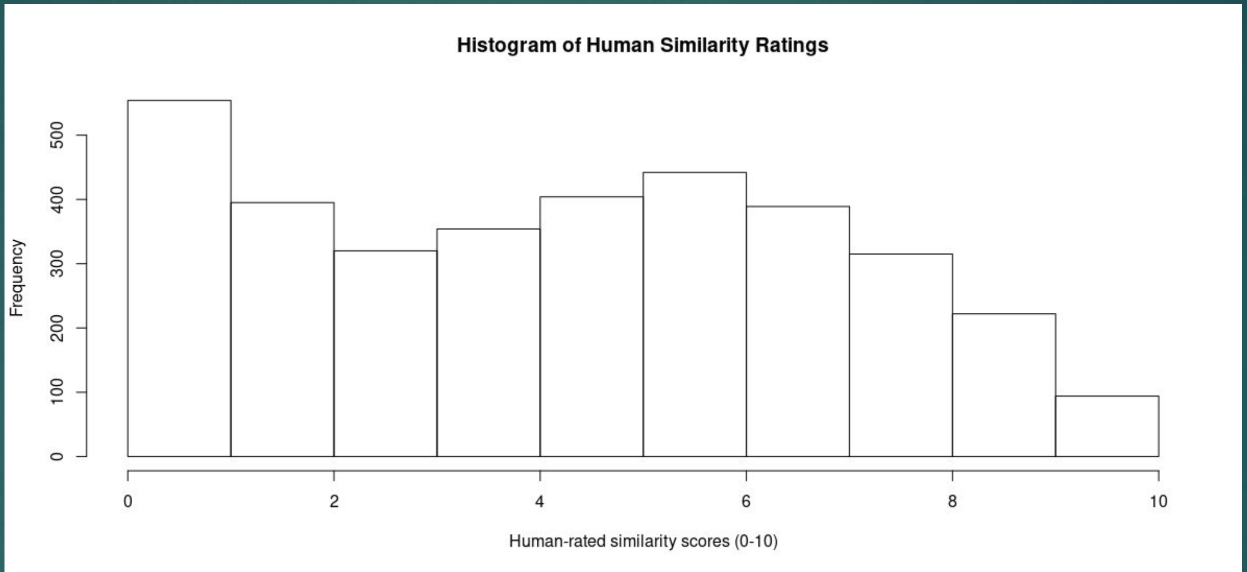
- Lexical database: WordNet
  - Lexical database of English constructed by linguists
  - Synsets: cognitive synonyms, groups of words
  - Similarity measured by jumps
  - Hill et al. claimed WordNet with Wu-Palmer similarity performed very closely to human performance on similarity tasks

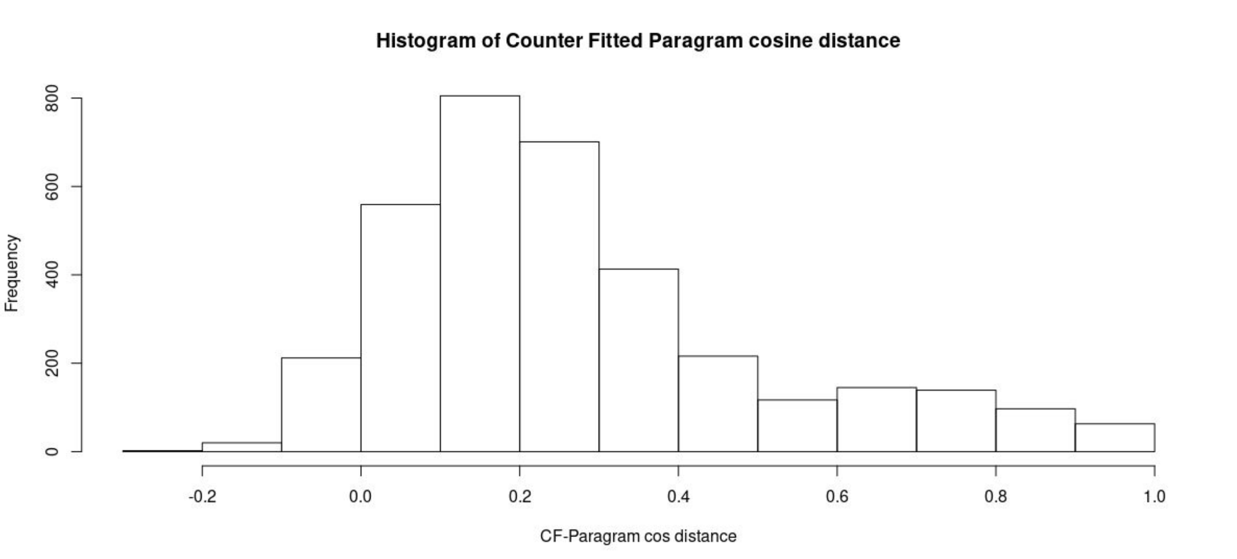


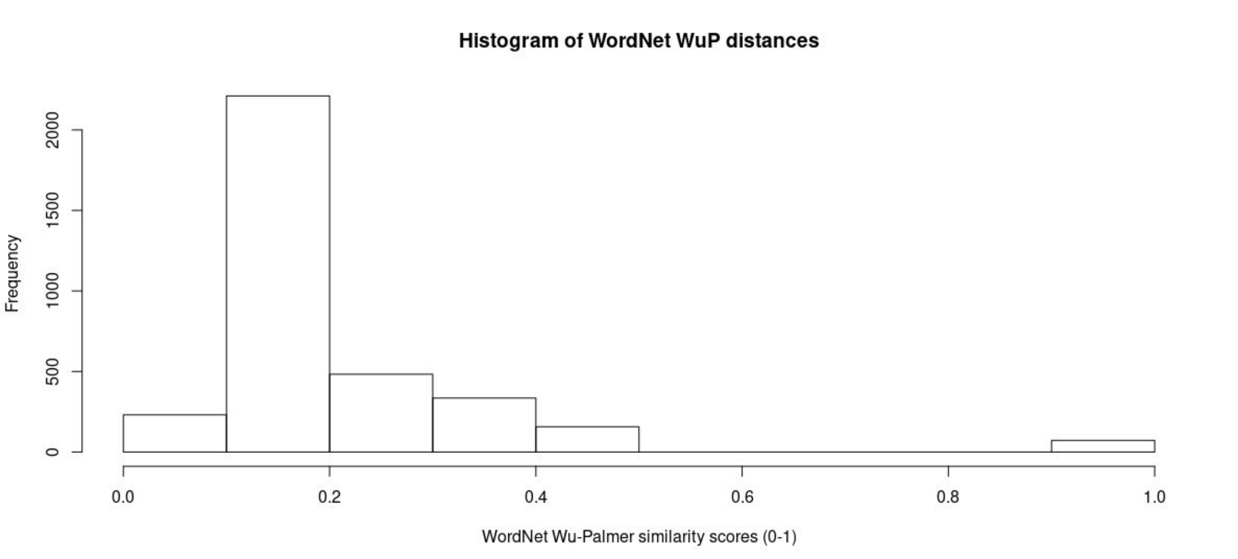
Design

- Verb pairs
  - SimVerb pairs* - selected from USF association norms and VerbNet verb lexicon
- Human judgements
  - SimVerb-3500 similarity rankings
  - *Prolific Academic* online crowdsourcing platform, 843 raters
- Computational Models
  - DSM: Counter-fitted paragrams
  - Lexical: WordNet
- Comparison
  - Visualization: Semantic network
  - Correlation between the human rankings and DSM
  - Confusion matrix

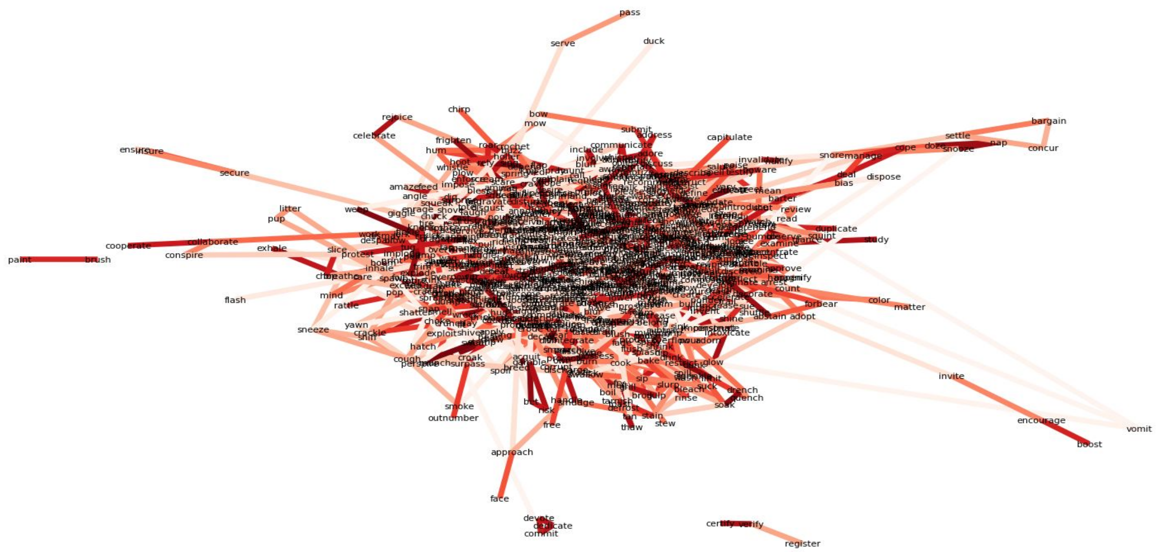
Results - similarity distribution



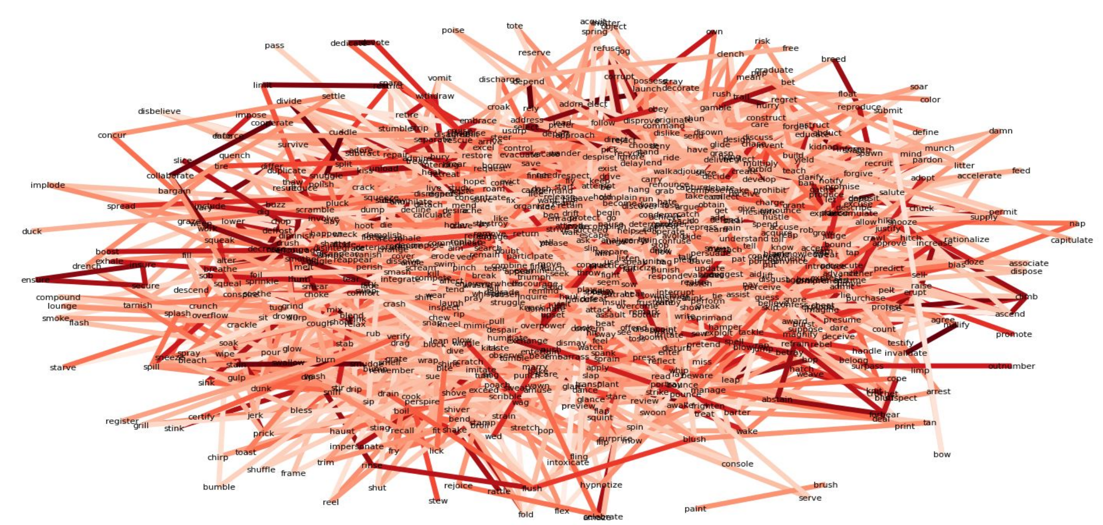




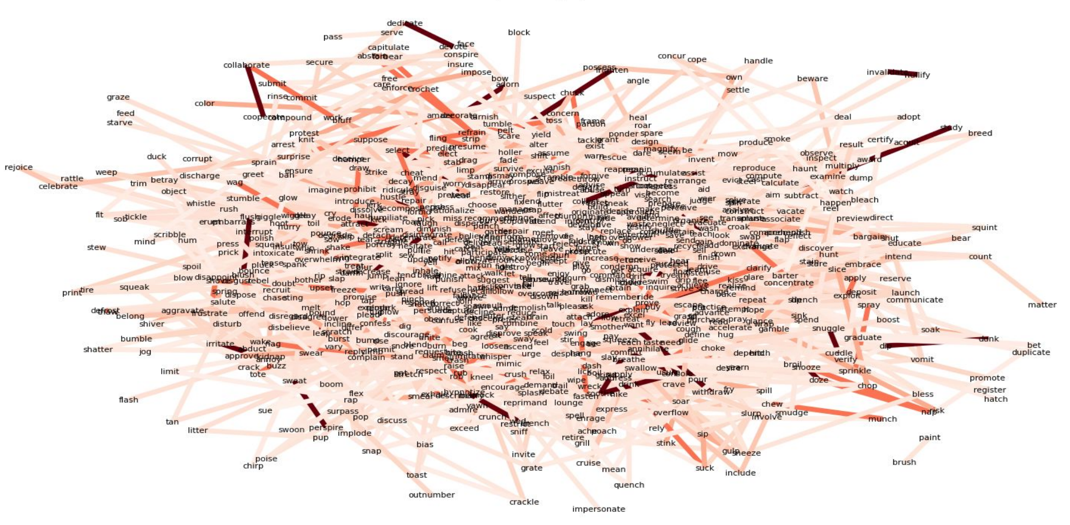
Results - semantic networks



Graph for sv_score

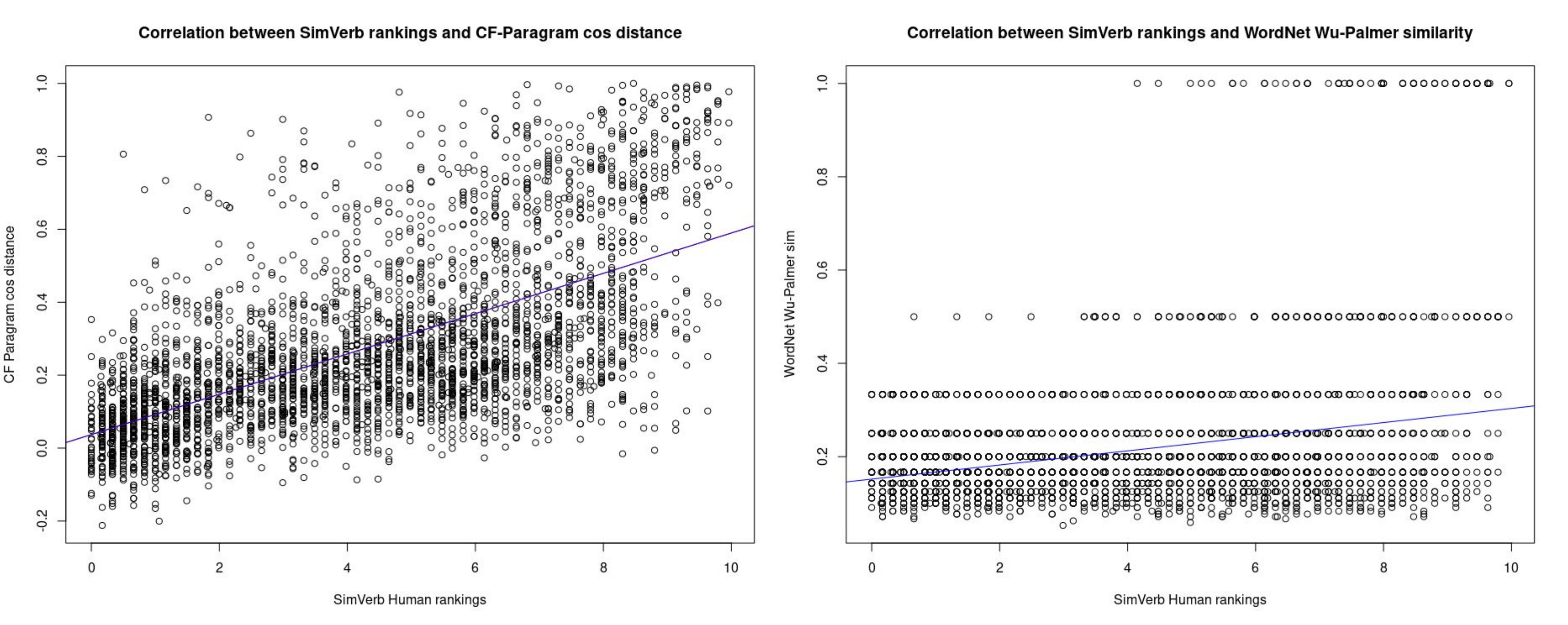


Graph for one_minus_cf



Graph for wn_wup

Results - correlation



Counter Fitted Paragram model (left):

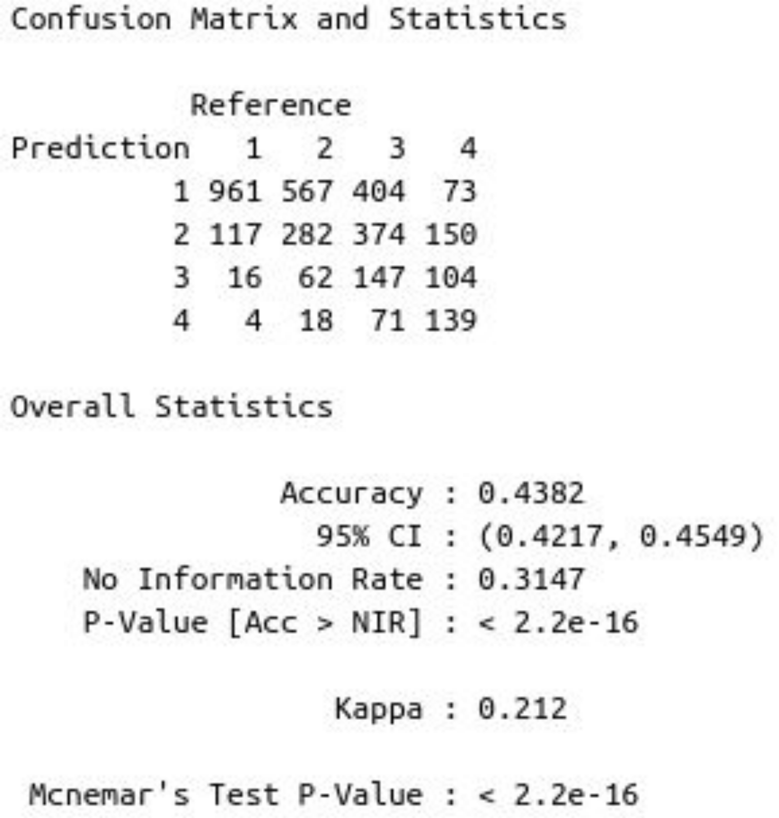
- High correlation of 0.616 between the two measures, t = 46.187, df = 3487, p-value < 2.2e-16
- Linear regression: Slope = 0.05527, intercept = 0.03720

WordNet WuPalmer (right):

- Low correlation of 0.2743 between the two measures, t = 16.846, df = 3487, p-value < 2.2e-16
- Linear regression: Slope = 0.01515, intercept = 0.15195

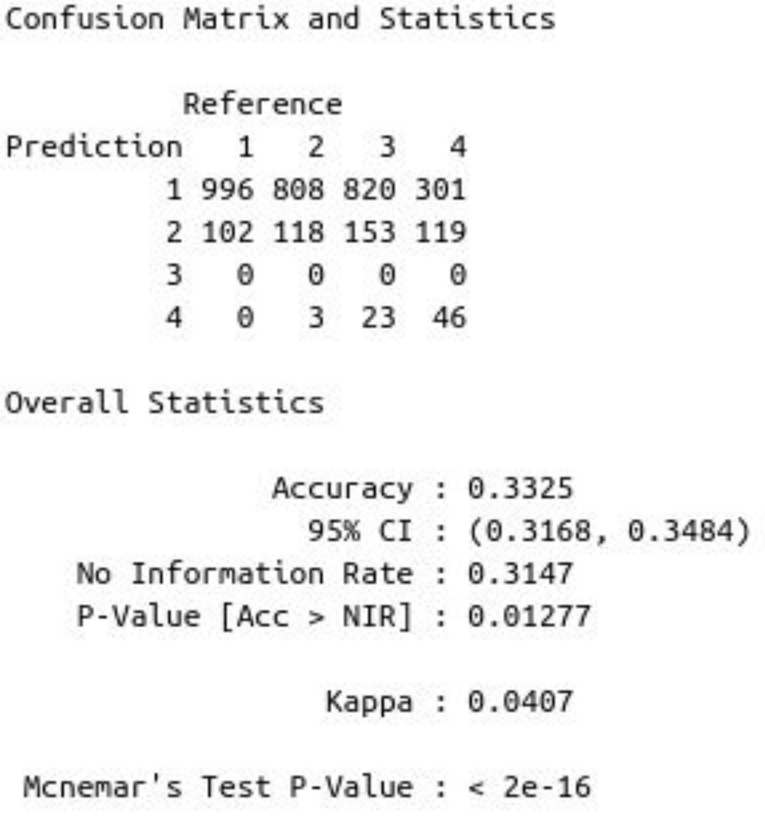
Results - Confusion Matrices

CF Paragrams



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 961 567 404  73
         2 117 282 374 150
         3  16  62 147 104
         4   4  18  71 139

Overall Statistics

               Accuracy : 0.4382
                 95% CI : (0.4217, 0.4549)
    No Information Rate : 0.3147
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.212

 Mcnemar's Test P-Value : < 2.2e-16
```

WordNet WuP



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 996 808 820 301
         2 102 118 153 119
         3   0   0   0   0
         4   0   3  23  46

Overall Statistics

               Accuracy : 0.3325
                 95% CI : (0.3168, 0.3484)
    No Information Rate : 0.3147
    P-Value [Acc > NIR] : 0.01277

                  Kappa : 0.0407

 Mcnemar's Test P-Value : < 2e-16
```

Discussion

- Accurate and precise word/verb similarity quantification methods are not only important for the study mentioned earlier, but also for tasks such as machine translation, automatic ontology generation
- Both correlation test and confusion matrix showed CF paragram model did better than wordnet
- Counter-fitting the paragrams seemed to do a good job of improving performance
  - Major improvement from GloVe
- WordNet Wu-Palmer similarity scores had low correlation to the human ratings even though Hill et al. referred to it as being very close to human ratings
  - Synset selection issue: some verbs have 20+ synsets, hard to automatically select for the 3500 verb pairs accurately
  - Potentially useful for cross-linguistic evaluation

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.

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