EasyChair Preprint

№ 7362

Improving Automatic Semantic Similarity Classification of the PNT

Alexandra Salem, Robert Gale, Gerasimos Fergadiotis and Steven Bedrick

January 21, 2022

Improving Automatic Semantic Similarity Classification of the PNT

Alexandra C. Salem[1,*], Robert Gale[2], Gerasimos Fergadiotis[3], Steven Bedrick[2]
[1]Department of Psychiatry, Oregon Health & Science University, Portland, Oregon, USA
[2]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA
[3]Department of Speech and Hearing Sciences, Portland State University, Portland, Oregon, USA
***corresponding author, salem@ohsu.edu**

Introduction

In the Philadelphia Naming Test (PNT; Roach et al., 1996), paraphasic errors are classified into six major categories according to three dimensions: lexicality, phonological similarity and semantic similarity to the target. Our team has developed software called ParAlg (Paraphasia Algorithms) for automatically classifying paraphasias by these three dimensions given a transcription (Fergadiotis et al, 2016, Mckinney-Bock & Bedrick, 2019). The classifier takes the form of a decision tree mirroring the scoring of the PNT, as illustrated in Figure 1. In ParAlg, the semantic similarity of a response to the target is determined with a binary classifier that uses a language model: a machine learning based model that produces meaningful representations of words in a vector space. Previously, the language model used in ParAlg was word2vec (Mikolov et al. 2013).

This work focuses on improving the semantic similarity classification in ParAlg. We fine-tune a modern language model called BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) alongside a binary classifier to categorize each transcribed response to a PNT item as semantically similar to the target or not. BERT produces contextual vectors, meaning the representation of a word changes based upon the context given to the model, in contrast to the static representations in word2vec. Therefore, BERT may allow for more accurate processing of polysemous words. Finally, we compare ParAlg classification results using word2vec and BERT.

Methods

Our dataset is a subset of the Moss Aphasia Psycholinguistic Database (MAPPD; Mirman et al., 2010) consisting of 11,999 clinician-transcribed and categorized paraphasias from 296 participants (mixed, semantic, abstruse neologism, phonologically-related neologism, formal, other). Errors are classified using ParAlg with word2vec or BERT to make semantic judgments. Performance is evaluated using metrics computed based on the corresponding classification matrices using 5-fold cross validation in order to prevent over-fitting.

Results

Overall, BERT outperformed word2vec when determining the semantic similarity of each error to the target (Table 1, top). Using BERT led to 556 semantic misclassifications compared to 1,084 with word2vec. The downstream effect of these improvements on categorization in the PNT is shown in the bottom of Table 1.

Further, a post-hoc qualitative analysis suggests that BERT's improved performance is associated with its ability to handle polysemy. For example, the most common word2vec error is marking the target/response pair *glass/cup* as semantically dissimilar. This is due to the fact that word2vec has one vector for each word despite polysemy; the closest word to *cup* in word2vec space is *championship*, since it is trained on news data. BERT, however, correctly classifies all 24 of those as semantically similar, since it produces contextual vectors and is able to refine to the appropriate meaning of *cup*.

Conclusions

Changing from word2vec to the contextual language model BERT makes substantial improvements to semantic similarity classification by reducing the number of semantic misclassifications by half. Moreover, BERT corrects a number of particularly "naïve" word2vec mistakes that affect the face validity of the system and may pose a significant implementation barrier for adoption by the clinical community.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805
2. Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic Classification of Five Characteristic Types of Paraphasias. *American Journal of Speech-Language Pathology*, *25*(4S). https://doi.org/10.1044/2016_AJSLP-15-0147
3. McKinney-Bock, K., & Bedrick, S. (2019). Classification of Semantic Paraphasias: Optimization of a Word Embedding Model. *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations For*, 52–62. https://doi.org/10.18653/v1/W19-2007
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546
5. Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. Cognitive Neuropsychology, 27(6), 495–504. https://doi.org/10.1080/02643294.2011.574112
6. Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, *24*, 121–133.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award NIDCD R01DC015999.

Figure 1: PNT decision tree example for target *cat*.

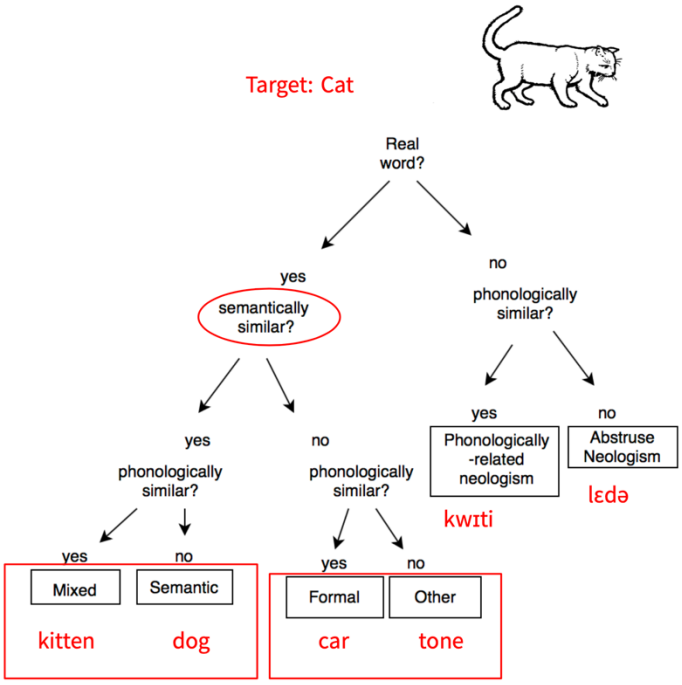

Table 1: Classification accuracy for binary judgements (top) and across the six PNT categories (bottom) for word2vec and BERT. Note that changes in performance are only reflected in the real word categories.

	word2vec				BERT			
	precision	recall	*f1*	accuracy	precision	recall	*f1*	accuracy
Binary Judgements	0.947	0.929	0.938	.910	0.961	0.975	0.968	0.953
Formal	0.789	0.771	0.780	0.910	0.806	0.887	0.845	0.933
Unrelated	0.596	0.748	0.663	0.946	0.670	0.836	0.744	0.959
Mixed	0.609	0.831	0.703	0.930	0.750	0.836	0.790	0.956
Semantic	0.857	0.744	0.797	0.936	0.913	0.796	0.850	0.953
Abstruse Neologism	0.856	0.803	0.829	0.972	0.856	0.803	0.829	0.972
Phonologically-related Neologism	0.956	0.898	0.926	0.947	0.956	0.898	0.926	0.947