ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Refining Semantic Similarity of Paraphasias Using a Contextual Language Model

Alexandra C. Salem,[a] Robert Gale,[a] Marianne Casilio,[b] Mikala Fleegle,[c] Gerasimos Fergadiotis,[c] and Steven Bedrick[a]

[a] Oregon Health & Science University, Portland [b] Vanderbilt University Medical Center, Nashville, TN [c] Portland State University, OR

ARTICLE INFO

Article History:
Received May 15, 2022
Revision received August 19, 2022
Accepted September 15, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Sarah Elizabeth Wallace

https://doi.org/10.1044/2022_JSLHR-22-00277

ABSTRACT

Purpose: ParAlg (Paraphasia Algorithms) is a software that automatically categorizes a person with aphasia's naming error (paraphasia) in relation to its intended target on a picture-naming test. These classifications (based on lexicality as well as semantic, phonological, and morphological similarity to the target) are important for characterizing an individual's word-finding deficits or anomia. In this study, we applied a modern language model called BERT (Bidirectional Encoder Representations from Transformers) as a semantic classifier and evaluated its performance against ParAlg's original word2vec model.
Method: We used a set of 11,999 paraphasias produced during the Philadelphia Naming Test. We trained ParAlg with word2vec or BERT and compared their performance to humans. Finally, we evaluated BERT's performance in terms of word-sense selection and conducted an item-level discrepancy analysis to identify which aspects of semantic similarity are most challenging to classify.
Results: Compared with word2vec, BERT qualitatively reduced word-sense issues and quantitatively reduced semantic classification errors by almost half. A large percentage of errors were attributable to semantic ambiguity. Of the possible semantic similarity subtypes, responses that were associated with or category coordinates of the intended target were most likely to be misclassified by both models and humans alike.
Conclusions: BERT outperforms word2vec as a semantic classifier, partially due to its superior handling of polysemy. This work is an important step for further establishing ParAlg as an accurate assessment tool.

Anomia, or word-finding difficulty, is a hallmark feature of aphasia, a language disorder primarily resulting from stroke (Goodglass & Wingfield, 1997). Anomia is typically assessed with picture-naming tests, which are used by researchers and clinicians alike to characterize deficit profiles, develop treatment plans, and monitor outcomes over time. The Philadelphia Naming Test (PNT; Roach et al., 1996) is a 175-item picture-naming test that offers a classification system for word production errors (paraphasias) based on well-supported models of spoken word production (e.g., Dell, 1986; Levelt et al., 1999). The system defines six major categories of paraphasias (formal, unrelated, mixed, semantic, abstruse neologism, and phonologically related neologism) that require examiners to make judgments along four linguistic dimensions: lexicality, phonological similarity, morphological similarity, and semantic similarity to the intended target word. The types of errors produced are understood to reflect strengths and weaknesses in core subcomponent language processes (i.e., lexical-semantic access, phonological encoding) as well as the degree of overall naming impairment (Dell et al., 1997; M. F. Schwartz & Brecher, 2000) and, as such, provide fine-grained, diagnostically valuable information about the nature of anomia.

Of the four dimensions that underly the PNT's classification scheme, judgments of semantic similarity are inherently the most challenging due to their subjective nature. For example, certain target–response pairs fall somewhere in the liminal space between semantically

Correspondence to Alexandra C. Salem: salem@ohsu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

related and unrelated (e.g., the response "pipe" for the target word "house"), making them particularly challenging to classify and consequently increasing the degree of measurement error in the PNT or other assessment tools reliant on these types of judgments.

Due to its ambiguity, the construct of semantic similarity has been criticized. Medin et al. (1993) reviewed criticisms of similarity, giving the example that "[I]f Mary says that John is similar to Bill, one may have no idea what she means until she adds the observation that they are both avid chess players." That is, a person's life experience and frame of reference affect whether two concepts are similar. Likewise, word-level semantic similarity is not a fixed entity; often, a frame of reference is needed. Chronis and Erk (2020) give the example that the antonyms "black–white" are judged to be maximally dissimilar in isolation but more similar when presented alongside "black–red." In children, variability in semantic judgments can also be influenced by cognitive development, theoretically due to differences in the amount of acquired relational knowledge and maturation of executive functions such as working memory and inhibition control (Lu et al., 2022). Further evidence on the subjective nature of semantic judgments also comes from large corpus studies of human-generated semantic judgments. For example, SimLex-999 (Hill et al., 2015) is a corpus of word pairs, where semantic similarity was rated on a 7-point scale by 500 Mechanical Turk workers. Even with adjusting the scores of 32 raters who were biased in one direction or the other, the reported interrater Spearman correlation between scores was ρ = .67. This correlation was favorable in comparison with other similarity corpora but still indicated the human variability of semantic similarity determination.

Machine learning approaches offer a probabilistic and efficient alternative to human judgments of semantic relationships. One such machine learning approach is the use of a vector space language model, which assigns words in a vocabulary to a point in a high-dimensional metric space in order to facilitate its processing by computers. Many of these models are inspired by a concept from distributional semantics known as the "distributional hypothesis": A word's meaning may be characterized by "the company that it keeps" (Firth, 1957; Harris, 1954), that is, by the words that co-occur in context with the word in question. In language models based on this hypothesis, words that are used in similar contexts will be geometrically "closer together" in the vector space; as a result, words that are closer together in the vector space will presumably be more semantically similar than more distant pairs of words.

Our team has developed a software tool called ParAlg (Paraphasia Algorithms) for automatically classifying paraphasias from the PNT using machine learning methods (Casilio et al., in press; Fergadiotis et al., 2016; McKinney-Bock & Bedrick, 2019). The PNT is labor intensive to score, and thus, the motivation of ParAlg is to automate that process in order to increase its clinical utility. The original ParAlg software used a language model called word2vec (Mikolov, Sutskever, et al., 2013) to determine the semantic similarity between a participant's response and the intended target word.

Although word2vec can accurately represent some semantic (e.g., city/state) and syntactic (e.g., past/present tense) relationships between pairs of words (Mikolov, Chen, et al., 2013), it has its limitations. Notably, word representations in word2vec are static; there is only one vector for the word "seal" despite its multiple meanings. This inflexibility on the part of word2vec likely contributes, at least in part, to instances of semantic misclassification in the context of a clinical tool like ParAlg. For example, Casilio et al. (in press) analyzed the paraphasia classification performance of ParAlg (with word2vec) using two different configurations of transcription input. For their best-performing configuration, further qualitative analyses of human versus algorithm paraphasia misclassifications revealed that the two primary sources of disagreement were over- and under-assignment of semantic similarity. Target–response pairs whose semantic similarity was due to an associative relationship or a category coordinate relationship most often contributed to misclassifications as compared with other semantic relationship subtypes outlined by the PNT (i.e., superordinate, subordinate, synonym, or diminutive). These results indicate that ParAlg would further benefit from a new and improved semantic classifier, particularly one that can better handle associative and categorical relationships.

Some recent language models produce contextual representations: The vector for a word changes depending on the linguistic context in which the word occurs. For example, "cup" in a sentence on baking will have a different vector than "cup" in a sentence on football. One such modern language model is BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019), which has proven successful in Google search, text summarization, question answering, and a variety of other tasks (Devlin et al., 2019; Liu & Lapata, 2019; B. Schwartz, 2020). One reason for its wide applicability is that it is designed to be fine-tuned for a specific task: A classifier is trained on top of BERT in such a way that the produced word vectors are tuned for the task at hand. For example, BERT can be fine-tuned to detect spam e-mails or classify whether a movie review was positive. Here, we fine-tune it to determine semantic similarity, which is outlined in more detail below. Like word2vec, BERT also has been shown to represent semantic relationships. For example, when fine-tuned for the semantic textual similarity benchmark (Cer et al., 2017), a collection of sentence pairs with human-annotated similarity judgments on a 5-point scale, BERT obtained a Spearman rank correlation of ρ = .865

with human-annotated scores, which became the new state-of-the-art score. Importantly, Reif et al. (2019) tested BERT's ability to disambiguate word sense and found that it was able to correctly categorize a word's sense in a sentence with a 0.711 F1 score (a common metric of performance agreement in machine learning on a 0–1 scale), which was higher than the previous state-of-the-art score. Thus, BERT shows promise as a modern alternative to word2vec for use in automatic semantic similarity determination.

The purpose of this study was to compare the performance of BERT and word2vec within the context of ParAlg and PNT paraphasia classification. Specifically, we evaluated binary semantic judgments and downstream paraphasia classification on a variety of metrics to determine which semantic model performed better and to what degree. We hypothesized that updating ParAlg from its pretrained language model word2vec to the contextual and fine-tuned language model BERT would result in fewer semantic classification errors. For this change to be successful, we wanted to see not only an improvement in the accuracy of the system but also a qualitative improvement in the types of errors the system makes. These aims can be summarized in three main research questions: (a) Does BERT improve the automatic semantic similarity classification of paraphasias? (b) Can we attribute the improvement partially to better word-sense disambiguation? (c) How many of ParAlg's semantic similarity misclassifications can be attributable to inconsistency in human judgment rather than language model error, and what subtypes of semantic similarity are most difficult for both humans and the two language models?

Method

Data

Our data set is a subset of the Moss Aphasia Psycholinguistics Project Database (MAPPD; Mirman et al., 2010), which consists of 11,999 single-word paraphasias produced by 296 participants with aphasia. On average, sample participants were 58.8 years old (SD = 13), were 29.9 months postonset (SD = 46.8), and had a range of severity but were overall skewed toward the higher end of the Western Aphasia Battery–Aphasia Quotient (M = 73.3, SD = 17.8, range: 25.2–99.3). Paraphasias were phonemically transcribed in the International Phonetic Alphabet (IPA) and assigned a paraphasia classification by MAPPD human annotators according to the PNT guidelines. Orthographic transcriptions were added to real-word paraphasias, and only the six most common paraphasia classification types (formal, unrelated, mixed, semantic, abstruse neologism, and phonologically related neologism) were included in our data set. Formal, semantic, mixed, and unrelated are all real-word errors that are phonologically similar, semantically similar, both, or neither, respectively, in relation to the target word. The two neologism categories account for nonword paraphasias; phonologically related neologisms are phonologically related to the target word, and abstruse neologisms are not. Due to their lack of lexicality, nonwords or neologisms are not judged or further classified along the semantic dimension. The distributions of each of these six categories and their descriptions are summarized in Table 1.

A representation of the MAPPD data set that we feed into ParAlg is shown in Figure 1. *Target* is the intended target word for a given paraphasia. *Response (orthographic)* is the lexical form of the paraphasia (if it is a real word), and *response (phonemic)* is the transcribed paraphasia in IPA. *Code* is the MAPPD human-annotator classification, which we use as the ground truth in training our models.

ParAlg

The ParAlg software used for these experiments includes four separate classifiers—lexicality, phonological similarity, morphological similarity, and semantic similarity—plus a decision tree for determining which of the six major PNT categories a paraphasia belongs to. The lexicality of a paraphasia (whether or not it was a real word) was determined by whether or not an orthographic transcription existed, followed up by confirming that the word was used frequently enough in a corpus of word usage (SUBTLEX; Brysbaert & New, 2009) to be considered lexical (and not an accidental but archaic real word). Phonological similarity was determined by checking a finite set of rules from the PNT, and morphological similarity

Table 1. Moss Aphasia Psycholinguistics Project Database code distributions and descriptions.

Code	Count	Description
Formal	2,471 (20.593%)	Real word, phonologically similar, not semantically similar
Unrelated	49 (7.076%)	Real word, not phonologically similar, not semantically similar
Semantic	2,031 (16.296%)	Real word, not phonologically similar, semantically similar
Mixed	1,198 (9.984%)	Real word, phonologically similar, semantically similar
P.R. neologism	4,450 (37.086%)	Nonword, phonologically similar
A. neologism	1,000 (8.334%)	Nonword, not phonologically similar

Note. P.R. neologism = phonologically related neologism; A. neologism = abstruse neologism.

Figure 1. Moss Aphasia Psycholinguistics Project Database data set.

Target	Response (orthographic)	Response (phonemic)	Code
lion	tiger	/taɪgɚ/	Mixed
pineapple		/brepfrut/	A. Neologism
pirate	parrot	/pɛɹət/	Formal
plant	flowers	/flaʊɚz/	Mixed
queen	princess		

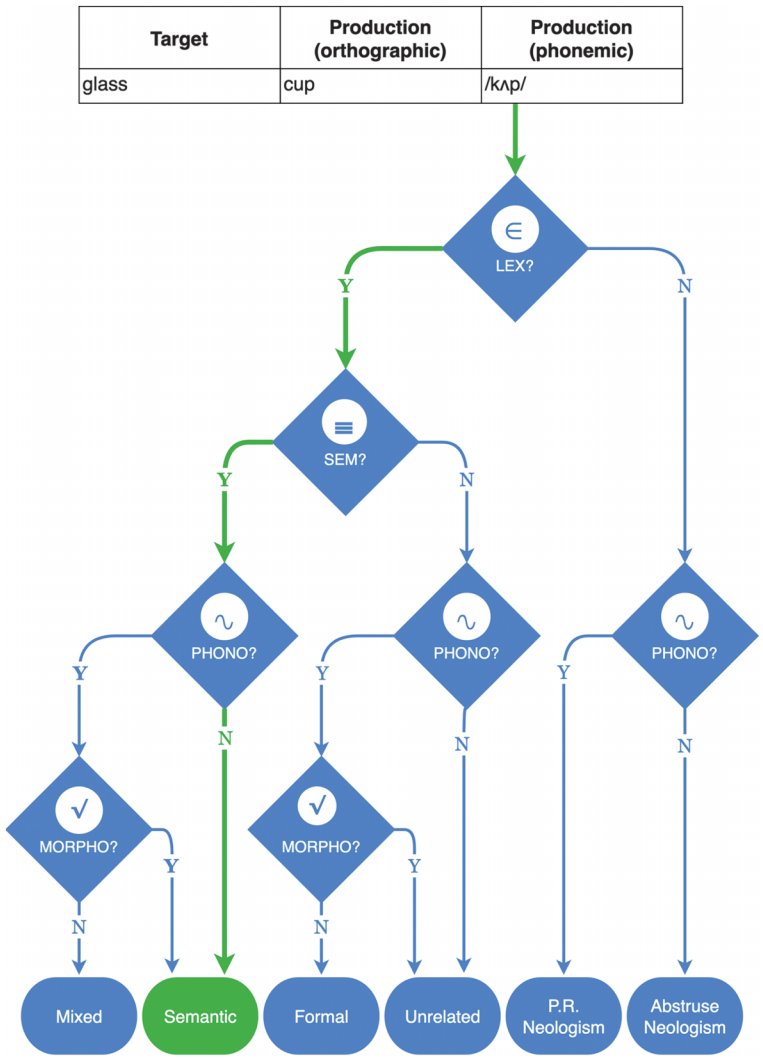
was determined using a corpus of morphology (CELEX; Baayen et al., 1995) and a finite set of rules from the PNT. Finally, semantic similarity was determined with either word2vec or BERT, described in more detail below. More details on the lexical, phonological, and morphological classifiers can be found in our previous work (Casilio et al., in press; Fergadiotis et al., 2016; McKinney-Bock & Bedrick, 2019).

The ParAlg decision tree, which uses these features, is illustrated in Figure 2. The example provided is the response "cup" for the target word "glass." Following the decision tree down, we see that "cup" is a real word; it is semantically similar to "glass," but it is not phonologically similar to "glass," making it a semantic paraphasia.

Language Models

In ParAlg, the semantic similarity of a response to the target is determined with a binary classifier that uses a language model. As mentioned previously, our original

Figure 2. ParAlg (Paraphasia Algorithms) decision tree. LEX = lexicality; SEM = semantic similarity; PHONO = phonological similarity; MORPHO = morphological similarity; P.R. Neologism = phonologically related neologism.



semantic model was a word2vec (Mikolov, Sutskever, et al., 2013) model, closely based on a previously reported version of ParAlg (McKinney-Bock & Bedrick, 2019). Word2vec provides a way of encoding words into numeric vectors to represent their meanings. To build our model, a vector is initialized for each word in the vocabulary and arranged as a matrix of weights for a (very shallow) deep neural network (DNN). The DNN is trained on a large corpus of data composed of a massive amount of text data—in our case, the *New York Times* subset of the Gigaword corpus (Graff & Cieri, 2003) combined with transcripts from the long-running public radio show *This American Life*. We removed from the text a set of about 180 stop words (i.e., words with minimal semantic content, such as determiners and pronouns), and affixes were removed from words in a process called "stemming." The model was trained using the continuous bag of words (CBOW) algorithm. In CBOW, each training sample consists of a target word and a window of N surrounding words. Given the surrounding words, the DNN is trained to predict the target word. Once the DNN has been trained on the entirety of the training data, the weights of the input layer are retained as vector representations for each word in the vocabulary, known as "word embeddings." This article replicated exactly the process described in our previous work (McKinney-Bock & Bedrick, 2019) with one exception: We removed the word "can" from the list of stop words because it is also one of the target words in the PNT. This changed the prediction for only two samples, one positive and one negative, having no effect on metrics. The model was trained with the best configuration from the previous article, with a window size of 1, a vector dimensionality of 750, and a word frequency threshold of 250.

The language model BERT (Devlin et al., 2019) is also a DNN but is designed and trained quite differently from word2vec. BERT is pretrained on a large data set consisting of the BookCorpus (Zhu et al., 2015) and the text from English Wikipedia in a two-step process. First, it trained using "masked language modeling." Here, BERT is given sentences from the corpus where 15% of the tokens are masked (i.e., removed and replaced with a special token [MASK]), and the model attempts to predict what those masked words were. By doing this on the whole corpus of sentences, BERT learns what words co-occur in what contexts. Next, it is pretrained further through "next-sentence prediction." In this process, the model is given two sentences, *Sentence A* and *Sentence B*. These two sentences are either a pair of sequential sentences or two randomly chosen sentences from the corpus. The model attempts to predict whether a given Sentence B followed Sentence A in the corpus. This allows it to capture information about sentence relationships, which the masked language modeling task does not. BERT can be used in the way it was released, as a pretrained model, but one major advantage of it is that it can also be fine-tuned for your specific classification task. That is, a classifier can be trained on top of it in such a way that the vectors BERT produces and the classifier at the end are tuned to the task at hand. In our case, the classifier we trained for fine-tuning was designed to determine when PNT target–response pairs were semantically similar or dissimilar.

The original BERT model was followed by a number of variations that are more efficient to train, are smaller, or have other advantages over the original model. One of these variations is called DistilBERT, which was released by Sanh et al. (2020). DistilBERT is a smaller and faster version of the BERT base model. We chose DistilBERT for this study because it reduces the number of parameters (from 66 million to 110 million) with minimal loss in performance, making it faster and reducing the likelihood of overfitting. It is also important to note that training "from scratch" versions of either BERT or DistilBERT is generally impractical due to the large amount of data and computation time required. As such, we followed standard practices in the field by beginning with a publicly available base model (Sanh et al., 2020), accessible through the open-source HuggingFace library (Wolf et al., 2020), that had been pretrained following the above methods and then fine-tuning from there on the PNT semantic similarity classification task as described. Note that we refer to the model in our experiments as "BERT" for simplicity.

Experiments

We trained ParAlg to classify paraphasias in the MAPPD using either word2vec or BERT for the semantic similarity classification and held all other model features fixed. In these two experiments, we used five-fold cross-validation in order to prevent overfitting. That is, we divided the 11,999 MAPPD items into five groups and trained five separate models for each experiment, in which each one group was held out as testing data. This was done in such a way that, for each of the five iterations, a participant's responses were only in either training data or testing data to prevent the models from learning participant-specific information. The same five-fold splits were used for word2vec and BERT experiments.

For the word2vec experiment, word2vec was pretrained as described in the Language Models section. Once it was trained, vectors for each of the MAPPD real-word responses and corresponding target were pulled out. Cosine similarities were calculated (a measure capturing angular distance) between the target and response vectors, which produced a value between 0 and 1, with higher values representing a higher degree of similarity between the two words. To use this for the semantic similarity classification in ParAlg, a receiver operating characteristic curve analysis (Hanley & McNeil, 1982) was used to

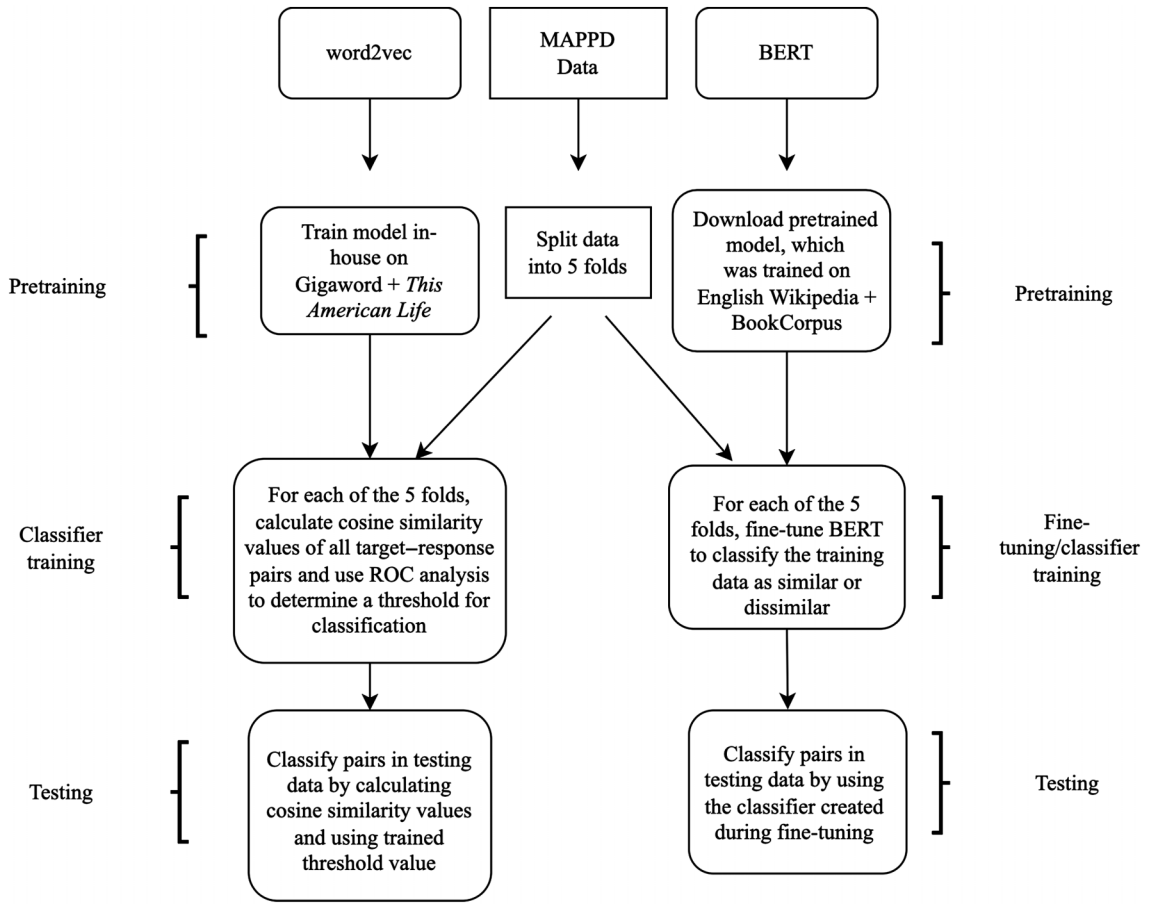
determine a threshold to use for the classifier based upon the cosine similarity values of MAPPD target–response pairs. For each of the five-fold data splits, we calculated the best threshold for classifying the semantic similarity of the training data. Once the optimal threshold for that training data was chosen (e.g., 0.55), then for target–response pairs in the test data, all cosine similarity values above 0.55 would be classified as semantically similar, and those below 0.55 would be classified as dissimilar.

In the BERT experiment, rather than training the model ourselves from scratch and then training a separate classifier, we used the openly available pretrained BERT model and fine-tuned it on the semantic similarity classification task from ParAlg, which produced a classifier for the task. Basing our methods on the BERT approach used by Gale et al. (2021), we gave BERT two sentences, where Sentence A was the target word (e.g., "glass") and Sentence B was the response (e.g., "cup") from the MAPPD. Because this task had similarities to the next-sentence prediction pretraining task described above, we theorized that BERT would be able to learn the semantic similarity relationship between target and response. In each of the five-fold data splits, once BERT was fine-tuned on training data, we then passed the target–response pairs in the testing data through the model to produce a classification result. A diagram summarizing the language model training and experiments is shown in Figure 3.

Evaluation

We evaluated the performance of our improved classifier in several ways. First, we compared classification matrices of binary semantic determination as well as downstream PNT code classification for the two configurations (word2vec and BERT). Then, we calculated several performance metrics on those matrices: positive predictive value, sensitivity, F1 (also known as F-measure), and accuracy, each of which is based on counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These four metrics are defined as follows (for the binary semantic similarity determination): *Positive predictive value* captured how often our

Figure 3. Training and experiment overview. MAPPD = Moss Aphasia Psycholinguistics Project Database; BERT = Bidirectional Encoder Representations from Transformers; ROC = receiver operating characteristic.



categorization as semantically similar was correct (TP/TP + FP), *sensitivity* captured how often we caught pairs that were truly semantically similar (TP/TP + FN), *F1* captured the harmonic mean of positive predictive value and sensitivity (TP/TP + 0.5(FP + FN)), and *accuracy* captured how often our models correctly classified both the semantically similar and dissimilar categories (TP + TN/TP + FP + TN + FN). We also calculated these same metrics for each PNT category. Taken together, these measures capture a wide picture of ParAlg's performance and allow us to weigh the BERT improvements to FP and FN alone and across the different PNT categories.

We then conducted an exploratory qualitative analysis of the geometry of word2vec and BERT vectors of target–response pairs most likely to have word-sense issues. We reduced the data to unique target–response pairs and separated the word2vec and BERT errors into FN and FP. We focused on the FN because we expected that most errors related to polysemy would be instances where a target–response pair was semantically similar, but because the model picked up on the wrong meaning of the target or response, it was incorrectly classified as dissimilar. We defined a BERT improvement as an instance where word2vec incorrectly classified a pair as dissimilar but BERT corrected it to be similar. We counted these BERT improvements for each of the 175 targets. Then, for the three targets with the highest number of FN improvements, for word2vec and BERT, respectively, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten & Hinton, 2008) dimensionality reduction for each of the responses to the target and the target itself to two dimensions and plotted them. t-SNE is a commonly used nonlinear dimensionality reduction technique that works well for high-dimensional data by preserving the local structure of data while revealing important global structure (van der Maaten & Hinton, 2008). Additionally, for each of those three targets, we also calculated the 10 other vocabulary words closest to them in word2vec and BERT space. That is, using the vocabulary we trained word2vec with, we calculated cosine similarity values between each of the 175 targets and all the words in the vocabulary, first using word2vec vectors and then using BERT vectors. Then, for each target, we filtered the 10 highest cosine similarity (so most similar) vocabulary words for each model.

Finally, we conducted an item-level discrepancy analysis of all cases, where BERT's or word2vec's semantic classification yielded a response different from that of human-annotated semantic judgments. Specifically, the 11,999-item data set was distilled to a subset of unique target–response pairs, where one or both models incorrectly classified semantic similarity, via the same process described by Casilio et al. (in press) and described in detail in the Appendix. Then, three research assistants reviewed the unique pairs by (a) judging whether the pair was semantically similar using the target and response orthographic transcriptions exclusively and, (b) if similar, identifying which of the one or more components of the PNT's semantic similarity criteria were met within a given pair (i.e., identifying the semantic similarity subtype). For example, the pair "apple–fruit" would be judged to be semantically similar because the response "fruit" is the superordinate of the target word "apple." The semantic similarity subtypes from the PNT scoring manual are the following: (a) *synonym*; (b) *category coordinate* (target and response share a category); (c) *superordinate*; (d) *subordinate*; (e) *associated*; (f) *diminutive*; (g) *semantically related proper name*; and (h) *shared morpheme*, defined as the addition of a lemma to a monomorphemic target or the addition/substitution of a lemma in a compound target. An exhaustive list of the semantic similarity criteria is freely available on the developer's website (see https://mrri.org/philadelphia-naming-test/).

In alignment with a similar discrepancy analysis conducted by Casilio et al. (in press, rater judgments were used in conjunction with the original MAPPD human-annotator judgment, as well as a given algorithmic judgment (either BERT or word2vec), to categorize discrepancies as being attributable to human-annotated error or algorithmic error. Specifically, discrepant target–response pairs were categorized as MAPPD human-annotated errors (called "Human") if all three research assistants and the algorithmic classification were in alignment with regard to semantic similarity classification but the original MAPPD human-annotated judgment was not. Pairs were categorized as algorithmic errors (called "Algorithm") if the opposite occurred: All research assistants and the original MAPPD human-annotated judgment agreed but the algorithmic classification did not. Pairs where the three research assistants failed to agree on semantic similarity were categorized into a third category, namely, "Uncertain." For both Human and Algorithm errors, the reason for the misclassification was further sorted into one of FP (actually not semantically similar) or FN (actually semantically similar). Then, the FN were sorted into subtypes (associated, superordinate, etc., or *subtype disagreement* if there was no consensus). These results were compared across configurations (word2vec vs. BERT), origins of error (Human error vs. Algorithm error vs. Uncertain), and (when applicable) subtype frequencies.

Results

Aim 1: BERT Improvement Over Word2vec

The classification metrics of the word2vec and BERT experiments are summarized in Table 5. The top

Table 2. Binary semantic similarity classification tables for word2vec and BERT (Bidirectional Encoder Representations from Transformers).

MAPPD classification	word2vec prediction		BERT prediction	
	Not similar	Similar	Not similar	Similar
Not similar	8,170	600	8,506	264
Similar	428	2,801	298	2,931

Note. MAPPD = Moss Aphasia Psycholinguistics Project Database.

row in Table 5 shows the performance of the semantic similarity binary classifier alone. Word2vec obtains a classification accuracy of 0.914, corresponding to 1,028 semantic similarity misclassifications. BERT obtains an accuracy of 0.953, which corresponds to 562 semantic similarity misclassifications. The binary semantic similarity classification matrices for the two experiments are shown in Table 2, where counts for correct predictions (TP and TN) are found on the diagonal and incorrect predictions (FP and FN) are found off the diagonal. BERT reduces the number of FN from 428 to 298 (−130) and the number of FP from 600 to 264 (−336). This change is reflected in an increase in sensitivity (+0.041) and a substantial increase in positive predictive value (+0.093). Overall, the BERT model had more equal weighting of FP and FN, whereas the word2vec model was overly biased toward saying pairs were similar when they were not.

We also see that, in the word2vec experiment, the majority of downstream errors in paraphasia classification into the six categories in ParAlg came from semantic similarity misclassifications. In the word2vec classification matrix (see Table 3), the largest off-diagonal number was 406 errors, where the pair was formal (just phonologically similar) but ParAlg categorized it as mixed (phonologically similar and semantically similar). This was followed by 258 errors, where the pair was semantic (just semantically similar) but ParAlg categorized it as unrelated (neither semantically nor phonologically similar). In contrast, in the BERT configuration classification matrix in Table 4, the largest number of downstream errors was caused by phonological similarity mistakes. The largest off-diagonal number was 205 errors, where pairs were semantic (just semantically similar) but ParAlg categorized them as mixed (semantically similar and phonologically similar). However, this was followed by 157 pairs that were formal (just phonologically similar) but ParAlg categorized as mixed (phonologically similar and semantically similar), indicating that misclassifications at the semantic level using BERT were still causing downstream errors in paraphasia classification.

The performance metrics of the downstream PNT category determination in the two configurations are also shown in Table 5. The largest change in positive predictive value out of the four lexical categories was seen in mixed, which improved from 0.616 to 0.736 (+0.120). This makes sense because word2vec was producing a disproportionate number of FP and, thus, classified a number of paraphasias that were actually formal (just phonologically similar) as mixed (phonologically similar and semantically similar) instead. Likewise, the largest change in sensitivity was seen in the formal category, which improved from 0.790 in word2vec to 0.891 in BERT (+0.101). This reflects the reduction in FN when changing to BERT. That is, word2vec was categorizing more paraphasias as formal (just phonologically similar), which, in actuality, were mixed (phonologically similar and semantically similar). Similarly, the largest change in accuracy was seen in the mixed category with an increase of 0.023, followed by the formal category with an increase of 0.022. The largest change in F1 was actually in the unrelated category (neither

Table 3. Downstream ParAlg (Paraphasia Algorithms) classification matrix using word2vec.

MAPPD classification	ParAlg with word2vec classification						
	Formal	Unrelated	Mixed	Semantic	A. neo	P.R. neo	All
Formal	1,952	61	406	30	0	22	2,471
Unrelated	45	638	10	154	2	0	849
Mixed	114	8	1,006	61	2	7	1,198
Semantic	19	258	210	1,524	15	5	2,031
A. neo	0	0	0	0	914	86	1,000
P.R. neo	0	0	0	0	92	4,358	4,450
All	2,130	965	1,632	1,769	1,025	4,478	11,999

Note. A. neo = abstruse neologism; P.R. neo = phonologically related neologism; MAPPD = Moss Aphasia Psycholinguistics Project Database.

Table 4. Downstream ParAlg (Paraphasia Algorithms) classification matrix using BERT (Bidirectional Encoder Representations from Transformers).

MAPPD classification	ParAlg with BERT classification						
	Formal	Unrelated	Mixed	Semantic	A. neo	P.R. neo	All
Formal	2,201	73	157	18	0	22	2,471
Unrelated	48	710	7	82	2	0	849
Mixed	90	5	1,030	64	2	7	1,198
Semantic	24	150	205	1,632	15	5	2,031
A. neo	0	0	0	0	914	86	1,000
P.R. neo	0	0	0	0	92	4,358	4,450
All	2,363	938	1,399	1,769	1,025	4,478	11,999

Note. A. neo = abstruse neologism; P.R. neo = phonologically related neologism; MAPPD = Moss Aphasia Psycholinguistics Project Database.

phonologically similar nor semantically similar), with an increase from 0.703 to 0.795 (+0.092), primarily reflecting BERT's decrease in FP.

Aim 2: Word-Sense Disambiguation Improvement

We counted how many times BERT corrected word2vec FN for each of the 175 targets, out of all the unique target–response pairs. The three targets for which BERT made the most FN corrections were "seal" (12 corrections), "can" (eight corrections), and "ruler" (seven corrections). Each of these target words is polysemous: "Can" can refer to an aluminum can (the intended target), but it can also refer to the modal verb meaning "be able to." "Seal" can refer to the aquatic mammal (as intended), but it can also refer to a letter embossing or closure. "Ruler" can be the straight-edge object for measuring (as intended), but it can also refer to someone who rules over a country. The 12 FN corrections for "seal," in no particular order, were "tiger," "zebra," "water," "cow," "snail," "sea," "otter," "animal," "monkey," "porcupine," "owl," and "elephant." The eight corrections for "can" were "sauce," "jar," "potato," "applesauce," "potatoes," "soup," "peas," and "pineapple." The seven corrections for "ruler" were "measure," "foot," "scale," "tape," "centimeter," "school," and "yardstick."

To observe patterns in the geometry of word2vec and BERT representations of these targets and responses, we performed t-SNE dimensionality reduction of the vectors of all the responses for "seal," "can," and "ruler" for each configuration. These are shown in Figures 4, 5, and 6, respectively. The FN corrections are responses that were incorrectly classified as dissimilar by word2vec (identified with green italic text with an asterisk in the figures) but correctly classified as similar by BERT (identified with green italic text without an asterisk in the figures).

Ten out of the 12 BERT FN corrections for "seal" were responses that were an animal ("tiger," "zebra," "otter," etc.), and the remaining two were responses "water" and "sea." The t-SNE reduction of BERT responses in Figure 4 shows a clear separated cluster of animal and water-related terms around "seal." Word2vec also clustered some animal terms, but the response "wheel," for example, was just as close to "seal" as "walrus." In Figure 5 of "can," we see little clustering for word2vec. In fact, word2vec classified all responses besides "canned" as dissimilar to "can," as shown by all the dissimilar responses being correct (no asterisk) and all the similar responses being incorrect (asterisk). In contrast, the BERT reduction

Table 5. Performance metrics of binary semantic similarity determination and downstream ParAlg (Paraphasia Algorithms) classification for word2vec and BERT (Bidirectional Encoder Representations from Transformers).

Paraphasia classification	word2vec				BERT			
	Pos pred value	Sens	F1	Acc	Pos pred value	Sens	F1	Acc
Sem +/–	0.824	0.867	0.845	0.914	0.917	0.908	0.913	0.953
Formal	0.916	0.790	0.849	0.942	0.931	0.891	0.911	0.964
Unrelated	0.661	0.751	0.703	0.955	0.757	0.836	0.795	0.969
Mixed	0.616	0.840	0.711	0.932	0.736	0.860	0.793	0.955
Semantic	0.862	0.750	0.802	0.937	0.909	0.804	0.853	0.953
A. neo	0.892	0.914	0.903	0.984	0.892	0.914	0.903	0.984
P.R. neo	0.973	0.979	0.976	0.982	0.973	0.979	0.976	0.982

Note. Pos pred value = positive predictive value; Sens = sensitivity; Acc = accuracy; Sem +/– = binary semantic similarity determination; A. neo = abstruse neologism; P.R. neo = phonologically related neologism.

Figure 4. t-Distributed Stochastic Neighbor Embedding (t-SNE) of word2vec and BERT (Bidirectional Encoder Representations from Transformers) responses to the target word "seal." Responses color-coded red are not semantically similar to the target. Responses color-coded green and in italic typeface are semantically similar to the target. The target word ("seal") is color-coded blue and in all-uppercase typeface. An asterisk (*) indicates that the model (word2-vec or BERT) incorrectly classified that response, and lack of an asterisk indicates that the model was correct. Sem Sim = semantically similar; TARG = target word.

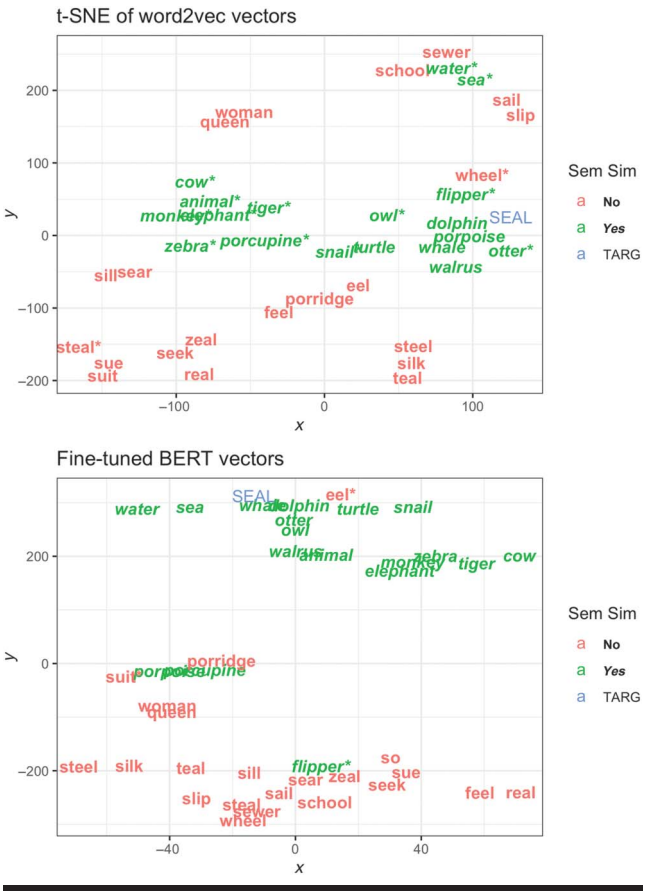
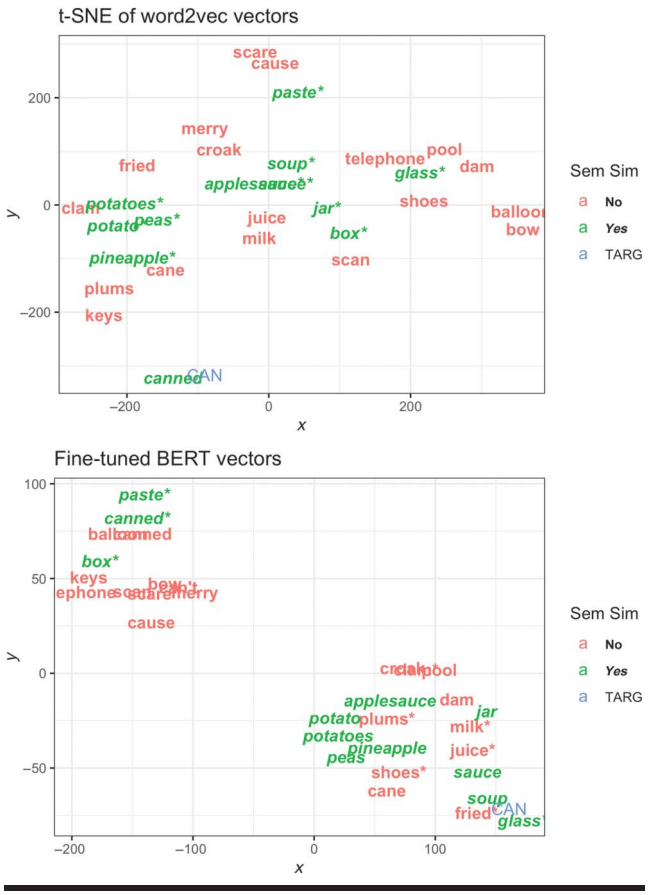


Figure 5. t-Distributed Stochastic Neighbor Embedding (t-SNE) of word2vec and BERT (Bidirectional Encoder Representations from Transformers) responses to the target word "can." Responses color-coded red are not semantically similar to the target. Responses color-coded green and in italic typeface are semantically similar to the target. The target word ("can") is color-coded blue and in all-uppercase typeface. An asterisk (*) indicates that the model (word2-vec or BERT) incorrectly classified that response, and lack of an asterisk indicates that the model was correct. Sem Sim = semantically similar; TARG = target word.



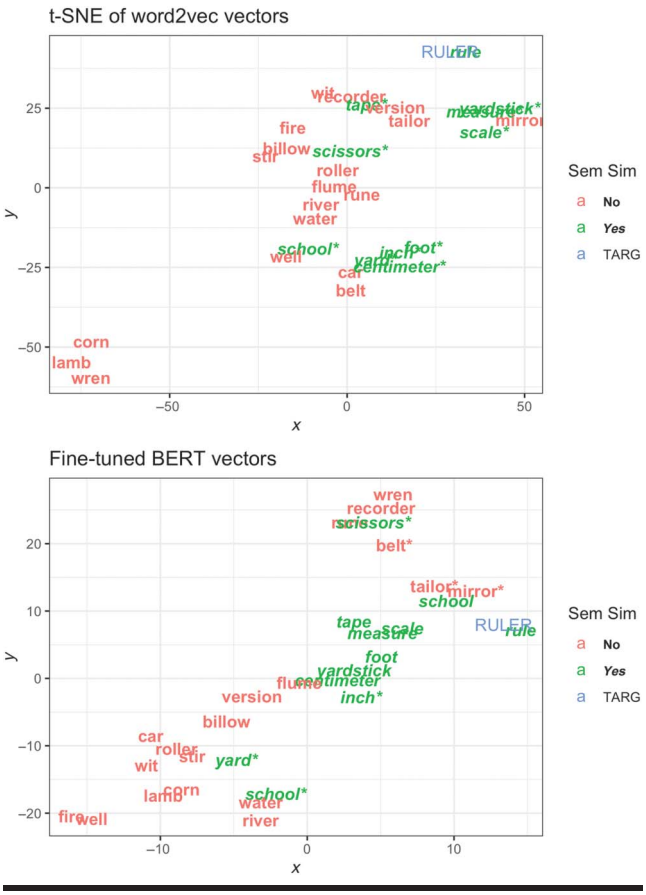
shows two clear clusters of responses. The cluster around the target word "can" consists mostly of food-related responses and includes all the FN corrections BERT made. Neither word2vec nor BERT had very clear clustering for "ruler" in Figure 6; however, there was a little more clustering of similar (green and italic) responses around "ruler" for BERT, including the corrections BERT made such as "tape," "measure," and "scale."

Finally, we calculated the 10 most similar words (out of the entire vocabulary used to train word2vec) to each of these targets for both configurations. These most similar words are listed in Table 6. In BERT space, the top two words most similar to "seal" were ambiguous or capturing the wrong meaning ("seals" and "seales"), but the following three terms ("whale," "dolphin," and "shark") were similar to the "aquatic mammal" meaning. For "can," BERT's most similar words all contained the stem "can"—one of which ("canister") was similar to the intended meaning—whereas words such as "able" and "didn't" were closest in word2vec space. Words most similar to "ruler" were still capturing the wrong sense in both word2vec and BERT space.

Aim 3: Comparison With Human Discrepancies

Of the 1,028 word2vec and 562 BERT semantic discrepancies, 790 and 536, respectively, were unique target–response pairs that were included in the discrepancy analysis, results of which are summarized in Table 7. There were 229 word2vec pairs and 176 BERT pairs categorized as Uncertain, meaning that the research assistants disagreed on whether they were similar pairs or not. In both

Figure 6. t-Distributed Stochastic Neighbor Embedding (t-SNE) of word2vec and BERT (Bidirectional Encoder Representations from Transformers) responses to the target word "ruler." Responses color-coded red are not semantically similar to the target. Responses color-coded green and in italic typeface are semantically similar to the target. The target word ("ruler") is color-coded blue and in all-uppercase typeface. An asterisk (*) indicates that the model (word2vec or BERT) incorrectly classified that response, and lack of an asterisk indicates that the model was correct. Sem Sim = semantically similar; TARG = target word.



experiments, a considerable chunk of errors was definitively due to Human error (60 or 10.7% of word2vec errors and 56 or 15.6% of BERT errors), where the MAPPD human-annotated semantic similarity was different from what the three research assistant raters and ParAlg decided on. This left 501 and 304 Algorithm errors for word2vec and BERT, respectively. There were more FP than FN in both configurations, although it was clear once again that BERT had a better balance of FP and FN (67.1% of errors in word2vec vs. 53.0% of errors in BERT). When further subtyping the FN, in the BERT configuration, the largest subtype was subtype disagreement, where the research assistants agreed that a pair was similar but disagreed on the reason why (meaning these were more difficult pairs). This subtype was followed by associated and then category coordinate. For word2vec, associated was the most common subtype of the FN, followed by subtype disagreement and then category coordinate. After that, the remaining subtypes were sparsely represented.

Human-related errors (i.e., the original MAPPD human-annotated judgment disagreed with ParAlg and the three human raters; 60 word2vec errors and 56 BERT errors) also demonstrated a similar ordering of subtypes for FN, and it is notable that humans were more biased toward FN than toward FP. Out of the FN, associated was the most common subtype, followed by shared morpheme and then category coordinate. No other classification error categories were represented.

Discussion

We compared the performance of two language models, namely, word2vec and BERT, to automatically classify semantic similarity of responses to items in the PNT. We found that BERT outperformed word2vec by reducing the number of semantic similarity misclassifications by almost half. We explored whether BERT's improvement was related to better handling of polysemy

Table 6. Top 10 most similar words to the target words "seal," "can," and "ruler" in word2vec and BERT (Bidirectional Encoder Representations from Transformers) space.

	Seal		Can		Ruler	
Rank	word2vec	BERT	word2vec	BERT	word2vec	BERT
1	reseal	seals	ca	'can	dictatorship	tyrant
2	clinch	seales	could	cana	monarch	emir
3	airtight	whale	'll	canister	strongman	dictator
4	wrap	dolphin	able	cannas	emperor	rulers
5	unseal	shark	didn't	canins	despot	throne
6	weatherstrip	sealant	how/why	canapes	tyrant	governorship
7	watertight	sealys	wo	caning	autocrat	rule
8	cordon	sealer	darndest	'cans	leader	king
9	padlock	seali	doesn't	cani	strongmen	governors
10	sealant	sealy	veeg	cantate	viceroy	queen

Table 7. Breakdown of the types of semantic similarity errors attributable to humans, word2vec, and BERT (Bidirectional Encoder Representations from Transformers) according to the Philadelphia Naming Test.

Category	Category subtype	word2vec		BERT	
Uncertain	**N/A**	**229**		**176**	
		Human $n = 60$	Algorithm $n = 501$	Human $n = 56$	Algorithm $n = 304$
False positive	No relationship	23 (0.383)	336 (0.671)	26 (0.464)	161 (0.530)
False negative	Related proper name	0 (0.000)	2 (0.004)	0 (0.000)	3 (0.010)
	Shared morpheme	5 (0.083)	3 (0.006)	3 (0.053)	7 (0.023)
	Associated	17 (0.283)	65 (0.130)	12 (0.214)	39 (0.128)
	Category coordinate	2 (0.033)	21 (0.042)	4 (0.071)	27 (0.089)
	Diminutives	0 (0.000)	1 (0.002)	0 (0.000)	2 (0.007)
	Subordinate	0 (0.000)	5 (0.010)	0 (0.000)	2 (0.007)
	Superordinate	0 (0.000)	13 (0.026)	0 (0.000)	3 (0.010)
	Synonym	0 (0.000)	2 (0.004)	0 (0.000)	1 (0.003)
	Subtype disagreement	13 (0.217)	53 (0.106)	11 (0.196)	61 (0.200)

Note. Uncertain indicates three research assistants disagreed on the semantic similarity of pairs. These Uncertain pairs have N/A (not applicable) category subtype. False positive represents cases where a model classified a pair as similar but the Moss Aphasia Psycholinguistics Project Database (MAPPD) said it was dissimilar. False negative represents cases where a model classified a pair as dissimilar but the MAPPD said it was similar. Human refers to cases where all three research assistants and ParAlg (Paraphasia Algorithms) agreed on semantic similarity but the original MAPPD human annotator did not. Algorithm refers to cases where all three research assistants and the MAPPD agreed on similarity but ParAlg did not. More information on the subtypes can be found at https://mrri.org/philadelphia-naming-test/.

by plotting vectors of responses to PNT targets, where BERT made the most FN improvements over word2vec, and exploring words closest to those targets in each space. We saw patterns indicating that BERT corrected some word2vec issues with word meaning. Finally, we conducted an item-level review of word2vec and BERT errors. We found that many word2vec and BERT errors were difficult for human annotators as well and that humans and the two language models struggled with similar semantic similarity subtypes.

The improvement in performance from switching to BERT was substantial. Although we saw a large reduction in FP, there was also a reduction in FN, which was reflected in improvements to both positive predictive value and sensitivity of both the binary semantic similarity decision and the downstream ParAlg classification. The BERT model had similar counts of FP and FN, whereas the word2vec model had substantially more FP than FN, in spite of the fact that there were many more semantically dissimilar pairs than similar ones in the data set. Additionally, looking at the downstream ParAlg performance in classifying PNT target–response pairs, semantic similarity was no longer the primary source of errors when switching to BERT, and instead, errors relating to the determination of phonological similarity errors predominated.

The top three targets that saw the largest reduction in FN were polysemous words ("seal," "can," and "ruler"). From examining the t-SNE dimensionality reduction plots, BERT appeared to cluster similar and dissimilar responses to these targets to a greater extent than word2vec did. Moreover, examining the most similar words to each target reiterated that BERT may be disambiguating word sense. In BERT space, at least three of the words most similar to "seal" were related to the meaning intended in the PNT (aquatic mammal), whereas in word2vec space, the most similar words instead related to a different meaning of that target word (adhesive seal). Neither word2vec nor BERT fully picked up on the "aluminum can" meaning of "can," but BERT at least had the correctly similar word "canister." The words closest to "ruler" in both word2vec and BERT space were related to the wrong meaning (country ruler rather than measuring object), but BERT still performed better on that target than word2vec; for example, it correctly identified "foot," "measure," and "yardstick" as semantically related. BERT's improvement of spatial relationships, through clustering of similar and dissimilar words and moving target words closer to other words that capture the PNT's intended meaning, was likely due to the fine-tuning process and helped BERT perform better on polysemous words.

This analysis only addressed improvements to FN, but it is possible some FP improvements (which was the larger source of improvement) were related to word sense as well. For instance, consider the target word "seal." We could imagine a situation where a participant says "peel" as the response, which a language model (which does not know what the target image is) could classify as similar due to its association with, say, peeling tape off a sealed package. Those accidental similarities would be due to random chance of the participant coming up with a word that happens to be similar to a different meaning of the word in the target image and would be difficult to account for scientifically. Still, more investigation is needed. Moreover, the majority of BERT's improved performance came from a reduction in FP. Thus, the entire story of why BERT improves the semantic similarity classification has not been elucidated.

In the analysis of classification discrepancies for word2vec and BERT, we saw that a large number of these pairs for both language models were also difficult for humans. As noted, 229 and 176 unique pairs for word2vec and BERT, respectively, were Uncertain, meaning the three post hoc human raters could not agree if they were similar or not. Moreover, 60 word2vec and 56 BERT unique errors were pairs where the three post hoc human raters and ParAlg disagreed with MAPPD, indicating that they were actually mistakes in the MAPPD data set rather than true errors. Thus, 289 out of 790 unique word2vec errors (36.6%) and 232 out of 536 unique BERT errors (43.3%) were not fairly "errors" but, instead, were inherently ambiguous pairs. Additionally, subtype disagreement (agreement on similarity but not on the subtype) was a common subtype of errors for both humans and ParAlg, though more so for BERT than for word2vec. Moreover, humans and both language models had fairly similar patterns for types of errors. Taken together, these results demonstrate the inherent difficulty of this task and further show that both models (but particularly BERT) made very few true errors and performed remarkably well.

Both word2vec and BERT were prone to overidentification of semantic similarity and similarly failed more frequently at classifying pairs that share the same category (category coordinate) or demonstrate an associated relationship (associated). For example, both BERT and word2vec diverged with human annotators for the target–response pair "star" and "rectangle," both of which share the category shape, and "saw" and "sander," which both belong to the tool category. As with our prior work (Casilio et al., in press), these patterns of error were also observed in human annotators and likely reflect a more universal ambiguity among such pairs, which makes them inherently more subjective to classify, either algorithmically or manually.

Beyond BERT's superior balancing of instances of FP and FN, BERT appears to additionally capitalize on pattern recognition among the target–response pairs within the data set. Specifically, only nine duplicates were present in the instances of semantic discrepancies within BERT, whereas there were 238 duplicates for word2vec. This improvement stems from the fact that BERT can be easily fine-tuned and, thus, can learn to correctly recognize common target–response pairs. The recognition of common patterns such as this is particularly advantageous in the context of scoring a clinical test such as the PNT: Item responses, although open-ended, tend to cluster around a small number of options, and stakeholders (speech-language pathologists, patients) place a high value on consistent and predictable scoring. As such, the strong performance of BERT in this regard, among others, shows its promise for permanent integration into the larger ParAlg system.

With regard to ParAlg, future development will include other elements, such as automatic speech recognition and computer-adaptive testing algorithms, each of which will contribute some measurement error; thus, eliminating as much unreliability in each component will be critical for the precision of the larger system. Here, we have identified one important source of noise within the semantic classification component of the system: inflexible handling of polysemy and suboptimal recognition of patterns on the part of word2vec. Through the use of BERT, we reduce the amount of noise coming from this source, thereby reducing the overall amount of unreliability within the entire system. Additionally, the use of BERT as a component of ParAlg holds promise because it improves the face validity of the system, where *face validity* can be loosely defined as the degree to which an assessment appears effective in capturing the intended construct of interest. This is because BERT, unlike word2vec, yields fewer obvious errors that a human annotator would be unlikely to make. Even though compromised face validity is not a critical psychometric property for an assessment in terms of leading to valid clinical inferences, it can serve as a major implementation barrier preventing acceptance and adoption of the system by the clinical community. As such, optimization of face validity, as can be done with BERT, may increase the likelihood of clinicians adopting ParAlg as part of routine practice.

We have demonstrated that using BERT leads to a highly accurate automatic semantic similarity determination of responses to items on the PNT. However, this work also has clinical applications for naming measures involving semantic similarity more broadly. One such application is the Boston Naming Test (BNT), a widely used picture-naming test consisting of 60 pictures (Kaplan et al., 2001). Two of the error codes on the BNT are a "verbal paraphasia, semantically related to the target word" and a "verbal paraphasia unrelated to the target word." As with the PNT, BERT could be trained to categorize paraphasias according to that code. Another potential application is the Quick Aphasia Battery (QAB; Wilson et al., 2018), which includes a confrontation naming test (based on the BNT) as one of its subtests. In this subtest, paraphasias are ranked on an ordinal scale where semantic relatedness is one of the components. Although both the BNT and the QAB are shorter and less labor-intensive to score than the PNT, there are still potential benefits of automation to make them even more accessible. Moreover, because the BNT and the QAB have simpler classification systems than the exhaustive PNT rules, it is possible that automated scoring could have even higher accuracy on those tests than the PNT. The exact generalizability of this work to other tests such as the BNT and the QAB is an empirical question to be explored in the future.

There are several limitations and many future directions for this work. Although we were able to explore patterns in both language models, FN and FP were difficult

to analyze in either the polysemy analysis or the item-level discrepancy analysis. This work explored only one implementation each of word2vec and BERT, but it is possible that different training methods could improve upon results using either model type. Additionally, the item-level analysis identified a number of errors that are inherently ambiguous; it would be useful to retrain both language models with those pairs removed to see if performance improves.

Overall, fine-tuning BERT leads to a much-improved semantic similarity classifier with high accuracy at 0.953. BERT appears to disambiguate polysemous words more than word2vec, but word sense is still a difficult problem for both language models. Semantic similarity is an inherently subjective task that no human or algorithm could do perfectly, but fine-tuning BERT for the task approaches a degree of performance that is within human range and makes fewer repeated and obvious mistakes. This work is an important step for establishing ParAlg as a useful tool for assessing anomia in aphasia research and clinical practice.

Data Availability Statement

A copy of our 11,999-item Moss Aphasia Psycholinguistics Project Database subset is available as supplemental material in Casilio et al. (in press).

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grant R01DC015999 (principal investigators: Steven Bedrick and Gerasimos Fergadiotis). The authors thank the study participants who donated their time; Adelyn Brecher for her assistance with the interpretation of the Philadelphia Naming Test guidelines; Brooke Cowan for her work in code development; Alex Swiderski for his preliminary extraction of the Moss Aphasia Psycholinguistics Project Database data set; Katy McKinney-Bock and Linying Li for their initial development using BERT (Bidirectional Encoder Representations from Transformers); and Hattie Olson, Khanh Nguyen, Emily Tudorache, and Mia Cywinski for their efforts in reviewing paraphasia discrepancies.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2* [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/GS6S-GM48

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Casilio, M., Fergadiotis, G., Salem, A. C., Gale, R., McKinney-Bock, K., & Bedrick, S. (in press). ParAlg: A paraphasia algorithm for multinomial classification of picture naming errors. *Journal of Speech, Language, and Hearing Research*.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity—Multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 1–14). https://doi.org/10.18653/v1/S17-2001

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 227–244). https://doi.org/10.18653/v1/2020.conll-1.17

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology, 14*(2), 133–142. https://doi.org/10.1080/026870300401513

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology, 25*(4S), S776–S787. https://doi.org/10.1044/2016_AJSLP-15-0147

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis*. Blackwell.

Gale, R., Bird, J., Wang, Y., van Santen, J., Prud'hommeaux, E., Dolata, J., & Asgari, M. (2021). Automated scoring of tablet-administered expressive language tests. *Frontiers in Psychology, 12,* 668401. https://doi.org/10.3389/fpsyg.2021.668401

Goodglass, H. & Wingfield, A. (Eds.). (1997). *Anomia: Neuroanatomical and cognitive correlates*. Academic Press.

Graff, D., & Cieri, C. (2003). *English Gigaword* [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/0Z6Y-Q265

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics, 41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Kaplan, E., Goodglass, H., & Weintraub, S. (Eds.). (2001). *Boston Naming Test* (2nd ed.). Lippincott Williams & Wilkins.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1–38. https://doi.org/10.1017/S0140525X99001776

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3728–3738. https://doi.org/10.18653/v1/D19-1387

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000358

McKinney-Bock, K., & Bedrick, S. (2019). Classification of semantic paraphasias: Optimization of a word embedding model. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 52–62). https://doi.org/10.18653/v1/W19-2007

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

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

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., & Kim, B. (2019). Visualizing and measuring the geometry of BERT. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*. https://proceedings.neurips.cc/paper/2019/hash/159c1ffe5b61b41b3c4d8f4c2150f6c4-Abstract.html

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.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108

Schwartz, B. (2020). *Google: BERT now used on almost every English query*. Search Engine Land. https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193

Schwartz, M. F., & Brecher, A. (2000). A model-driven analysis of severity, response characteristics, and partial recovery in aphasics' picture naming. *Brain and Language, 73*(1), 62–91. https://doi.org/10.1006/brln.2000.2310

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9,* 2579–2605.

Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE, 13*(2), e0192773. https://doi.org/10.1371/journal.pone.0192773

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 19–27). https://doi.org/10.1109/ICCV.2015.11

Appendix

Data Set Preparation for the Item-Level Discrepancy Analysis

The following describes the data set preparation procedure for the item-level discrepancy analysis of this study.

Data Set Preparation

In an effort to reduce coding burden, the 11,999-item paraphasia data set from the MAPPD (Mirman et al., 2010) was reviewed by the third author (M.C.), and all duplicate target–response pairs were identified. Duplicates were operationally defined as pairs that were identical in their (a) target orthographic transcription, (b) response orthographic and phonemic transcription, and (c) human-annotated paraphasia code. Extraneous punctuation (with the exception of diacritics) and capitalization differences were not considered. All duplicate target–response pairs were then removed prior to completing the item-level discrepancy analysis, resulting in a subset of 9,280 unique target–response pairs.