AUTOMATED CLASSIFICATION OF SEMANTIC PARAPHASIAS

Katy McKinney-Bock Steven Bedrick

(Oregon Health and Science University)

1 Introduction

In clinical assessment of people with aphasia, impairment in the ability to recall and produce words for objects (*anomia*) is assessed using a confrontation naming task, where a target stimulus is viewed and a corresponding label is spoken by the participant. Vector space word embedding models have had inital results in automating this task [1]; however, the resulting models are also highly dependent upon training parameters. To select an optimal family of models, we fit a beta regression model to the distribution of performance metrics on a set of 2,880 grid search models and evaluate the resultant first- and second-order effects to explore how parameterization affect model performance. A central methodological question in natural language processing research is how to use extrinsic evaluation to measure what semantic relations are encoded by a model. In this paper, we engage in the interdisciplinary question of how lexical relations can be modeled in a clinical domain, and present an application of word embedding models for assessing semantic impairment.

2 Natural Language Processing and Models of Lexical Semantics

In natural language processing, one approach to modeling a vocabulary of lexical items is to quantify their distributional properties in a large corpus of language data by taking the immediately local context of words around some target word. Then, words that occur in similar contexts/share distributional properties are taken to have some type of semantic similarity. Current methods for quantifying word vectors have roots in Brown clustering [2], Latent Semantic Analysis [3, 4], among others, and the more recent neural network approaches such as the *word2vec* Skipgram architecture rely on a similar underlying distributional hypothesis [5]. Recent research into word embedding models has shown that different hyperparameters used to train models changes the resulting embedding space such that the relationship between word vectors appears to capture different lexical relations. For example, the window size around a target word the immediately local context of words around a target word in a corpus appears to capture different information regarding word association vs. synonymy, as well as functional properties vs. topicality. Word embedding models have been adapted to capture synonymy, association, and hypernymy [6, 7, 8, 9]. Evaluation of these models involves an extrinsic data source, such as a list of word pairs with human ratings of similarity or a list of analogies, and the embedding space compares cosine similarity measures between the word vectors to see whether the embedding space correlates with human ratings.

3 Clinical Databases as Extrinsic Evaluation

The Ahiladelphia Aaming Aest APNT) As A Aonfrontation Aaming task that Aas Aeveloped Aor Asycholinguistic And clinical Aesearch; the Acoring Af this test Anvolves A Aarge taxonomy Af Aoding Aesponses Aased An phonological And Aemantic Aimilarity Af the Aesponse to the target Abject [10]. Ahe taxonomy As Aotivated Ay Dell's two-step Aodel Af Aphasia, Ahere Anomia Aesults Arom A Aisruption An Accessing Aoth the Ahonological representation As Aell As Aemantic Aroperties Af the Abject [11]. Aonventional Acoring Af the ANT Aefines A criteria Aor Aemantic Arrors that Anvolves A Aeal Aord Aoun Aroduction that As An Ane Af Aix Aemantic relations Aith the target Aord; A.g. Aynonymy: AOILET → Acommode"; Aategory Aoordinate: BANANA → Aapple"; superordinate: APPLE → Afruit"; Aubordinate: ALOWER → Arose"; Associated: BENCH

→ "park"; diminutive: DOG → "doggie" [10]. Canonical word embedding tasks used in NLP research strive to model semantic relations that are used in the definition of PNT semantic errors such as synonymy and association (e.g. [6, 7]), and should be well suited for application to identifying/classifying semantic paraphasias in the PNT.

The PNT consists of 175 items, represented by a set of black-and-white images, and selected based on a series of controls, involving varying word frequency based on [12], word length (1 to 4 syllables), and high name performance by control participants [10]. The Moss Aphasia Psycholinguistic Project Database (MAPPD) contains transcribed responses from over 300 administrations of the PNT, and is often used in aphasiological research; in this work, we use a subsample of 152 administrations selected on the basis of clinical characteristics. The frequency and length controls for targets on the PNT, in addition to the relations that define semantic paraphasic errors on the naming test, establish a paradigm for target-production word pairs that is quite similar to the structure of certain external evaluation datasets developed for word embedding models. For example, SimLex-999 [6] is a benchmark dataset that balances word association strength using the USF Free Association norms, samples from both associated and unassociated word pairs, and controls for features such as the concreteness and part-of-speech of the word pairs. Additionally, the PNT involves human evaluation of these semantic relations – in this case, two trained clinicians – with instructions much like SimLex that train evaluators to look for specific dimensions of semantic similiarity when evaluating whether a word pair is semantically similar. Comparing results from MAPPD, which depends on a clinician's identification of a word pair as semantically similar, with results from SimLex-999 is should establish whether clinical data is a reliable evaluation metric for embedding models.

4 Experiment

The current study tests whether model architecture, corpus preparation, and training parameters influence the semantic content of the word embedding model, as measured via the downstream classification task of scoring paraphasic errors on the PNT. We performed a grid search over these sets of parameters, and we evaluate the resultant models on both the PNT dataset as well as the SimLex-999 dataset [6], to evaluate and compare what patterns both evaluation methods find in the data. In doing this, we ask whether the Philadelphia Naming Test can be used as a valid extrinsic evaluation for word embedding models.

Design and Methods: 2,880 word embedding models were trained using Gensim v3.4.0 on the English Gigaword corpus of newsire text, varying the following parameters: the type of model architecture (CBOW vs. Skipgram), corpus preparation (stemming and stopword removal), the size of the symmetrical context window (0-25), *dimensionality* of word embedding vectors (100-750 dimensions), and *minimum word frequency* threshold (100-5000). We evaluated the word embedding models using a semantic classification task for all trials in the MAPPD database. We took the orthographic representation of the visual target item and the produced response to the naming task to be a target-production word pair in the embedding model, and used cosine similarity scores as input to the classifier to determine semantic similarity of target-production pairs in MAPPD. Word pairs involve an out-of-vocabulary word were assigned a similarity score of 0. For all cosine similarity scores for a given grid search model, we calculated the Area Under the Curve for the Receiver Operating Characteristic (AUC for ROC; [13] to determine the best cut-off as to whether a similarity score was considered a semantic paraphasia or a non-semantic paraphasia. We take AUC score as a broad, threshold-independent evaluation of model performance [14] and use this as a criteria for selection of our optimal family of models from the above parameter settings. We used beta regression [15] to model the distribution of the AUC scores from our grid search, and used the resulting coefficients to find optimal settings for each parameter.

Results: For all models, optimal parameters are minimal frequency threshold=100 and maximal dimensions=750. Skipgram models are optimal when the corpus is stopword removed/not stemmed; window size $n=1$. CBOW models are optimal when the corpus is stemmed/stopword-removed. CBOW models are generally optimal with large window sizes; an exception is window $n=1$, where the CBOW models have high performance. Optimization over the SimLex dataset, using Spearman's rank correlation coefficient between human and model scores, shows similar parameter settings as the clinical MAPPD dataset for dimensionality, model type and window size. Key differences in frequency threshold are related to differences in out-of-vocabulary items. Stemming is dispreferred across the SimLex dataset, which differs from the MAPPD CBOW models. As MAPPD utilizes only a limited vocabulary of nouns, the stemmed corpus might have a smaller effect than on the more morphologically varied SimLex word pairs. An additional qualitative investigation related to neighborhood density of the 175 PNT target words across different models results in a very different geometry of the resulting embedding space. Qualitative investigation of the linguistic similarities for different models is in progress, and shows that word sense ambiguities play a role in model performance; we will report results of qualitative investigation as well.

Conclusion: Using beta regression to explore how parameterization affects model performance, we show that performance on MAPPD and SimLex-999 datasets depends on similar optimal parameters. However, results also reveal the importance of further investigation into the geometry of resulting vector spaces and the importance of qualitative linguistic analysis of lexical relations. We demonstrate that the MAPPD dataset, based on a carefully constructed clinical assessment, is useful as an evaluation task for word embedding models and sheds additional insight onto the sensitivity of training parameter selection.

References

[1] Gerasimos Fergadiotis, Kyle Gorman, and Steven Bedrick. Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology*, 25:S776–S787, 2016.

[2] Peter Brown, Peter deSouza, Robert Mercer, Vincent Della Pietra, and Jennifer Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 1992.

[3] S. Deerwester, Furnas G. W. Dumais, S. T., T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[4] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

[6] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[7] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[8] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 302–308, 2015.

[9] Pierre Lison and Andrey Kutuzov. Redefining context windows for word embedding models: An experimental study. *CoRR*, abs/1704.05781, 2017.

[10] A. Roach, M.F. Schwartz, N. Martin, R.S. Grewal, and A. Brecher. The philadelphia naming test: Scoring and rationale. *Clinical Aphasiology*, 24:121–133, 1996.

[11] G. S. Dell. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93:283–321, 1986.

[12] W. Francis and H. Kučera. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin, 1982.

[13] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[14] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[15] S.L.P. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.