Fine-tuning Semantic Vectors With Semantic Fluency Data

Jeffrey C. Zemla (jczemla@syr.edu)
Department of Psychology
Syracuse University, Syracuse, NY

Nichol Castro (nicholca@buffalo.edu)
Department of Communicative Disorders and Sciences
University at Buffalo, Buffalo, NY

Abstract

Semantic vectors derived from training on large text corpora (e.g., word2vec, BERT) are widely used as a methodological tool to model similarity of concepts. Recent work has demonstrated that a small amount of human training data can be used to fine-tune these vectors for modeling specific tasks. For example, human ratings of pairwise similarity can be used to estimate a set of dimensional weights, and these weights can improve estimates of human similarity ratings for held-out pairs. We applied this methodology to the semantic fluency task (listing items from a category) and find that category-specific weights can be used to identify the semantic category of a fluency list. The results have methodological implications for modeling retrieval in semantic fluency tasks, estimating semantic representations, and identifying semantic clusters and switches in fluency data.

Keywords: semantic fluency; semantic spaces; word embeddings; semantic retrieval

Semantic fluency

The semantic fluency task refers to the timed listing of items from a semantic category, such as listing animals for one minute. A robust finding is that similar concepts tend to be clustered together with occasional transitions to dissimilar concepts (Troyer, Moscovitch, & Winocur, 1997). For example one might list *lion*, *tiger*, *cat* (all felines) in sequence before switching to an unrelated animal like *whale*. This has been taken as evidence that associative and controlled search are two components of semantic retrieval in the fluency task (Marko et al., 2023; Shao et al., 2014; Troyer et al., 1997).

Many computational models of the semantic fluency task make use of large semantic spaces derived from text corpora, such as word2vec (Mikolov et al., 2013), in which each concept is represented as a vector. Semantic spaces are widely used as a means to infer when participants move between associative search (clustering; e.g., *cat*, *lion*) and controlled search (switching; e.g., *lion*, *whale*), by using cosine similarity to identify low-similarity transitions (Kumar et al., 2024). The probability of transitioning between two concepts in the fluency task can also be estimated under the assumption that the next response to be recalled is proportional to its similarity to the previous response, i.e., using softmax (Hills et al., 2012).

Large semantic spaces have had a big impact on how people analyze semantic fluency data and model retrieval. One advantage of these spaces is that they can be easily applied to virtually any fluency category (e.g., animals, foods, furniture) and can accommodate nearly any response generated by participants. However, similarity estimates from semantic spaces are noisy and only moderately correlated with human similarity ratings (Richie & Bhatia, 2021). One reason for this is that algorithms like word2vec are trained on text corpora to predict word co-occurrence, but co-occurrence is only a rough proxy for similarity. This is especially true when a word is used in multiple distinct contexts. For example, *dove* and *pigeon* might be rated as similar by humans, but they have low cosine similarity likely because *dove* has many other meanings (e.g., past tense of dive, brand of soap, symbol of peace) and the word *pigeon* is unlikely to co-occur with *dove* in these contexts.

In contrast, classical approaches to analyzing semantic fluency data rely on human raters to either manually identify cluster switches, or to generate large taxonomies that can be used to infer cluster switches (Troyer et al., 1997; Zemla et al., 2020). This approach is very labor intensive and can be biased by individual differences across human coders, but it often provides a better fit to participant-designated switches than semantic spaces (Kumar et al., 2025). This is in part because semantic vectors are generic: each vector encodes a prototype of the word's meaning that collapses over multiple uses (i.e., they lack context).[1]

Here, we describe a method for using semantic fluency data to fine-tune semantic vectors derived from text corpora. Fluency data is used as training data to estimate a set of dimensional weights that are applied to semantic vectors in order to improve estimates of similarity and transition probabilities. One intended application is to estimate and compare dimensional weights across *populations* (e.g., cognitively healthy and dementia), but here we use *fluency category* as a means to validate the method. We estimate a unique set of weights for each of 27 different semantic fluency categories and use cross-validation to predict each fluency list's category (i.e., whichever set of weights maximize the likelihood of the fluency list transitions). We

[1] Recent large language models, particularly those using transformers, have demonstrated that vectors can be adjusted based on context to provide context-specific word meanings. Our goal here is conceptually similar: we aim to fine-tune vectors for the specific context of modeling semantic fluency. However, our methods are very different than large language models.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

predict categories well above chance and find that weighted vectors perform substantially better than unweighted vectors.

We also explore whether weights from different categories are correlated with each other. Additionally, we tested whether these weights can be transferred to another task. Specifically, we test whether the weights derived from fluency data improve estimates of human pairwise similarity ratings, relative to unweighted vectors. However, we did not find convincing evidence of this.

Weighted semantic vectors

The approach of fine-tuning semantic vectors with human training data in order to improve correspondence with human judgment has been applied to many domains (Lu, Wu, & Holyoak, 2019; Peterson, Abbott, & Griffiths, 2018; Richie, Zou, & Bhatia, 2019). A typical procedure is as follows (Richie & Bhatia, 2021): First, participants provide pairwise similarity ratings for a large number of word pairs. The ratings are then regressed on the product of two normalized word vectors to estimate a set of regression coefficients (sometimes using regularization to avoid overfitting). These coefficients act as dimensional weights that transform semantic vectors. Similarity ratings of held-out pairs can then be predicted using cosine similarity on the weighted vectors, outperforming unweighted vectors (Richie & Bhatia, 2021).

A plausible reason for why weighted vectors outperform unweighted vectors is that human similarity ratings are often influenced by a small number of specific features (Navarro & Perfors, 2010), whereas unweighted semantic spaces collapse across many different features and contexts. For example, *lion* and *elephant* are often judged to be similar by humans because they share an ecosystem, even though they are dissimilar in many other ways (e.g., size, color, diet). In contrast, *lion* and *kangaroo* might be judged as less similar, though they are roughly the same size (but do not share an ecosystem). This comparison would suggests that ecosystem is more relevant than size when judging the similarity of animals. Though dimensions of a semantic vector do not map neatly to interpretable features such as size or ecosystem, some dimensions will encode (e.g.) 'size' more than others. The semantic vector weights can be viewed as representing a prototype of the features used by humans to judge similarity for that category, though the weights do not perfectly align human judgments and vector-based similarity.

The weight vectors estimated by Richie and Bhatia (2021) using pairwise similarity ratings are conceptually similar to those that we derive from fluency data. In both our work and theirs, the weights encode the relevance of each dimension for estimating similarity within a category. However, there are several key differences. While Richie and Bhatia (2021) rely on explicit similarity ratings, transitions in fluency data are guided by implicit similarity (i.e., associations that come to mind automatically). The features that guide explicit similarity ratings may not be the same as those that account for implicit mental associations. The probability of generating a response in semantic fluency is also associated with unitary features, such as word frequency in a corpus (De Marco, Blackburn, & Venneri, 2021), and other lexical features like typicality, concreteness, imageability, and age-of-acquisition (Rofes et al., 2020). Additionally, the goal of participants in semantic fluency is to generate as many responses as possible, which encourages exploration. For example, people often cluster pets together when listing animals, but when the rate of listing pets decreases participants will switch to another subcategory that has yet to be explored (Hills et al., 2012; Zemla, Gooding, & Austerweil, 2023).

Model for estimating weights from fluency data

In semantic fluency, participants do not provide explicit ratings of word pair similarity. We propose a model that instead finds the set of weights that maximize the probability of observed transitions in the data. We model similarity between word vectors *i* and *j* as:

$$sim(i,j) = \sum_d w_d * i_d * j_d$$

where *w* is a weight vector estimated from the data, and *d* is the number of dimensions in each semantic vector. When *i* and *j* are L2-normalized, this is equivalent to the cosine similarity of the weighted vectors. In our experiment, we compare normalized and unnormalized approaches.

The transition probability between *i* and *j* is modeled using the softmax function:

$$P(i,j) = \frac{e^{sim(i,j)}}{\sum_k e^{sim(i,k)}}$$

where *k* iterates over all possible known responses (e.g., the set of animals listed in animal fluency, collapsed over all participants). We note two simplifications of this model, which we discuss later: 1) We assume that all transitions are a product of associative search (i.e., all transition probabilities are proportional to similarity), which is unlikely to be the case; 2) We omit a temperature parameter commonly used in softmax to increase or decrease stochasticity of the model.

We estimate a weight vector by minimizing the loss function:

$$L = \sum_{s=1}^{N} \sum_{t=1}^{T_s - 1} -\log P(F_t^s, F_{t+1}^s) + \lambda \sum_d w_d^2$$

where F_t^s denotes the *t*-th response in the fluency list of the *s*-th participant. N is the number of participants and T_s is the

length of the fluency list generated by participant *s*. The left term represents the joint probability of all transitions in the fluency data, and the right term is a ridge penalty to avoid overfitting (λ is a free parameter). Below, we compare a penalized model ($\lambda = 10$) to an unpenalized model ($\lambda = 0$). Weights were estimated using the L-BFGS-B optimization algorithm (Byrd et al., 1995).

Experiment

Dataset and participants

We used data from Castro, Curley, and Hertzog (2021). The original dataset contains data from 246 participants who vary in age (younger, middle-aged, and older adults). Each participant completed a semantic fluency task for each of 70 different categories, such as *a bird*, *a color*, *a musical instrument*, *a sport*, *a fruit*, and others. Some categories resulted in a small average number of responses across participants. Similarly, some participants generated a small average number of responses across categories. We iteratively removed categories that had the lowest average number of responses until we had a sizable number of participants that had at least five responses[2] for each of the remaining categories. Our final dataset consists of 63 participants who generated at least five responses in each of 27 different categories (see Figure 1 for a full list of categories).

These participants had a mean age of 50.0 years (range 20-74). The sample consisted of 51 female and 12 male participants. Three participants had a high school education, 23 had some college or technical training, 26 had a Bachelor's degree, 8 had a Master's degree, and 3 had a PhD. All participants were native English speakers.

Procedure

Participants were recruited from Amazon Mechanical Turk. For each category, participants had thirty seconds to list exemplars from that category. A category label was shown on the screen while participants typed their responses into a text box. After thirty seconds, participants proceeded to the next category.

Data analysis

For an initial semantic space, we used a model pre-trained using word2vec on the roughly one hundred billion word Google News corpus (GoogleNews-vectors-negative300.bin.gz, available from Google). Each semantic vector has 300 dimensions. The decision to use word2vec is arbitrary; we expect that performance would be impacted by both algorithm (e.g., BERT, GloVe) and dataset (e.g., Wikipedia). However, we tested only a single semantic space in order to focus on theoretical aspects of the model presented here, rather than comparing the merits of various semantic spaces.

We estimated separate weight vectors for each of the 27 categories. This was done in two ways: 1) using a leave-one-participant-out design in order to predict the category of each list of the left-out participant (i.e., 27 * 63 = 1701 different weight vectors); and 2) using all available participant data (i.e., 27 different weight vectors) to examine correlations between the weight vectors. We also used unweighted vectors directly from the initial semantic space for comparison.

We repeated this process using four different parameterizations (model types). We estimated a weight vector after first L2-normalizing the initial semantic vectors (i.e., using cosine similarity) and again without normalizing the initial vectors (i.e., using dot-product similarity). We crossed this with a model that used a ridge penalty to reduce overfitting ($\lambda = 10$)[3] and one without a penalty ($\lambda = 0$; i.e., a linear model).

Finally, we tested whether applying weights estimated from fluency data improves the correlation between human similarity ratings and semantic vector similarity, compared to unweighted vectors. For this analysis, we used pairwise similarity ratings collected by Richie and Bhatia (2021) from eight categories that overlap with ours: *birds*, *clothing*, *fruit*, *furniture*, *professions*, *sports*, *vegetables*, and *vehicles*.

Results

Predicting fluency category with cross-validation

We tested four models (leaving-one-participant-out) and all models performed well above chance at (minimum accuracy 23.4% compared to chance performance of 3.6%). The best performing model used dot-product similarity with no ridge penalty (40.5% accuracy), followed by cosine similarity with no ridge penalty (27.3%), dot-product similarity with ridge penalty (26.9%), and cosine similarity with ridge penalty (23.4%).

Results from the best model are presented in Figure 1A (results from other models are available at https://osf.io/esyf9/?view_only=1a30b0d31e0644af8a1db4f28e87ab5a). The best performing weights were typically trained on the test category, with three exceptions: fluency data from *a female first name*, *a type of vehicle,* and *an occupation or profession* were most often predicted to be *a male first name*. Striations in Figure 1A indicate that weights trained on *a male first name* or *a female first name* provide a good fit to many categories, while other category weights are more selective. We omit unweighted vectors from the list of training categories in the figure because the unweighted vectors never provided the best fit across the entire dataset. This may reflect sensitivity to category-general features, such as visual features (e.g., as opposed to relational features).

[2] For participant selection, we counted the number of responses generated in each category. Later, during analyses, we found that a small number of responses were unusable because they had no corresponding semantic vector. The most common reason for this was compound words, e.g., *blue jay* has no vector in the pre-trained semantic space. Transitions to or from words with no corresponding vector were omitted during training and testing.

[3] We made no attempt to optimize this parameter, and chose this value after cursory testing because it appeared to work adequately here and in similar applications (Richie & Bhatia, 2019).

However this was not true for all models tested. For example, the unweighted vectors often provided the best fit in the linear cosine model.

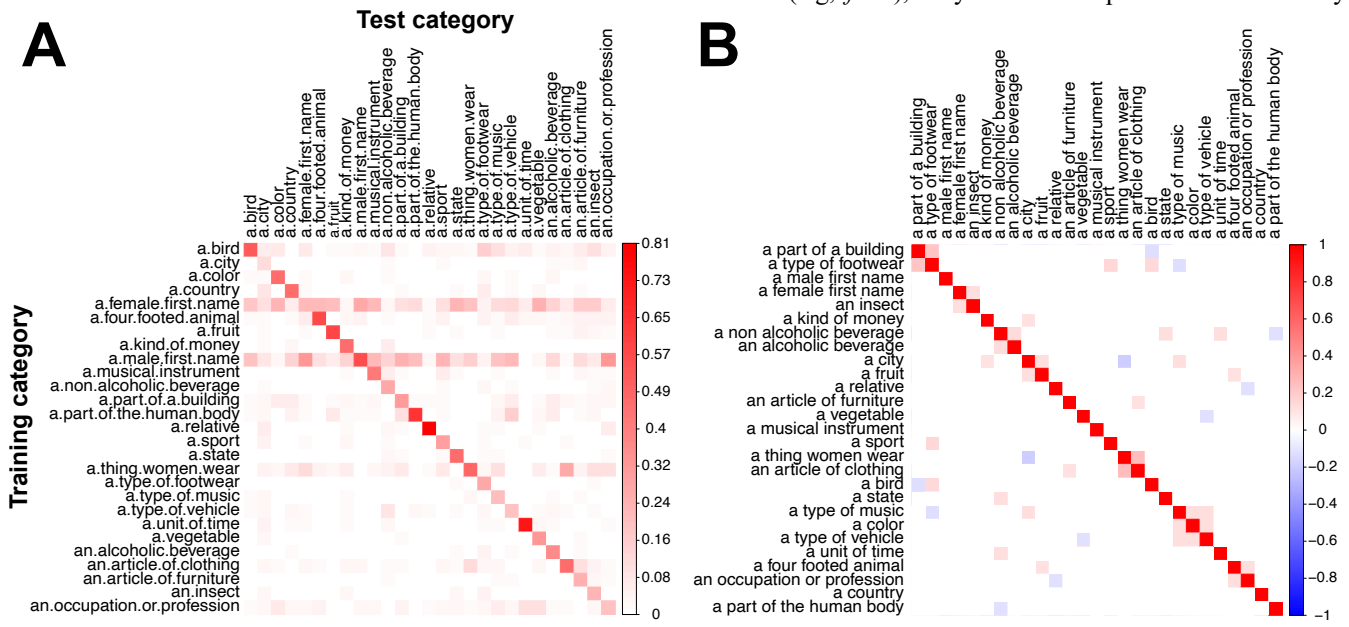

Figure 1. (A) Each cell indicates the proportion of fluency lists from a test category (column) that are predicted to be from the training category (row). (B) A category weight correlation matrix, plotting only correlations where $p < .05$. A hierarchical clustering algorithm is used to group similar weights together.

Similarity of weights across categories

The weight vectors encode the importance of each vector dimension for predicting transitions in fluency data. If the same features are associated with transitions in two categories, then weights for those two categories should be correlated. For example, *an alcoholic beverage* and *a non-alcoholic beverage* might both be clustered by similar taste properties, such as sweetness or bitterness. Figure 1B indicates the correlations for the best-fit model (i.e., dot-product similarity with no ridge penalty), while the other models can be found on https://osf.io/esyf9/?view_only=1a30b0d31e0644af8a1db4f28e87ab5a.

The weight vectors tended to be weakly correlated, at best. However, several correlated weight vectors reflected category similarity (and perhaps an overlapping set of responses). For example, *a non-alcoholic beverage* weights were significantly correlated with *an alcoholic beverage* weights, and *a thing women wear* weights were significantly correlated with *an article of clothing* weights. Models with a ridge penalty produced some higher correlations and (arguably) a more interpretable structure, though most category pairs were still uncorrelated.

Task transfer

We applied the weights derived from fluency data to predict pairwise similarity ratings. Results are shown in Figure 2. The unweighted vectors tended to provide the best fit to human similarity ratings. While weighted vectors occasionally provided modestly better fits than unweighted vectors (e.g, *fruit*), they sometimes provided dramatically worse fits (e.g., *vegetables*). While the weight vectors are conceptually similar to those estimated from human similarity ratings, these results indicate a strong task-dependence that makes them poor for transfer.

Discussion

We describe and implement a model for fine-tuning semantic vectors for the purposes of modeling semantic fluency data. Results from cross-validation indicate that the model weights encode category-specific information about the nature of associations that guide transitions in semantic fluency. When provided a fluency list of an unknown category, the true category is predicted well above chance levels. However, the results also show some degree of category-generalness in the task, in that any weights, regardless of category, tended to predict transitions better than unweighted vectors. This has been previously observed in similar applications (Richie & Bhatia, 2021).

We did not find compelling evidence that these weights can be transferred to a different task (similarity rating), despite both tasks relying on conceptual similarity. In general, the weight vectors estimated from fluency data did not improve the correlation between human similarity ratings and similarity estimates derived from the semantic space. One

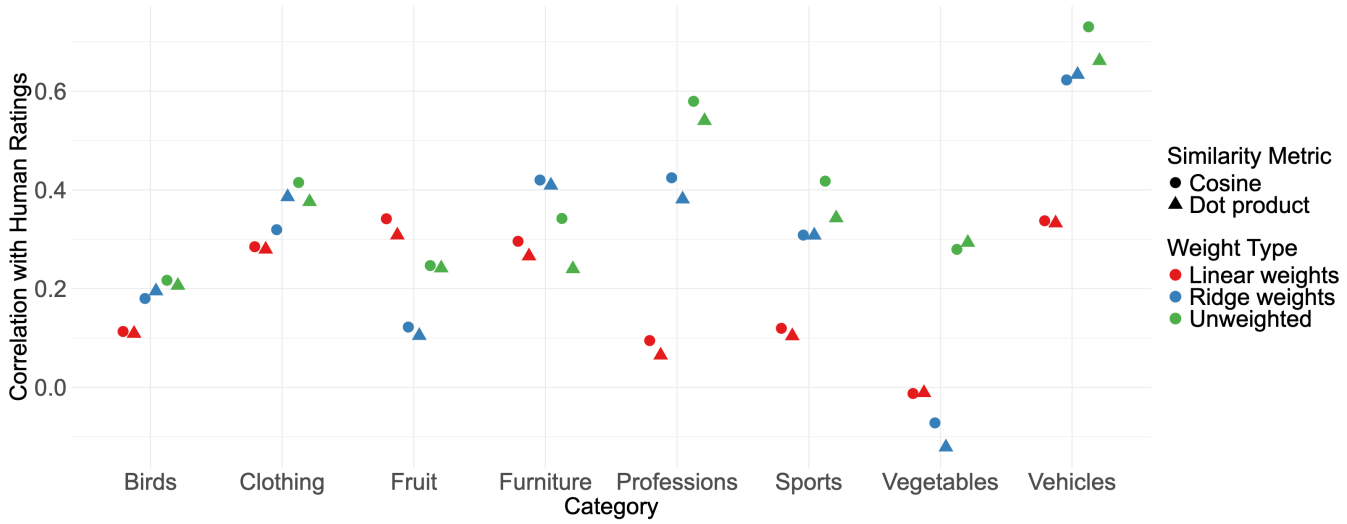

Figure 2. Correlations between human similarity ratings and similarity estimates from each of the four models, plus unweighted vectors.

reason for this may be that associations in the fluency task are implicit: a participant may transition from *dog* to *wolf* without explicitly considering how or whether the two concepts are similar – *wolf* simply pops into mind after retrieving *dog* without effort. In contrast, explicit similarity ratings allow participants to compare concepts on many different features, even if those features are not invoked automatically. For example, one might rate *wolf* and *orca* as similar because they are both pack hunters, but it seems unlikely that this association would come to mind without explicit comparison. Another reason why the weights did not transfer may have to do with the tasks themselves. In particular, similarity ratings are influenced entirely by association, while some fluency transitions reflect non-associative processes (e.g., word frequency).

Differences between models

We tested four models that differed in their similarity function (cosine similarity versus dot-product similarity) and use of ridge penalty. To our surprise, we found that dot-product similarity with no ridge penalty performed substantially better than the next-best model (40.5% accuracy versus 27.3%).

We allowed for a ridge penalty to avoid overfitting, given that the semantic vectors each have 300 dimensions, but this did not improve performance. A plausible reason for this is that we used a single, static value for lambda (penalty weight). There are two potential issues with this. One is that the value was chosen somewhat arbitrarily, though this could be remedied by optimizing lambda. A more difficult issue is that the impact of the penalty term depends on how much data (i.e., the number of fluency transitions) are used to estimate the weights, which varies across all categories and all possible training sets (i.e., 1701 training sets). If lambda is kept constant, then the penalty has less influence as the amount of training data increases and more influence as the amount of training data decreases. Optimizing the penalty term across all possible dataset sizes is beyond the scope of our work.

More surprising is that dot-product similarity outperformed cosine similarity, despite cosine similarity being widely used in the literature. One explanation, though tentative, is that dot-product similarity may account for word frequency effects that occur in fluency data. Cosine similarity is identical to dot-product similarity, except that semantic vectors are first normalized to length one. However, vector length encodes information about word frequency: typically longer vectors are associated with higher word frequency in the training corpus (in this case, Google News) (Wilson & Schakel, 2015). As a result, dot-product similarity sometimes (though not always) increases when substituting a low-frequency word with a highly related high-frequency word. For example, in a cursory test, we calculated the cosine similarity between *camel* and twenty common animals, and then did the same for *dromedary* (a low-frequency term for a type of camel). *Dromedary* had higher cosine similarity for 14 of 20 animals. However, when using dot-product similarity, this dropped to 6 of 20.

In most applications, one would not want similarity estimates to be biased by word frequency. However, transitions in fluency data *are* associated with word frequency (De Marco, Blackburn, & Venneri, 2021; Plant, Webster, & Whitworth, 2011), with high-frequency words being listed more often than low-frequency words (e.g., it is more common for someone to list *cow* than *aardvark* when listing animals). The tendency to list high-frequency words is even more pronounced during early search (Crowe, 1998), which is relevant in our data because participants were given only thirty seconds for each list. Some models of the fluency task explicitly include a process for generating transitions based on frequency, independent of similarity (Hills, Jones, & Todd, 2012). However, the models we describe ostensibly use only similarity to guide transitions. Future work should explore whether it is necessary to include a separate process

for generating non-associative transitions based on word-frequency, or whether the word-frequency information encoded in semantic vector length is sufficient.

Limitations

One limitation is that our model omits a temperature parameter used to control stochasticity in the softmax function. We experimented with a model that includes a temperature parameter and found that including this parameter did not interfere with convergence. However, the estimated temperature values tended to be low (high stochasticity). In addition, when each category is trained separately, temperature parameters can vary. It is an open question as to whether category-specific temperature parameters are psychologically justifiable. Moreover, it complicates the interpretation of cross-validation results, because it is difficult to discern the extent to which predictions are influenced by the semantic weights versus the temperature parameter. As a result, we did not attempt a systematic analyses of a model with a temperature parameter.

Another limitation is that our training and test data may overlap, superficially resulting in better model predictions. While we did not train on the test data specifically, the held-out participant may have transitions in their data that were also generated by other participants. For example, many participants list *dog*, *cat* when listing *four footed animals*, so the weight vector may be both trained and tested on this transition. An alternative approach to cross-validation could entail segmenting the data by transitions (word pairs), rather than by participant.

Future directions

Semantic vectors are commonly used in the fluency literaetur to discern clusters in the data (Hills, Jones, & Todd, 2012; Kumar et al., 2024). One method identifies a cluster switch as a transition of low similarity flanked by transitions of higher similarity (e.g., in the list *lion*, *tiger, mouse*, *rat*, the transition *tiger*, *mouse* is identified as a cluster switch). However, identifying cluster switches with the assistance of human raters is sometimes a more meaningful measure (though more labor intensive) than identifying cluster switches with semantic vectors alone. Our method offers a hybrid solution: semantic vectors are fine-tuned by training on human-generated fluency data, which in turn affects the segmentation of clusters when using vector-based procedures. This has potential as a methodological tool for identifying clusters and switches in fluency data that correlate highly with participant-designated switches, but with lower human labor costs than manual annotation.

However, the method's potential is not merely about convenience or cost. In traditional clustering analyses, semantic associations are assumed to be identical for all participants: what counts as a cluster switch for one participant would count as a cluster switch for any participant. As a result, clustering differences across populations are often assumed to reflect differences in retrieval processes, not semantic organization. One intriguing application of our method is to address this by comparing semantic representations of different populations.

Comparing different populations of participants on the same fluency category is common, for example, older adults versus younger adults (Castro et al., 2021), healthy older adults versus those with Alzheimer's disease (Zemla & Austerweil, 2019), high- versus low-creative individuals (Kenett et al., 2016), and so forth. Our method is easily extendable to estimate weight vectors associated with groups, rather than categories, though we expect that differences between weight vectors from different populations would be less stark than weight vectors between different categories. Our method provides a way to test whether differences in fluency performance can be attributed to differences in semantic association. Some group comparisons are prima facie more likely to show differences than others; for example, it seems plausible that an expert ornithologist will cluster *birds* differently than a novice, because they have knowledge of bird features that a novice does not. Using separate weight vectors for different groups also raises the possibility that differences in clustering and switching that are commonly attributed to retrieval processes might actually reflect a difference in semantic representation. For example, a novice might think an *ostrich* and *chicken* are dissimilar, while an expert might consider them similar (recognizing that these are both flightless birds). For the expert, perhaps these birds belong to the same cluster (high weighted similarity), while for the novice they belong to different clusters (low weighted similarity).

One limitation is that weight vector dimensions are not readily interpretable. In other words, while associative differences can be identified, the weights do not help us to understand which *features* affect theses associations. Future work should explore whether it is possible to use semantic vectors that are more interpretable, for example vectors derived from a feature norm dataset (e.g., McRae et al., 2005).

Finally, our approach applies a weight vector for the purposes of estimating pairwise similarity, but does not transform the entire semantic space all at once. This can be done, but only if the weight vector is constrained to be positive. In that case, multiplying each vector by the square root of the weight vector transforms the entire space while preserving the optimized similarity metric. Such transformations could be useful to compare semantic spaces across populations (e.g., Chan et al., 1993; Sung et al., 2012). It is not clear whether constraining the weight vector to be positive has any effect on the optimization process.

Our work demonstrates that pre-trained semantic vectors can be fine-tuned with human fluency data, and that fluency category can be predicted from these weights. These vectors have a variety of potential applications for modeling semantic fluency data that remain to be tested.

References

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained

optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190-1208.

Castro, N., Curley, T., & Hertzog, C. (2021). Category norms with a cross-sectional sample of adults in the United States: Consideration of cohort, age, and historical effects on semantic categories. *Behavior Research Methods*, *53*, 898-917.

Chan, A. S., Butters, N., Paulsen, J. S., Salmon, D. P., Swenson, M. R., & Maloney, L. T. (1993). An assessment of the semantic network in patients with Alzheimer's disease. *Journal of Cognitive Neuroscience*, *5*(2), 254-261.

Crowe, S. F. (1998). Decrease in performance on the verbal fluency test as a function of time: Evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 391-401.

De Marco, M., Blackburn, D. J., & Venneri, A. (2021). Serial recall order and semantic features of category fluency words to study semantic memory in normal ageing. *Frontiers in Aging Neuroscience*, *13*, 678588.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431-440.

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(4), 377-388.

Kumar, A. A., Apsel, M., Zhang, L., Xing, N., & Jones, M. N. (2024). forager: A Python package and web interface for modeling mental search. *Behavior Research Methods*, *56*(6), 6332-6348.

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2025). What's in my cluster? Evaluating automated clustering methods to understand idiosyncratic search behavior in verbal fluency. *Journal of Memory and Language*, *141*, 104606.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176-4181.

Marko, M., Michalko, D., Dragašek, J., Vančová, Z., Jarčušková, D., & Riečanský, I. (2023). Assessment of automatic and controlled retrieval using verbal fluency tasks. *Assessment*, *30*(7), 2198-2211.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, pp. 3111-3119.

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*(3), 256-268.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648-2669.

Plant, C., Webster, J., & Whitworth, A. (2011). Category norm data and relationships with lexical frequency and typicality within verb semantic categories. *Behavior Research Methods*, *43*(2), 424-440.

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), e13030.

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), 50.

Rofes, A., de Aguiar, V., Jonkers, R., Oh, S. J., DeDe, G., & Sung, J. E. (2020). What drives task performance during animal fluency in people with Alzheimer's disease?. *Frontiers in Psychology*, *11*, 1485.

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 772, 1-10.

Sung, K., Gordon, B., Vannorsdall, T. D., Ledoux, K., Pickett, E. J., Pearlson, G. D., & Schretlen, D. J. (2012). Semantic clustering of category fluency in schizophrenia examined with singular value decomposition. *Journal of the International Neuropsychological Society*, *18*(3), 565-575.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138-146.

Wilson, B. J., & Schakel, A. M. (2015). Controlled experiments for word embeddings. *arXiv preprint arXiv:1510.02675*.

Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The semantic network and fluency utility. *Behavior Research Methods*, *52*, 1681-1699.

Zemla, J. C., & Austerweil, J. L. (2019). Analyzing knowledge retrieval impairments associated with Alzheimer's disease using network analyses. *Complexity*, *2019*(1), 4203158.

Zemla, J. C., Gooding, D. C., & Austerweil, J. L. (2023). Evidence for optimal semantic search throughout adulthood. *Scientific Reports*, *13*(1), 22528.