When Polysemy Matters: Modeling Semantic Categorization with Word Embeddings

Elizabeth Soper and **Jean-Pierre Koenig**
Department of Linguistics
State University of New York at Buffalo
{esoper,jpkoenig}@buffalo.edu

Abstract

Recent work using word embeddings to model semantic categorization have indicated that static models outperform the more recent contextual class of models (Majewska et al., 2021). In this paper, we consider polysemy as a possible confounding factor, comparing sense-level embeddings with previously studied static embeddings on both coarse- and fine-grained categorization tasks. We find that the effect of polysemy depends on how one defines semantic categorization; while sense-level embeddings dramatically outperform static embeddings in predicting coarse-grained categories derived from a word sorting task, they perform approximately equally in predicting fine-grained categories derived from context-free similarity judgments. Our findings highlight the different processes underlying human behavior on different types of semantic tasks.

1 Introduction

A great deal of work has been devoted in recent years to creating computational models of meaning (Landauer and Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019). Such models have been evaluated on a variety of semantic tasks, from word pair similarity judgments to document classification. One task that has received relatively little attention is semantic categorization. Besides making pair-wise judgments about the similarity between two words, humans can also reason about higher-order structures; we can tell not only that *robin* and *sparrow* are similar to each other, for example, but also that they belong in a group with other birds (e.g. *ostrich* and *pigeon*). Based on the impressive performance of embedding models on other semantic tasks, we expect such models to excel at identifying semantic categories as well.

Our particular interest is on the role of polysemy in semantic categorization. Because words generally have multiple distinct senses, categorization decisions will depend on which sense of a word is being considered. Representing the distinct senses of polysemous words, then, should be important to modeling how humans categorize words. For this reason, we expect contextual embeddings, which represent each instance of a word in context as a unique embedding, to model semantic categorization better than static models, which conflate every use of a word into a single representation. But, in fact, recent work evaluating different word embedding models on verb categorization suggests just the opposite; Majewska et al. (2021) found that contextual models perform poorly compared to older static models.

In the following paper, we challenge this result. First, we extend the evaluation from Majewska et al. (2021), who compare word embedding clusters to coarse-grained semantic categories generated by humans in a word sorting task, by evaluating sense-specific embeddings in addition to the static embeddings previously reported. We find that retaining sense-level information from contextual BERT embeddings more than doubles its F1 score, outperforming static embeddings by a large margin. This result suggests that the reported under-performance of BERT in Majewska et al. (2021) was due not to the irrelevance of context to categorization or an inherent weakness of contextual embedding models, but rather to the fact that information about polysemy was thrown away in generating static embeddings from contextual models.

Next, we evaluate the same set of models on fine-grained categorization, using categories derived from human similarity judgments. Contrary to the coarse-grained setting, we find that static and contextual models perform about the same in predicting fine-grained categories. We surmise that humans use different cognitive processes to perform word sorting vs similarity judgment tasks. Choosing the best word embeddings thus depends on the type of behavior one is trying to model.

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 123 - 131
July 14-15, 2022 ©2022 Association for Computational Linguistics

2 Background

Since both static and contextual embeddings have been shown to model pairwise similarity between words well (Pereira et al., 2016; Chronis and Erk, 2020), and since similarity is a primary criterion for categorization, it seems intuitive that word embeddings should perform well at categorization tasks. Some previous work supports this intuition; word embeddings have excelled at word sense disambiguation (Giulianelli et al., 2020; Soler and Apidianaki, 2021; Chronis and Erk, 2020) and topic modeling (Sia et al., 2020; Aharoni and Goldberg, 2020) when cast as categorization problems.

In the present paper, we are interested in semantic category induction. Instead of grouping instances of a word into distinct senses, or documents into topics, the goal of semantic categorization is to group unique words into semantically related clusters. This more abstract type of categorization has received less attention in the word embedding literature; a few probing studies have tested whether different models encode a pre-defined set of categories (Senel et al., 2018; Yaghoobzadeh et al., 2019; Michael et al., 2020), but in all cases these categories were stipulated by the researchers and had not been experimentally validated.

Majewska et al. (2021) recently published a more empirical categorization dataset, based on judgments from non-expert native speakers, rather than stipulated by trained researchers. The dataset, SpA-Verb[1], contains data from two tasks. The first is a sorting task, where participants grouped a set of verbs into broad semantic classes. The second task involves spatial multi-arrangement, which provides finer-grained judgments about the similarity between words within a single semantic domain. SpA-Verb is valuable as an evaluation resource for modeling categorization because it allows for a more direct comparison between human categorization behavior and model behavior than previous datasets. Also, SpA-Verb contains 825 verbs in 17 semantic classes, which is much more comprehensive than other available category datasets.

Most of the verbs in SpA-Verb are polysemous. While many words belong to more than one class (corresponding to distinct senses of those words), the dataset has so far only been used to evaluate static word embeddings (either from static models or extracted static representations from contextual models). Our goal with the following study is to find out when polysemy matters in modeling natural language semantics, in particular, whether sense-specific representations are better predictors of human behavior on some semantic tasks, but not others.

3 Models

Below we describe the word embedding models we evaluate on SpA-Verb:

3.1 Word2vec

The first model we evaluate is a word2vec model trained on part-of-speech-tagged data (Fares et al., 2017). POS tagging allows the static model to distinguish between senses which have different parts of speech (e.g. *duck_NOUN* and *duck_VERB*), although senses which have the same POS are still conflated into a single vector (e.g. *get#ACQUIRE* and *get#UNDERSTAND*). Skip-gram with negative sampling was used to train the model on Gigaword 5th Edition (Parker et al., 2011), with a context window size 5 and 300 dimensions.

3.2 BERT

We evaluate three methods of extracting BERT embeddings: two baseline methods, which create one representation per word form, and a multiprototype method which generates one representation per word sense. For all methods we use BERT Base Uncased from HuggingFace's transformers package (Wolf et al., 2020).

Decontextualized (Decont). First and most simply, we extract embeddings from BERT by feeding each word to the model in isolation. This creates a single, static embedding for each word. This strategy has been used previously as a way to easily extract 'context-free' representations from BERT (Liu et al., 2019; Vulić et al., 2020).

Aggregated (Aggr). Next, we create static embeddings from BERT by averaging a word's embeddings across 100 unique contexts. This aggregated approach still reduces a word to a single representation, but has been shown to produce higher quality representations than the decontextualized strategy (Bommasani et al., 2020).

Multiprototype (MPro). Finally, to test whether sense-specific information is important to semantic categorization, we distill token-level BERT embeddings into multiple prototype embeddings. We use the method of Chronis and Erk (2020) to generate representations which corre-

[1] https://github.com/om304/SpA-Verb

Model	F1-optimal	F1-gold
Random baseline	0.204	0.161
Majewska word2vec	0.355	0.326
Majewska best BERT	0.340	0.322
POS-tagged word2vec	0.442	0.433
Decont. BERT	0.309	0.191
Aggr. BERT	0.398	0.346
MPro BERT	**0.743**	**0.687**

Table 1: Average F1 across models on coarse-grained categories. 'Gold' is for k=17, as in the ground truth. 'Optimal' is best result for k in the range (5, 50).

spond to different senses of a word, without collapsing every token into a single representation (see Appendix A).

3.3 Random Baseline

Finally, we generate random vectors and evaluate them in order establish a baseline for random chance performance.

4 Evaluation

To evaluate the performance of each model on the ground truth classes, *k*-means clustering is used to group verbs into predicted classes. We use the same metrics as Majewska et al. (2021): modified purity and weighted class accuracy are combined in an F1 score, calculated as their balanced harmonic mean. Modified purity is the mean precision of predicted clusters, while weighted class accuracy targets recall (see Appendix B).

Because MPro BERT has multiple representations for a single word, the same word form may show up more than once within a single cluster. To prevent artificially inflating the recall in evaluating MPro BERT, we eliminate duplicates within each cluster before evaluation.

5 Coarse-grained Categorization

Next we describe our evaluation of each model on coarse-grained categorization.

5.1 Dataset

The Phase 1 data of SpA-Verb contains 825 verbs in 17 broad classes (see Appendix C). 116 verbs belong to more than one class. No words were assigned to more than 3 classes.

5.2 Results

Table 1 shows the results of each embedding type, compared to results reported in Majewska et al. (2021). The baseline models (Decont. and Aggr. BERT) perform comparably to previously reported results. POS-sensitive word2vec model scores about 10 points higher than reported for a similar model architecture without POS information. MPro BERT performs dramatically better than other embeddings, achieving more than double the F1 score of the best previously reported BERT results. This suggests that polysemy does play an important role in modeling semantic categorization.

When we look more closely at MPro BERT, we find that embeddings from later layers are better predictors of the ground truth categories than earlier layers (see Appendix D). Interestingly, layer 0 performance is about on par with the static BERT baselines. Earlier layers of BERT have been shown to contain less contextual information than later layers (Ethayarajh, 2019), so this result further supports the idea that contextual information is important to semantic categorization, and that averaging over all contexts or feeding a word in isolation essentially neutralizes the benefit of contextual models over static models for this task.

The benefit of sense-specific embeddings for this task is clear in the example of *freeze*. In the ground truth data, *freeze* belongs to just one class, related to cooking (along with words like *bake, fry, melt,* and *thaw*). *Freeze* has another figurative sense, meaning to stop or suspend. Because the word is polysemous, static embedding clusters struggle to categorize it appropriately. In the aggregated BERT clusters, *freeze* appears in a cluster predominated by verbs related to violence (*whip, shoot, choke, crush, smash*). Decontextualized BERT puts *freeze* in a heterogeneous cluster with a few cooking words (*melt, stew, fry*) but also many seemingly unrelated words (*knit, greet, disturb, wander*). It appears that the different senses of the word skew its static representation and prevent accurate classification. MPro BERT, by contrast, puts *freeze* in two clusters: one related to cooking (as in the ground truth) and another cluster with words like *stop, delay, arrest* and *restrict*, which seems to correspond to the figurative sense of *freeze*. Thus factoring out different senses allows MPro BERT to give a more accurate and reasonable categorization.

MPro BERT tends to capture more distinct senses per word than human participants did, as they generally focused on a single sense when categorizing. On average, each word form appeared in

3.02 MPro BERT clusters, but only in 1.14 ground truth classes. For example, the word form *jump* occurs in one MPro BERT cluster corresponding to violence (*jump#ATTACK*), another cluster corresponding to physical movement (*jump#HOP*), and a third one related to change (*jump#INCREASE*). In the ground truth data, *jump* only occurs once, in a class related to physical movement. Perhaps this is the most salient sense of the word *jump*, and therefore participants were more likely to be thinking of this sense during the word sorting task and ignore its other possible senses. But although the other two senses of *jump* counted against MPro BERT in our evaluation, the fact that embeddings for *jump* were assigned three separate clusters is not necessarily a weakness: the MPro BERT clusters are more thorough as they represent each sense of the word separately and appropriately assign them to separate clusters.

This example demonstrates that F1 scores do not give a full picture of the quality or reasonableness of the word embedding clusters. Categorization is a relatively flexible task; there may be many possible criteria for sorting a group of words, especially when given such a large set of words to sort (Tversky, 1977; Barsalou, 1982). This might explain the low inter-annotator agreement between two initial test participants on Majewska et al. (2021)'s verb sorting task (0.400 B-Cubed score), suggesting that humans don't perform very consistently in creating broad semantic categories from a large group of words. As a result, it's possible for induced categories from word embeddings to be reasonable, but still correlate poorly with our ground truth data.

6 Fine-grained Categorization

Next, we examine how word embeddings fare on finer-grained categories. We speculated that given a smaller, more focused set of words, there is less ambiguity about the relevant criteria for categorizing words, and so evaluating word embeddings on fine-grained categorization may be a better test of model quality than coarse-grained categorization. This section describes how we created a benchmark for fine-grained categorization from the SpA-Verb Phase 2 data, and evaluated the same models on this new benchmark.

6.1 Dataset

In addition to the broad semantic classes created in Phase 1, SpA-Verb also contains Phase 2: a set of fine-grained similarity data from a spatial multi-arrangement task, where participants arranged all words within a single Phase 1 class on a screen according to their relative similarity. The result is a complete matrix of semantic distances for all words within each Phase 1 class. While the original authors use this as resource for evaluating models on standard pair-wise similarity, it can also serve indirectly as a resource for evaluating category structure. In order to use this similarity data to evaluate embedding clusters, we take each row of a class' distance matrix as the vector representation for that word. We run k-means clustering on these representations, and use these clusters as the ground truth to compare with word embedding clusters.

In the fine-grained categorization setting, we assume that only one sense is relevant for each word; the other words in the class implicitly disambiguate between possible senses of a polysemous word, since they were all assigned to a single semantic class in Phase 1. For example, when *stew* occurs in a class with other words related to cooking, the sense of *stew* meaning to worry or fret is not relevant. Since there is only one relevant sense per word for the fine-grained categorization task, in order to evaluate our MPro BERT embeddings in this setting, we need to automatically decide which of a word's sense embeddings is the most relevant given a particular class. To do this, we apply the MAXSIM method used by Chronis and Erk (2020): for each pair of words in a given class, we find the MPro embeddings that yield the highest similarity between the two words. Then, for each word, the prototype that produced the MAXSIM for the most other class members is selected as its most relevant sense, and all other sense embeddings are discarded.

6.2 Results

Table 2 shows the average F1 scores across all 17 classes for each type of embedding. Unlike in the coarse-grained setting, there is not a significant difference between models. Aggregated BERT has a slight advantage with an average F1 of 0.643. All three types of static embeddings do significantly better on fine-grained than coarse-grained categorization. By contrast, F1 for BERT MPro embeddings is 15 points lower in the fine-grained compared to the coarse-grained setting. Furthermore, the opposite pattern appears across BERT layers, with earlier layers performing better and later lay-

Model	Average F1
Random baseline	0.033
word2vec	0.626
BERT decontext.	0.586
BERT aggregated	**0.643**
MPro BERT	0.582

Table 2: Average F1 across all classes for each embedding type on fine-grained categorization.

ers performing worse. It seems that accounting for polysemy makes little difference in the ability of embeddings to identify fine-grained categories.

The ground truth classes with the highest F1 across models were related to sound (*buzz, boom, chirp, rattle*) and physiological processes (*sweat, cough, breathe, yawn*). The classes with the lowest F1 across models were transitive verbs related to physical movement (*drag, fling, tow, throw, lift*) and verbs of communication (*announce, discuss, explain, tell*). In general, smaller and more specific classes were easier to categorize than larger, broader classes (see Appendix E for detailed breakdown of model performance by category).

This stark difference in the relative performance of static and contextual embeddings on two different levels of category granularity is surprising. One possible explanation for this result is that the ground truth for fine-grained categorization was derived from similarity judgment data, and thus may reflect a fundamentally different cognitive process than the coarse-grained ground truth, which came from a sorting task. Phase 2 data was obtained by asking participants to make similarity judgments among a group of words. Our assumption was that since similarity is the primary criteria for categorizing words, similarity data would yield the same categories as a sorting task. However, in the absence of any disambiguating context, participants may have made decisions about similarity based on all exemplars of a word, rather than focusing on one particular sense. By contrast, participants in the Phase 1 sorting task were asked to make explicit category judgments. Categorizing words forces participants to select criteria or features for membership in a particular category. Because of this, participants in the sorting task may have singled out a particular sense of a word in making their decision. Evidence from psycholinguistics supports the idea that human performance on different semantic tasks may derive from very different cognitive processes (Kumar, 2021).

If context-free similarity judgments activate all exemplars of a word, this would explain why static embeddings (in particular the aggregated BERT embeddings, which average over many exemplars) would better fit the Phase 2 data. On the other hand, if semantic categorization activates specific criteria and forces participants to focus on a particular sense of words in making a decision, this would explain why MPro BERT better predicts the Phase 1 data. In order to make a more direct comparison between coarse- and fine-grained categorization, we plan to replicate the Phase 1 sorting task for each individual semantic class.

7 Conclusion

Majewska et al. (2021) found that contextual BERT embeddings performed more poorly than static word2vec on the SpA-Verb semantic categorization benchmark. In this paper, we challenged their analysis, testing the effect of sense-specific contextual information on model performance on two different levels of category granularity, and find that the rich sense-specific information contained in BERT, if properly exploited, allows BERT to excel in predicting coarse-grained human semantic categories. Our results suggest that polysemy affects coarse-grained categorization, and that accounting for polysemy can significantly improve the predictions of embedding models.

On the other hand, contextual information seems to be less relevant in modeling finer-grained categories derived from similarity judgments. It seems that humans rely on different underlying processes in making context-free similarity judgments between words than when making decisions about category membership. While similarity is judged based on a summary of all of a word's exemplars, categorization requires choosing specific criteria for membership and thus focuses attention on a particular sense of a word.

While using sense-specific embeddings seems best for performing category induction, static representations are still desirable for some applications. For example, in making a cross-linguistic or historical comparison of word meanings, clustering average representations may be more appropriate than many sense-specific ones. Ultimately, both types of behavior are of interest within NLP, but it's important to choose an approach carefully, by considering exactly what type of behavior one is trying to model.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763.

Lawrence W Barsalou. 1982. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10(1):82–93.

BNC Consortium. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Gabriella Chronis and Katrin Erk. 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*, pages 227–244, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, 131:271–276.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Abhilasha A Kumar. 2021. Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28(1):40–80.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43.

Olga Majewska, Diana McCarthy, Jasper JF van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2021. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics*, 47(1):69–116.

Julian Michael, Jan A Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pages 1–12.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. Linguistic Data Consortium.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33(3-4):175–190.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(10):1769–1779.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736.

Aina Gari Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *arXiv preprint arXiv:2104.14694*.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yadollah Yaghoobzadeh, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753.

A Multi-Prototype BERT embeddings

Multi-prototype embeddings were generated as follows:

1. For each verb in the dataset, we sampled up to 100 sentences from the British National Corpus (BNC Consortium, 2007), excluding non-verbal uses of the target word. A few words in the set occurred in BNC fewer than 100 times. Four words (*broil*, *corrupt*, *exhale*, and *misspend*) did not occur as verbs at all in the BNC and were excluded from our analysis. The average number of occurrences sampled for a word was 95.6.

2. We extract BERT token embeddings for each collected occurrence of a word. For words which BERT tokenizes into multiple word pieces, we average over all component pieces.

3. We cluster the token embeddings for each verb. Like Chronis and Erk (2020), we use *k*-means clustering to group tokens into 'sense' clusters. We use the number of verb senses listed in WordNet (Miller, 1995) to determine the appropriate *k* for each word. Verbs in the dataset had on average 5.9 senses. (min: 1, max: 59, for *buzz*).

4. After identifying clusters, we take the *k* cluster centroids for each word. These are the embeddings we evaluate against the SpA-Verb categorization data.

B Evaluation metrics

As in Majewska et al. (2021), we evaluate performance of word embeddings on semantic categorization using modified purity and weighted class accuracy, which are combined in an F1 score, calculated as their balanced harmonic mean. Modified purity is the mean precision of automatically induced verb clusters:

$$\text{MPUR} = \frac{\sum_{C \in Clust, n_{prev(C)} > 1} n_{prev(C)}}{\#test\_verbs} \quad (1)$$

where each cluster C from the set of all K_{Clust} induced clusters *Clust* is associated with its prevalent gold class, and $n_{prev(C)}$ is the number of verbs in an induced cluster C taking that prevalent class, with all other verbs considered errors. $\#test\_verbs$ is the total number of verbs in the dataset. While modified purity is a measure of precision, weighted class accuracy targets recall:

$$\text{WACC} = \frac{\sum_{C \in Gold} n_{dom(C)}}{\#test\_verbs} \quad (2)$$

where for each class C from the set of gold standard classes *Gold*, we identify the dominant cluster from the set of induced clusters having most verbs in common with C ($n_{dom(C)}$).

C Ground truth coarse-grained categories

The ground truth categories used for evaluating models on coarse-grained categorization come from Phase 1 of SpA-Verb. 825 verbs are grouped into 17 broad semantic classes. Table 3 gives an overview of the classes.

D MPro BERT Cross Layer Analysis

The MPro BERT embeddings from later layers of BERT are better predictors of the ground truth categories than earlier layers. As shown in Figure 1, F1

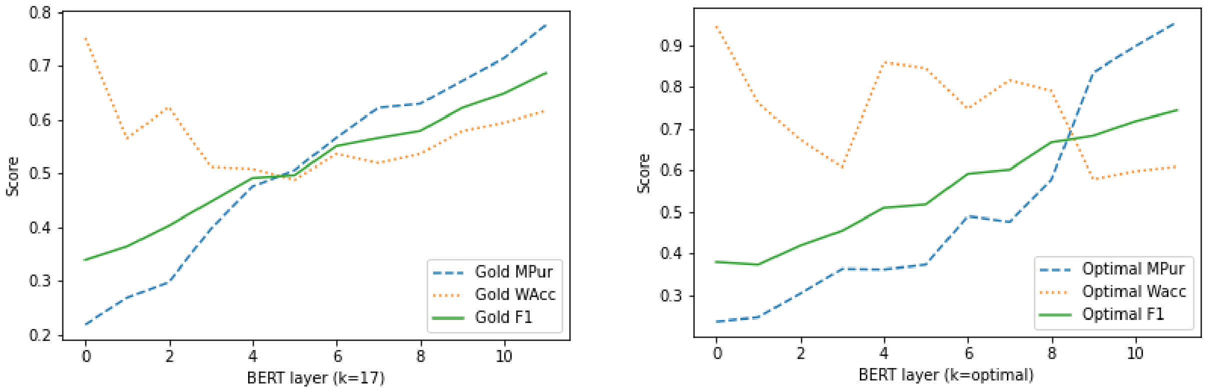

Figure 1: Performance of multi-prototype BERT embeddings from each layer. Left: gold case (*k*=17), right: optimal case

Cluster label	**Example verbs**
movement	*wander, fly, glide, roam*
communication	*persuade, command, tell*
crime & law	*beat, abduct, abuse, shoot*
negative emotion	*offend, aggravate, enrage*
positive emotion	*admire, respect, adore, like*
cognitive process	*suppose, assume, realize*
cooking	*cook, slice, stew, boil*
possession	*belong, obtain, acquire*

Table 3: A sample of the 17 gold classes in SpA-Verb dataset (labels are given for descriptive purposes only)

scores increase virtually monotonically from the first to last layer of BERT. Layer 0 performance is about on par with the static BERT baselines.

In general, recall (WAcc) decreases from earlier to later layers of BERT, while the precision measure (MPur) increases. The increase in precision is steeper than the decrease in recall, leading the F1 scores to trend up in later layers. The optimal *k* value for the middle layers is very low (5-10) but much higher for early and later layers (20-30). As can be seen in Figure 1, there is a spike in recall in the middle layers, likely due to the lower *k* values. Having a few large clusters means that clusters are more likely to overlap with gold classes, even if they contain extra irrelevant members.

E Fine-Grained Categorization Results

Table 4 shows a breakdown of the F1 scores for each model by class. The classes which all models did best at categorizing were Class 13 (which contains words describing sounds like *boom, buzz, crunch, rattle, squeak*), Class 3 (related to change: *accelerate, diminish, grow*) and Class 12 (physiological processes: *sweat, cough, breathe, yawn*). The classes which models struggled most with were Class 15 (physical movement: *catch, grab, fling, jerk*), Class 7 (communication: *announce, discuss, explain, tell*), and Class 9 (cognitive processes: *analyze, describe, ponder, think*).

Class	**word2vec**	**BERT decontext.**	**BERT aggreg.**	**MPro BERT**	**Average**
1	0.624	0.521	0.547	0.541	0.558
2	0.563	0.606	0.619	0.563	0.588
3	0.679	0.660	0.685	0.629	0.663
4	0.535	0.498	0.654	0.545	0.558
5	0.610	0.676	0.673	0.671	0.657
6	0.600	0.589	0.697	0.61	0.625
7	0.498	0.532	0.605	0.556	0.548
8	0.649	0.542	0.649	0.586	0.606
9	0.579	0.521	0.578	0.539	0.554
10	0.504	0.59	0.587	0.598	0.570
11	0.788	0.624	0.60	0.585	0.651
12	0.722	0.581	0.727	0.616	0.661
13	0.742	0.647	0.764	0.573	0.682
14	0.603	0.499	0.653	0.58	0.584
15	0.508	0.572	0.531	0.561	0.543
16	0.740	0.629	0.672	0.545	0.646
17	0.694	0.658	0.682	0.595	0.657
Average	0.626	0.586	0.643	0.582	0.609

Table 4: F1 for each class and embedding type on fine-grained categorization.