Modelling the Role of Polysemy in Verb Categorization

A great deal of work has been devoted in recent years to developing computational models of meaning based on the distribution of words in text. Traditional static embeddings [Mikolov et al. 2013] represent each word type as a unique vector, while more recent contextual models [Devlin et al. 2019] generate a unique representation for every instance of a word in context. In this paper, we focus on the role of polysemy in verb categorization. Because verbs generally have multiple possible senses, categorization decisions depend on which sense of a word is being considered. Representing the distinct senses of polysemous words is thus important to modelling how humans categorize sets of verbs' denotations. This paper shows that the contextual information implicit in recent distributional semantic models makes them a good approximation of polysemy and of verb categorization.

To understand the unique role contextual embeddings can play in modeling effects of polysemy on verb categorization, it is important to note that there are at least two different approaches to the relationship between polysemy and context. One account holds that words have a static set of possible senses, and while context helps disambiguate between possible senses of a word, those senses exist independently of context. A stronger claim, though, has been made (for example, by Elman 2009) that different senses of a word are not simply reflected in, but actually *created by* context. Proponents of this claim believe that word meaning is fundamentally context-dependent. Contextual language models like BERT implicitly take this view of word meaning, as they represent each instance of a word in a particular context as a unique embedding. We can then treat classes of contexts as equivalent to senses in these models. This is why contextual embeddings seem well-suited to test Elman-style conception of polysemy and its role in verb categorization. If this view of polysemy is correct, contextual word embeddings should model verb categorization better than static embeddings. This is the hypothesis we test in this paper.

Interestingly, 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 when approximating the verb categorization done by participants in their experiments. We argue that this result is due not to the irrelevance of context to categorization, but rather to the way the contextual embeddings were extracted from the model in Majewska et al. [2021]. Although many of the words in Majewska et al. [2021]'s ground truth data are polysemous and are assigned to multiple 'gold' classes by participants, they evaluate models in a one-representation-per-word-form manner. Even when evaluating BERT, which has been shown to encode sense-specific information, this information was thrown away by averaging over all contexts. Because they use polysemous data to test representations which do not encode sense information, Majewska et al. [2021]'s results may not reflect the full potential of contextual architectures to model categorization.

Our paper shows that by accounting for polysemy in the model representations, we can significantly improve the correlation between word embedding clusters and human categories. In particular, retaining sense-level information from contextual BERT embeddings more than doubles its performance, outperforming static embeddings by a large margin. These results demonstrate that sense-specific information is crucial even for categorization of words in isolation, and suggests that contextual embedding models are a good approximation of both polysemy and verb categorization, supporting a contextual account of word meaning. We evaluated two ways of handling polysemy in word embeddings:

1. **Static word2vec embeddings trained on POS-tagged data.** Part-of-speech tagging allows the model to distinguish between, for example, *duck_NOUN* and *duck_VERB*. This strategy

Model	F1-optimal	F1-gold
Majewska et al. [2021] word2vec	0.355	0.326
Majewska et al. [2021] BERT	0.340	0.322
Our POS-tagged word2vec	0.442	0.433
Our multi-prototype BERT	**0.755**	**0.731**

Table 1: Comparison of our methods with results reported in Majewska et al. [2021]. F1 scores reported. 'Gold:' k=17, as in the ground truth. 'Optimal:' best result for k in the range (5,30).

is simplistic as different senses which have the same part of speech are still conflated into one vector (like *get#ACQUIRE* and *get#UNDERSTAND*), but it at least factors out noise from non-verb senses.

2. **Multi-prototype BERT embeddings.** Following the methods of Chronis and Erk [2020], we distill BERT embeddings representing individual tokens into multiple prototype embeddings, which represent each sense of a word, without collapsing every token into a single representation, as in Majewska et al. [2021]. This allows for different senses of a word to be assigned to different clusters, while still generalizing beyond individual instances of a word. Multi-prototype BERT embeddings have been used to improve word pair similarity estimates, but to our knowledge this is the first time that such sense-level representations have been used to study the role of polysemy in other domains.

To evaluate the performance of each method, we use the verb categorization data from SpA-Verb [Majewska et al., 2021] as a ground truth, which comprises 825 verbs in 17 semantic classes. This data was derived from a sorting task performed by 10 participants. We use k-means clustering to group verb embeddings into predicted classes. To compare the induced clusters with the ground truth, we use the same F1 metric used by Majewska et al. [2021], which balances precision and recall. Table 1 shows the results of the two methods compared to the results reported in Majewska et al. [2021]. The F1 value for the POS-sensitive word2vec model is slightly higher than reported for a similar model architecture without POS information. Multi-prototype BERT, by contrast, performs dramatically better than any of the results previously reported.

While Majewska et al. [2021] found that contextual models performed poorly compared to static models, our results indicate that properly exploiting its contextual information allows BERT to predict verb categories very well. Retaining sense-level information from BERT by generating multi-prototype representations, rather than generating one representation per word form, more than doubles its F1 score. This boost in performance shows that contextual, sense-specific information is important to human verb categorization, and supports a strongly contextual view of polysemy and word meaning. On a more general level, these results suggest that linguistic input encodes a great deal of information about semantic categories, independently of other perceptual input that humans receive, and that this category information can be extracted from embedding models which are trained on linguistic data alone. Future work is needed, though, to further explore the role of language in forming semantic categories, and whether models like those discussed here can model the flexible, goal-dependent nature of human categorization.

References

G. Chronis and K. Erk. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *CoNLL*, Online, Nov. 2020.

O. Majewska, D. McCarthy, J. J. van den Bosch, N. Kriegeskorte, I. Vulić, and A. Korhonen. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics*, 47(1):69–116, 2021.