When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships

Gabriella Chronis
Department of Linguistics
The University of Texas at Austin
Austin, TX 78705 USA
gabriellachronis@utexas.edu

Katrin Erk
Department of Linguistics
The University of Texas at Austin
Austin, TX 78705 USA
katrin.erk@utexas.edu

Abstract

This paper investigates contextual language models, which produce *token* representations, as a resource for lexical semantics at the word or *type* level. We construct multi-prototype word embeddings from `bert-base-uncased` (Devlin et al., 2018). These embeddings retain contextual knowledge that is critical for some type-level tasks, while being less cumbersome and less subject to outlier effects than exemplar models. Similarity and relatedness estimation, both type-level tasks, benefit from this contextual knowledge, indicating the context-sensitivity of these processes. BERT's token level knowledge also allows the testing of a type-level hypothesis about lexical abstractness, demonstrating the relationship between token-level behavior and type-level concreteness ratings. Our findings provide important insight into the interpretability of BERT: layer 7 approximates semantic similarity, while the final layer (11) approximates relatedness.

1 Introduction

The rampant success enjoyed by contextualized language models (CLMs) like CoVe (McCann et al., 2017), ElMo (Peters et al., 2018), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019b) has precipitated a deluge of research into analyzing and interpreting their functionality. But to date, there has been little work analyzing their lexical semantic knowledge. This paper seeks to answer two questions: 1) Is it possible to generate useful static word-type embeddings from BERT activations for word tokens? 2) What sort of semantic relations are represented in embeddings generated from BERT?

'Useful' word embeddings are those which successfully represent target semantic relations and enable the testing of linguistic and cognitive hypotheses. Many linguistic tasks and questions concern word meanings at the type level—that is, what a word means in general, abstracted away from any particular context. Similarity, relatedness, and abstractness are often construed as properties at the type-level: most similarity and relatedness datasets contain judgments on isolated word pairs, and abstractness datasets contain judgments on isolated words.

CLMs produce representations at the token level: the vector representation of a word varies depending on its context of occurrence. How can contextual representations be helpful for type-level tasks? Similarity and relatedness are context-sensitive processes. For example, the relevant features of *water* used for calculating its similarity to *land* are different from those for calculating its similarity to *coffee*. BERT's contextual knowledge is useful for dealing with the effects of this variation on similarity and relatedness judgments. Relative abstractness/concreteness is another property often treated as a type-level phenomenon. However, this property may be detectable through token-level interactions. We hypothesize that a high degree of contextual variation may be an indicator of type-level abstractness. In other words, abstract words are more 'heterogeneous' than concrete words. Traditional static representations, which represent words as infinitesimal points, are not suitable to test this hypothesis. In contrast, BERT enables the representation of a single word as a constellation of points, with each point corresponding to a different usage or usage type. In BERT, heterogeneity translates to the dispersion of tokens or prototypes in space.

The most obvious extension of contextualized word embeddings to the type level is to build exemplar models from token representations. Such models represent a word as the sum total of observed occurrences, i.e. the set of all token vectors. Full exemplar representations are computationally expensive, and subject to noise from outliers. At the other end of the spectrum, some have tried av-

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 227–244
Online, November 19-20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

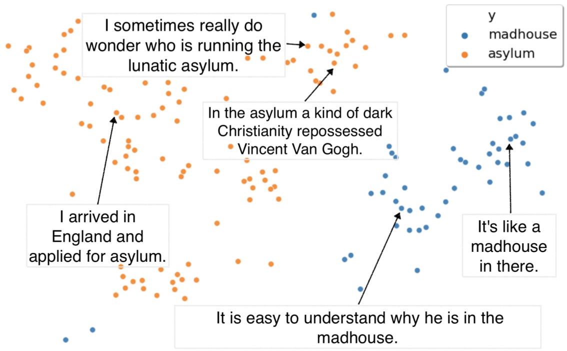

Figure 1: 2D t-SNE visualization of layer 8 vectors for tokens of *asylum* and *madhouse* sampled from the BNC.

eraging over exemplar models to generate a single vector for each word (Bommasani et al., 2020). However, these models do not leverage the contextual knowledge resident in BERTs later layers.

Our approach aims to capture the regularities in contextual variation while reducing the noise it introduces. A set of BERT token vectors for a single word naturally tends to separate spatially into groups of similar usages, or 'usage types' (Giulianelli et al., 2020). Usage types often correspond to polysemous or homonymous senses, idiomatic constructions, and affordances.[1] Figure 1 shows a t-SNE visualization of tokens for *madhouse* and *asylum*. The dominant usage of asylum is as political refuge, but there is a cluster corresponding to *asylum* as an institution of confinement for people with psychiatric diagnoses, which shares similar negative connotations to *madhouse*, a derogatory colloquialism for such institutions.

To test whether these usage types retain enough contextual information to aid in context-sensitive lexical tasks, we use K-means clustering of BERT token representations to derive multi-prototype lexical embeddings. The embeddings are evaluated on the standard lexical tasks of similarity and relatedness estimation. Clustered BERT-based representations provide high-quality predictions of human judgments for both tasks. They are also employed to test a cognitive hypothesis which holds that one of the factors contributing to relative abstractness/concreteness a word is how much its meaning varies in context. We find that the average dispersion of tokens in a cluster bears a significant relationship to abstractness.

The contributions of this paper are as follows:

1. Application of BERT for type-level lexical modeling, and the testing of type-level lexical-semantic hypotheses.

2. Clustering of contextualized representations into multi-prototype embeddings, which maintain the advantages of contextualization without the complexity burden of exemplar models, leading to improved performance on similarity and relatedness estimation.

3. Insight into the semantic interpretability of BERT, most notably that middle layers best approximate similarity while the final layer approximates relatedness.

2 Related work

The pre-trained BERT language model is a bidirectional Transformer (Vaswani et al., 2017) encoder. It has either 12 fully-connected layers (`bert-base`) or 24 (`bert-large`). The model is trained on two tasks: masked token prediction and next sentence prediction. For `bert-base`, the representation of an input sequence consists in a 12-layer activation network with 768-dimensional vectors for each input token (where tokens correspond to sub-word WordPieces [Schuster and Nakajima, 2012]) at each of 12 layers. The preponderance of research surrounding analysis and interpretability of BERT has been dubbed BERTology, after the poster-child of the contextual revolution (cf. Rogers et al., 2020 for an excellent survey).

Probing Tasks Most studies of word meaning in BERT follow an agenda of extrinsic evaluation (Artetxe et al., 2018). Tenney et al. (2019a) found that CLMs improve over non-contextual counterparts largely on syntactic tasks, with smaller gains on semantic tasks. Tenney et al. (2019b) introduced *edge*-probing tasks to analyze the layer-wise structure of BERT, and found that early layers perform syntactic tasks like part-of-speech and dependency tagging, while later layers encode information pertinent to semantic tasks like coreference resolution, relation labeling, and semantic proto-role labeling. We aim to advance this kind of structural analysis of BERT through the *intrinsic* evaluation of representations at different layers.

[1] Affordances are the different ways a thing can present itself to an individual (Gibson, 2015 [1979]). For instance, water can manifest as a drink, a chemical substance to be studied, a span to be travelled across, a medium for recreational sports, or a municipal resource, just to name a few.

Contextual Word Embeddings Existing applications of CLMs to lexical tasks use exemplar models or single-prototype models. Wiedemann et al. (2019) successfully employed a K-nearest-neighbor approach to BERT exemplar models for word sense disambiguation (WSD). Coenen et al. (2019) created a visualization tool that generates a 'word cloud' from BERT tokens, browsable by layer.[2] They also achieved a state-of-the-art F1 score on a WSD task with the simple scheme involving sense-annotated training data from Peters et al. (2018).

Ethayarajh (2019) generated static embeddings from CLM models, using the first PCA component sets of token representations. Bommasani et al. (2020) experimented with averaging tokens for context-agnostic word vectors. This approach performs well on similarity and relatedness tasks compared to traditional static embeddings. However, averaging evaporates much of BERT's contextual variation, cutting off its potential to aid in similarity estimation. It's no surprise that averages derived from earlier layers of BERT performed best at similarity estimation. Later layers demonstrate more contextual variation (Ethayarajh, 2019), making the mean less meaningful.

Similarity The most common method for intrinsic evaluation of word embeddings is similarity estimation. However, major critiques have been leveled at the standard similarity datasets, and even at the construct of similarity itself. Depending on which comparison a word enters into, there is variation in which senses (and/or features) of the word are considered for the calculation. Nelson Goodman's (1972) dismissal of similarity argues that the *selection* of properties to consider varies so widely as to be essentially arbitrary. Thus, similarity datasets which ask raters to assess the similarity between two words out of context are argued to be premised on the flawed notion that similarity is fixed (Faruqui et al., 2016).

Two words are never absolutely similar or dissimilar. Rather, in assessing the likeness between two words, one implicitly selects some grounds for assessing their likeness. In Tversky's classic (1977) feature-matching model, the similarity of two items is computed from the number of shared properties they have as compared to the number of properties they hold distinct. Tversky notes that property selection is subject to variation. To give an extreme example involving polysemy, the relevant features for comparing *bishop* with *rabbi* are different from those for relevant for comparing *bishop* and *rook*.

Given the complaints lodged against similarity, and the existence of contextual similarity datasets (Erk et al., 2013; Pilehvar and Camacho-Collados, 2019; Stanovsky and Hopkins, 2018), why bother applying BERT to non-contextual datasets at all? And for that matter, what is the value of modeling word-pair similarity judgments? The pared down format of word-pair judgments foregrounds the cognitive regularities of the underlying process by which humans select a grounds for comparison. Despite potential variability in similarity judgments, inter-annotator agreement for word-pair similarity ratings is fairly high (Medin et al., 1993). In one sense, an apple is like candy: both are sweet snack foods. However, they are consistently judged to be dissimilar (2.08 in Simlex999 on a scale of 1-10; SD=0.75). Together, *apple* and *candy* co-determine a grounds for likeness, narrowing the context or the features under consideration (one is healthy, the other rots your teeth).

Medin et al. (1993) argue that similarity is a *process*, and this process is governed by constraints which give rise to regularities. For example, antonyms such as *black-white* are judged to be maximally dissimilar in isolation, but are judged more similar when presented alongside a comparison such as *black-red* containing a related word. As the authors so vividly put it, "Nelson Goodman (1972) called similarity a chameleon, but we believe that similarity is more like two yoked chameleons: The entities entering into a comparison jointly constrain one another and jointly determine the outcome of a similarity comparison" (272).

Viewed as a context-sensitive, constraint-based process, it seems natural that similarity judgments should benefit from context-sensitive lexical representations. In single-prototype embeddings, the properties or features which are considered are necessarily constant. The same vector is considered in every similarity calculation for a single word. By representing a word as a set of vectors, each corresponding to a prototypical usage, we can access the usage types relevant to different comparisons separately.

Relatedness Similarity is paradigmatic: highly similar words are more likely to occupy the same 'slot' in a sentence (i.e., *The bug is on the*

[2] https://storage.googleapis.com/bert-wsd-vis/demo/index.html.

rug/carpet). Semantic relatedness, on the other hand is syntagmatic: two highly related words are likely to appear in succession, as in *She filled the car with gasoline*. As a comparison between two isolated words, the argument for context-sensitivity of similarity also applies to relatedness judgments. The next section describes a BERT-based approach to inject a degree of context-sensitivity into similarity and relatedness estimation.

3 Multi-prototype BERT embeddings

As stated, BERT token representations demonstrate greater contextual variation at each progressive layer (Ethayarajh, 2019). To capture the contextual knowledge of later layers, we construct multi-prototype embeddings from many token representations of a single word.

The usage types captured by clusters do not always correspond directly to dictionary word senses, but they often discriminate between senses of polysemous and homonymous words, metaphorical senses, as well as syntactic roles and constructions. (Giulianelli et al., 2020). For this reason, count-based multi-prototype models have long been recognized for their usefulness to WSD tasks (Schütze, 1998; Reisinger and Mooney, 2010; Pilehvar and Camacho-Collados, 2019). We observed that natural clusters often capture cognitive affordances. Whether or not they are linguistically significant, these loose categories likely play a role in cognitive processes like similarity. To recast our earlier observation: the affordance(s) of *water* that comes to mind in the comparison of *water-land* are different from the idea that surfaces when comparing *water-coffee*. We use clustering to approximate BERT's usage-types and the affordances they capture, and demonstrate their usefulness for lexical tasks.

3.1 Materials & Methods

The transformation of token-vectors into multi-prototype vectors requires several steps, described here for a single word w.

Data Collection First, a set S of up to 100 sentences containing a token t of w was sampled at random from the British National Corpus (BNC, Burnard, 2000).[3] As the human judgments in our evaluation datasets were collected agnostic to part of speech, and are particular to word forms, no lemmatization or tagging was used. Any sentences too large to input to BERT were discarded.

Then, each sentence $s \in S$ was passed to the pre-trained `bert-base-uncased` model, and the layer-wise network activations obtained. If BERT split the token t into subword WordPieces[4], we followed the now-standard practice of averaging t's subword vectors to obtain a single token vector for t at each layer. This process yielded a vector for each token at each of 12 layers.

Clustering At each layer, w's token vectors at that layer were clustered using K-means. Separate embeddings were calculated for each layer $l \in L = \{0..11\}$, and for each number of K-means clusters $K \in \{1, ..., 10, 50\}$.[5] For a given choice of K and l, the set of cluster centroids $\boldsymbol{\pi}(w)^l_1 \dots \boldsymbol{\pi}(w)^l_K$ constitute the multi-prototype representation for w. These centroids correspond to K prototypical usages of w. In the rare case that $|S| < K$ for w, it was not possible to construct a K-prototype representation for w. Consequently, scores could not be calculated for a handful of words for some parameter combinations. However, scores for at least 99% of the comparisons for all datasets were collected for K=1-10 (See A.2). This detail renders comparison between models less than ideal, but the differences are so minimal as to make the issue negligible.

Evaluation Cosine distance, the typical method of relation estimation, will not work for multi-prototype models, as it is a function of two vectors. Reisinger and Mooney (2010) compute distance using the centroids of clusters. MaxSim of words w, w' is the maximum cosine similarity of any cluster centroid of w to any centroid of w'. In our case, the layer l introduces an additional parameter, such that $MaxSim(w, w', l, K) =$

$$\max_{1 \leq j \leq K, 1 \leq k \leq K} cos(\boldsymbol{\pi}^l_j(w), \boldsymbol{\pi}^l_k(w'))$$

where $\boldsymbol{\pi}_{j,l}(w)$ corresponds to the centroid of the jth cluster for word w at layer l. Effectively, the similarity of two words is equal to the similarity

[3] Indices of the sampled sentences are available at https://github.com/gchronis/MProBERT

[4] Note that this process is referred to as 'tokenization', and the resulting WordPieces are typically referred to as tokens. To avoid confusion, this paper reserves the word 'token' to refer to an occurrence of an entire word in sentential context.

[5] Results for k=50 were evaluated for the sake of curiosity about the extreme case but are not discussed—such embeddings are more like exemplar models than multi-prototype models.

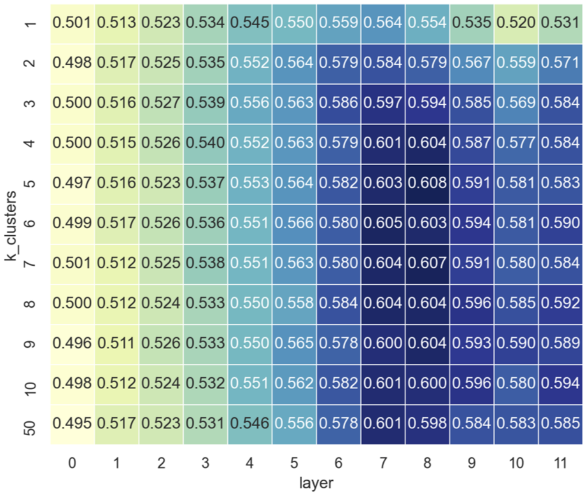

Figure 2: Spearman's ρ for BERT multi-prototype embeddings correlated against SimLex-999 **similarity** for each combination of layer and number of K-means clusters ($p < 1e^{-60}$).

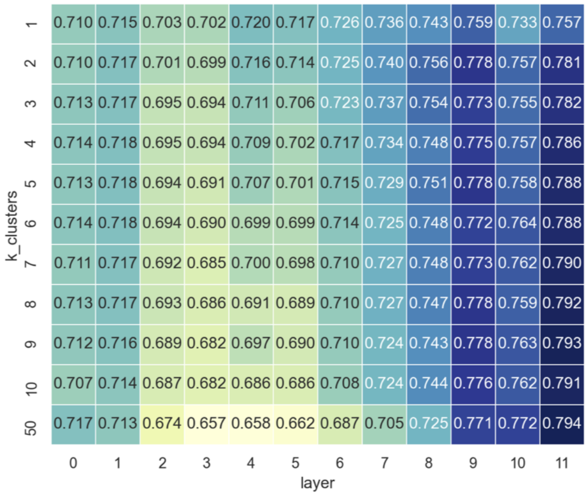

Figure 3: Spearman's ρ for BERT multi-prototype embeddings correlated against MEN **relatedness** ratings for each combination of layer and number of K-means clusters ($p < 1x10^{-100}$).

of their two closest prototypes. We also experimented with their AVGSIM, defined as the mean of all pairwise similarities between w's centroids against those of w'. For AVGSIM, two words are close if many of their prototypes are close. In stark contrast to Reisinger and Mooney (2010), we found MAXSIM to universally outperform AVGSIM, and so report results on the former only.

4 Similarity and Relatedness

4.1 Datasets

The embeddings were evaluated against three similarity datasets and four relatedness datasets. For similarity, we used SimLex999 (Hill et al., 2015), SimVerb3500 (Gerz et al., 2016), and WordSim353$_{sim}$ (Agirre et al., 2009). The latter is a partition of WordSim353 (Finkelstein et al., 2001) devised to approximate similarity. SimLex999 contains 999 word pairs balanced for concreteness and annotated specifically for similarity. SimVerb3500, containing 3500 verb pairs, was also designed specifically to target similarity. For relatedness datasets we used WordSim353$_{rel}$ (Agirre et al. (2009)'s complementary relatedness subset); MEN, containing 3000 word pairs (Bruni et al., 2014), and YP-130, consisting of 130 verb pairs (Yang and Powers, 2006). Though it was not collected with specific instructions to the raters, for the sake of comparison to other approaches, we also evaluate on the original WordSim353 dataset, and report results with the relatedness datasets.

4.2 Results

Overall, clustering yielded great improvement, especially at later layers.

Similarity Performance on all similarity datasets peaked between layers 7-9, somewhere between 5-8 clusters. This pattern is exemplified by the heatmap in Figure 1, which displays Spearman's ρ correlations against SimLex999 for embeddings at each layer and each choice of K. Performance on SimLex999 reaches a peak score at layer 8 with 5 clusters (ρ = 0.608). SimVerb3500 follows a strikingly similar pattern, with the highest performance at layer 7 with 7 clusters (ρ = 0.531). For WordSim353$_{sim}$, performance peaks at layer 9 with 8 clusters (ρ = 0.826), but the performance is nearly matched by layer 11 with 4/5 clusters (ρ = 0.825).

Relatedness Across the board, layer 11 achieves the best performance on relatedness tasks. The pattern is illustrated in Figure 3, which shows Spearman's ρ correlations between MAXSIM predictions and MEN ratings for each layer and choice of K. Clustering induces significant performance improvements at layers 8 and up. Layers 0-1 show little variation due to clustering, and layer 2-7 performance actually degrades. MEN performance peaks at layer 11 with 9 clusters ($\rho = 0.793$). The other datasets show a similar pattern, with peak performance at layer 11 between 7 and 9 clusters: K=7 for WordSim353$_{rel}$ (ρ = 0.665), K=9 for YP-130 ($\rho = 0.715$), and K=7 for WordSim353 ($\rho = 0.747$).

	Similarity			Relatedness			
	SL-999	*SV-3500*	*WS-353$_{sim}$*	*WS-353*	*WS-353$_{rel}$*	*MEN*	*YP-130*
Distributional	0.563 SP-15	0.364 CBOW	0.795 GloVe	0.738 FastText	**0.681** FastText	**0.801** GloVe	0.535 GloVe
CLM-based	0.550 XLNet-24 (4)	0.455 XLNet-24 (3)	- -	0.730 BERT-24 (6)	- -	0.200 BERT-pca-1 (1)	- -
MProBERT	**0.605**	**0.528** (layer=7, K≤9)	**0.807**	**0.741**	0.653 (layer=11, K≤9)	0.781	**0.711**

Table 1: Performance of best unioned multi-prototype BERT embeddings (M-ProBert) for both similarity and relatedness estimation tasks (Spearman's ρ) compared to other CLM-based word embeddings and to state-of-the-art corpus-based distributional models. Best results in bold. Parentheses for CLM models indicate layer number.

Model Selection It is possible to sidestep the issue of model selection by following Reisinger and Mooney (2010) in taking the union of all of the prototypes of different cluster sizes. This method works as well or nearly as well as selecting the best value for K, with the union of all clusters for $K \leq 9$ giving the best results overall.

As a general model for similarity estimation, we suggest layer 7 with the union of all clusters for $K \leq 9$. For relatedness, we suggest embeddings built from layer 11 using the union of all clusters $K \leq 9$. For both similarity and relatedness, $K = 7$ does provide marginally better results, but the success of the unioned model demonstrates that *post hoc* selection of K is not necessary to achieve good performance.

Table 1 compares the best unioned models for similarity and relatedness to state-of-the-art distributional approaches trained on running monolingual text (without the injection of structured knowledge), as well as to other CLM-based static embeddings. For distributional models we compared to Symmetric Pattern embeddings (SP-15, [Schwartz et al., 2015]), GloVe (Pennington et al., 2014), CBOW (Mikolov et al., 2013), and FastText (Bojanowski et al., 2017). For other CLM-based embeddings, we compared to Bommasani et al. (2020), who tested layer-wise token aggregations for numerous architectures: BERT, RoBERTa, GPT2, XLNet, and DistilBert. We also compared to Ethayarajh (2019), who examined the first PCA component of individual layers of several CLMs. Performance drastically improves over other CLM-based embeddings, and our generalized similarity estimation model surpasses the distributional state-of-the-art on all three datasets.[6]

4.3 Discussion

In contrast with Bommasani et al. (2020), who find performance to peak at early layers, our model's performance peaks at later layers, which we know to possess more fine-grained contextual information (Ethayarajh, 2019). Multi-prototype embeddings harness the power of this contextual information for the type-level tasks of similarity and relatedness estimation. The fact that contextual information aids in similarity estimation supports the hypothesis that similarity, even between isolated words, is a dynamic, context-sensitive process. Indeed, word-pair or 'context-free' similarity estimation is not truly a type-level task. Each word in a pair constitutes the linguistic context for the interpretation of the other word. By capturing the most typical contexts for a word, multi-prototype embeddings enable the selection of a grounds for likeness, approximated here with MAXSIM.

Multi-prototype vs Exemplar The optimal number of prototypes is relatively small. Performance increases with K up to a point, after which it begins to degrade minimally but steadily (Figure 4.2). This behavior can be explained in terms of the model taking into account more outliers with a very high K. The higher K is, the more likely we are to find small clusters very far from the rest of the tokens, representing a rare but highly specific usage type, or even totally unique usages in clusters all by themselves. As K is maximized towards an exemplar model, the likelihood increases that any pair of words will have exemplars that are near to each other, thus increasing the predicted similarity. We hypothesize that this kind of overestimation of similarity is the reason performance degrades.

More generally, we found multi-prototype models with relatively few (K≤10) prototypes better

[6] Benchmark scores are self-reported by the authors or obtained from Lastra-Díaz et al. (2019)'s reproducible survey. Future work should reimplement these models to test the statistical significance of the performance gains.

purposed than more exemplar-like models (K=50) for estimating semantic relationships. In addition to being more lightweight, multi-prototype vectors appear to be a beneficial abstraction over exemplar models, at least for the present task. Whether or not exemplars are stored in memory, knowledge of individual outliers is not relevant to these tasks, which deal with stereotypical or prototypical class relationships rather than relationships between individuals.

Similarity vs Relatedness The analysis also uncovered differences in the semantic relations approximated by different layers. The final layer of BERT approximate relatedness, while layer 7 is optimal for estimating similarity. This finding bears an interesting connection to recent insights in BERTology. The middle layers of `bert-base` (6-9) are consistently noted to be the most transferable, i.e., to perform the best across tasks (Hewitt and Manning, 2019; Goldberg, 2019; Jawahar et al., 2019; Liu et al., 2019a). The final layers, on the other hand, are the most specific to the next sentence prediction task. The connection suggests that successful representation of semantic similarity may be critical to many NLP tasks, moreso than relatedness. Our results support the thesis that static vectors cannot surfaces all aspects of lexical semantic meaning at once (Artetxe et al., 2018). But, perhaps when forced to compromise on one general purpose embedding for downstream applications, one which approximates similarity may be preferable over those which approximate relatedness—these embeddings appear to work best on a wide variety of downstream tasks.

The difference in preferred layer for different tasks confirms that similarity and relatedness, so important a distinction to distributional models, ought to be treated separately in CLMs as well.

Clusters While we do not undertake a systematic analysis into the types of usage-clusters captured by K-means, qualitative examination indicates that the gains from clustering come from expected behavior. For example, at layer 8 with K=3, the clusters for *river* correspond to the natural feature and associated sensory imagery, the name construction *river X*, and adjectival uses (e.g. *river warden, river dolphin*). The clusters for *stream* correspond to to a a fluid *medium* through which other entities pass (e.g. *blood stream, gas stream*), the natural feature, and a *substance in motion* constituting the stream (many but not all of which examples take the form *stream of X*, where X is not a typical fluid).[7] Among these clusters, MaxSim selects the two corresponding to the concrete natural feature.

The clusters selected by MaxSim do not always correspond to a clear cohesive usage-type. Sometimes it selects a catchall prototype, or one one that encompasses multiple distinct usage types. However, even in these cases, K-means separates out distractor usages such as proper names and specific constructions that would otherwise shift the mean of a single prototype into less relevant realms. The SimLex999 rating between *cat* and *lion* is 6.75 (SD=0.84). The single prototype model (l=8) predicts 5.79. The multi-prototype model (l=8, k=7) is more accurate at 6.33. MaxSim selects a *cat* cluster that lacks an obvious interpretation (though it does frequently contrast cats to other animal species). Importantly, MaxSim *excludes* a cluster containing cats as pets, a cluster in which a cat is a grammatical subject, and one for a colloquialism meaning 'obtain'. The selected *lion* cluster contains lions which interact directly with humans. They have surprising docile characteristics like being 'tamed' or 'lying down like a lamb'. The discarded clusters correspond to lion as wild animal, as a name, the idiom *lion's share*, and metaphorical human lions.

The *cat-lion* example shows that the model implements the principle of context-sensitivity for similarity estimation. The chosen prototypes are not the most stereotypical, but actually those which *downplay* the distinctive features of domesticity in cats and wildness in lions. The relatively high human rating may be attributed to the recognition that the two species have many shared biological traits compared to other animals, and MaxSim is able to find this common ground. Refer to Appendix B for example sentences from different clusters.

5 Abstractness and contextual variation

To demonstrate the potential of multi-prototype embeddings, we next leverage our model to test a cognitive hypothesis about concreteness. Specifically, we examine the difference between abstract and

[7] The blend between syntactico-grammatical and semantic prototypes underscores the relationship between form and meaning. The construction *stream of X* invites a focus on the movement of the substance and tends towards more metaphorical uses, allowing count nouns, whereas *X stream* invites a focus on the channel created by the substance and is limited to more 'fluidlike' fluids.

concrete words with respect to spatial dispersion among prototypes.

Roughly speaking, abstract concepts are classically characterized as those that are “neither purely physical nor spatially constrained” (Barsalou and Wiemer-Hastings, 2005, 129). Concrete words are recognized and comprehended faster (West and Holcomb, 2000), remembered longer (Paivio, 1971, 2013; Fliessbach, Weis, Klaver, Elger, and Weber, 2006), and more resilient to brain damage (Katz and Goodglass, 1990) than abstract words.

The dominant explanation for concreteness effects is called the Dual-Coding Hypothesis (Paivio, 1971, 2013; Crutch and Warrington, 2005). It holds that concrete and abstract concepts are organized differently in the brain: the former are grounded in experience, while the latter are based solely on other concepts. The sensory richness of concrete concepts explains their memory advantage. Significant distributional-semantic research has tested the Dual-Coding Hypothesis, with mixed results (cf. Hill et al., 2013, 2014a,b).

Dual-Coding Theory demands discrete types of conceptual representations, and raises an uncomfortable metaphysical issue of whether a concept can be about anything other than the purely physical. Embodied views of abstract representation (Kousta et al., 2011) hold that both linguistic and experiential information contribute to rich representations for *all* concepts, and that the apparent distinction arises from statistical patterns in the proportion of sensorimotor to affective experiential information undergirding the concepts. The question then arises of how to account for concreteness effects without positing distinct representations.

A recent psychological theory contends that concreteness effects are a consequence of ‘situational systematicity’ (Davis et al., 2020): abstract concepts are constituted by a larger and more complex set of relationships, dispersed through space and time, and are therefore more subject to contextual variation. If the hypothesis is valid, it might contribute to an explanation of concreteness effects without reliance on the claim that abstract concepts lack experiential grounding: simple, systematic, contextually invariant concepts would naturally be easier to remember.

Testing the situational systematicity hypothesis demands a way to measure the complexity and contextual variability in concepts. Multi-prototype BERT meets this demand, at least for modeling variability, which can be viewed a proxy for complexity. Single-prototype models cannot be used to test this hypothesis—they lack such fine-grained information about contexts of occurrence.

5.1 Setup

We analyze the 1028 unique words in the SimLex999 dataset, each of which is annotated with its USF concreteness norm (Nelson et al., 2004). As a measure of heterogeneity we use average pairwise token distance among clusters. For a given K and l, we calculate the average pairwise cosine distance between all tokens in a cluster, and then average that value across clusters. This value measures how heterogeneous a word’s usage-types tend to be.

5.2 Results

Results indicate a relationship between abstractness and dispersion. However, the nature of this relationship changes throughout the layers of the network. Figure 4 shows the correlation between concreteness and average pairwise token distance throughout the layers of `bert-base` with K set to 9. The strongest correlations are observed for this choice of K, reaching a maximum of $\rho = -0.264$ at layer 9. The same pattern of correlations, from significantly positive to significantly negative, was observed for all choices of K.

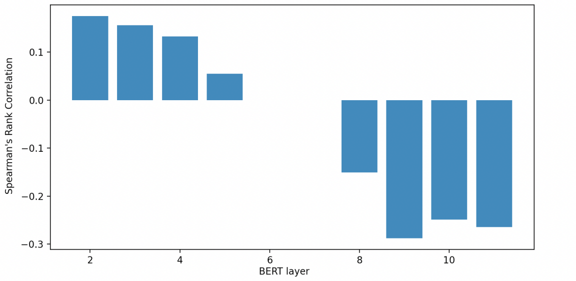

Figure 4: Spearman’s correlation between USF concreteness norms and average inter-token distance for K=9 multi-prototype vectors ($p < 0.1, N = 1028$). Correlations for layers 0, 5, and 6 were not significant.

The correlation changes direction depending on the layer: there is a significant positive correlation at layers 1-5, and a significant negative correlation at 7-11. A strong negative correlation means that abstract words tend to demonstrate greater variance at these layers than do concrete words.

5.3 Discussion

The heterogeneity hypothesis predicts a negative correlation between concreteness and variance.

This pattern is upheld at layers 7-11. At these layers, tokens in the clusters of an abstract word tend to be somewhat farther apart from one another than those in a concrete word. Ethayarajh (2019) noted that inter-token variance increases throughout the layers of BERT. While true in general, variance increases more for abstract words than concrete words. However, the hypothesis is not upheld at early layers. Early layers tend to separate concrete tokens while later layers tend to separate abstract tokens.

Where the model demonstrates more variance between tokens, it encodes more detailed knowledge of that word. Where it displays less variance, these internal differences are collapsed. That is, where variance is low, the model is 'focused', or zoomed in, on representing differences between words, and where variance is high, the model is 'focused' on representing variation in how one word is used. The reason for this shift in correlation from positive to negative, or in other words the shift in 'focus' from concrete to abstract words, remains opaque. The immediate conclusion to be drawn from this analysis is that abstractness, a type-level property, bears a significant connection to behavior at the token level. This supports the idea that properties at the word level are dependent on interactions at the token level, and demonstrates the utility of token-level representations for lexical semantics at the type level.

Implications for the situational systematicity hypothesis are inconclusive. Contextual variation represents just one dimension along which concrete and abstract concepts vary. On its own, the correlation between dispersion and concreteness does not capture the data explained by Dual-Coding Theory. The results of the present analysis, while preliminary, constitute a proof-of-concept for how at least some differences between abstract and concrete words might be accounted for without positing richer representations for one type over the other. Incorporating more properties and using of more sophisticated measures of contextual variation may prove to distinguish representations of concrete and abstract words even further.

6 Conclusion

We have presented evidence that the spatially complex lexical representations afforded by CLMs are useful in type-level lexical modeling, and in investigating type-level semantic questions. We first addressed the common tasks of predicting word similarity and relatedness, and demonstrated that BERT produces high-quality multi-prototype word embeddings. We then used these representations test a hypothesis about word-level abstractness, and uncovered a significant connection between this type-level property and the relationships between tokens.

Multi-prototype embeddings represent word meaning as a *constellation* of points, as opposed to a single point. The strength of multi-prototype embeddings (and other alternatives such as probabilistic, Gaussian, and exemplar-based models) lies in this word-internal complexity. A vector lacks internal complexity—it is infinitesimal. As such, it can only be compared to other vectors on the basis of their location in space. Internally complex representations, on the other hand, can be compared on the basis of their internal geometry. When words have a shape, we can ask how the shape of words, or of lexical categories, compare to one another. Computational lexical semantics ought to seek out and apply mathematical methods for comparing the higher-order structure or topology of word meaning in such models.

Future Work In this experiment we limit the sample size for each word to 100, and use a fixed K for each word. While increasing the number of data points would perhaps lead to more natural fine-grained clusters, and therefore more sensible K-means clusters for a higher K, we predict that outliers would still lead to the over-prediction of similarities. In order to maximize the contribution to the longstanding debate between exemplar and prototype models for semantic memory, the effects of increased sample size should be evaluated. In addition, the optimal number of prototypes is not the same for every word. We hypothesize that the application of non-parametric and hierarchical clustering methods will demonstrate a tendency towards a relatively low number of prototypes, further validating multi-prototype models of semantic memory as well as improving performance.

Future work will apply CLMs to to other lexical tasks and questions, such as metaphor. Metaphor interpretation is a context-sensitive process akin to similarity judgment. The Rational Speech Act formalization of metaphor interpretation (Kao et al., 2014) utilizes hand-crafted feature vectors; it might be extended by inducing metaphorical sense representations from BERT-based prototypes.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291. Association for Computational Linguistics.

L Barsalou and K Wiemer-Hastings. 2005. Situating abstract concepts. In *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *arXiv:1607.04606 [cs]*.

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. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *arXiv:1906.02715 [cs, stat]*. Version: 1.

Sebastian J. Crutch and Elizabeth K. Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Charles P. Davis, Gerry T. M. Altmann, and Eiling Yee. 2020. Situational systematicity: A role for schema in understanding the differences between abstract and concrete concepts. *Cognitive Neuropsychology*, 37(1):142–153.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

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*, pages 55–65. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv:1605.02276 [cs]*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. Association for Computing Machinery.

K. Fliessbach, S. Weis, P. Klaver, C. E. Elger, and B. Weber. 2006. The effect of word concreteness on recognition memory. *NeuroImage*, 32(3):1413–1421.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182. Association for Computational Linguistics.

James J. Gibson. 2015 [1979]. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, New York ; Hove, England.

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. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv:1901.05287 [cs]*.

Nelson Goodman. 1972. Seven strictures on similarity. In *Problems and Projects*. Bobs-Merril, Indianapolis.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138. Association for Computational Linguistics.

Felix Hill, Douwe Kiela, and Anna Korhonen. 2013. Concreteness and corpora: A theoretical and practical study. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 75–83, Sofia, Bulgaria. Association for Computational Linguistics.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014a. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, 38(1):162–177.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. Association for Computational Linguistics.

Justine Kao, Leon Bergen, and Noah Goodman. 2014. Formalizing the pragmatics of metaphor understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36).

R. B. Katz and H. Goodglass. 1990. Deep dysphasia: analysis of a rare form of repetition disorder. *Brain and Language*, 39(1):153–185.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 85:645–665.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. 1993. Respects for similarity. *Psychological Review*, 100(2):254–278.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*.

Douglas Nelson, Cathy McEvoy, and Thomas Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 36:402–7.

Allan Paivio. 1971. *Imagery and verbal processes*. Holt, Rinehart & Winston, New York.

Allan Paivio. 2013. Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General*, 142(1):282–287.

Jeffrey Pennington, Richard Socher, and Christopher 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. Association for Computational Linguistics.

Matthew 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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273. Association for Computational Linguistics.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs].*

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 258–267, Beijing, China. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Gabriel Stanovsky and Mark Hopkins. 2018. Spot the odd man out: Exploring the associative power of lexical resources. In *EMNLP*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv:1905.06316 [cs.CL].*

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv:1706.03762 [cs].*

W. Caroline West and Phillip J. Holcomb. 2000. Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience*, 12(6):1024–1037.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv:1909.10430 [cs.CL].*

Dongqiang Yang and David Powers. 2006. Word similarity on the taxonomy of WordNet. In *Proceedings of GWC 2006*, pages 121–128.

A Reproducibility Details

A.1 Data Collection

Tokens from the British National Corpus (Burnard, 2000) were collected using NLTK v3.4.5. The sentences in the corpus were shuffled to ensure a random sample of tokens for each word. Lists of token sentences for each word, along with their original BNC indices, are available in the `word_data` directory in the supplemental material. To generate token vector representations, we used the HuggingFace `pytorch_pretrained_bert` implementation of the pre-trained `bert-base-uncased`.

A.2 Evaluation Datasets

For a few words, the BNC did not contain enough tokens to generate multi-prototype embeddings for some choices of K. These words are not included in the analyses for that choice of K. For instance, if only 7 tokens of a word were collected, and K=8, predictions were not calculated for pairs containing that word. Fortunately, this was very rare, and most words in the evaluations datasets have at least 50 tokens in the BNC, if not more. Tables 2 and 3 give detailed information about which word-pairs from each dataset, if any, were not evaluated at each cluster. As a consequence of the pruning, the Spearman's ρ correlation was sometimes calculated on minimally different data from one cluster to the next. However, the differences are so minimal as to make the issue negligible. Importantly, for the unioned models which we compare to other approaches, over 99% of all word-pairs were evaluated for each dataset.

A.3 Supplemental Material

The codebase for this project, including scripts for collecting data, generating BERT representations, calculating clusters, evaluating models, and visualizing results, is available at `https://github.com/gchronis/MProBERT`. Here one can also find lists of the tokens included in this analysis along with their BNC indices.

B Clusters

K-means clustering of token BERT representations captures polysemy as well as different usage types. Tables 4 and 5 shows a selection of sentences in clusters for *stream* and *river* at layer 8 for k=3 clusters. Tables 7 and 6 show a representative selection of tokens from each of the clusters at layer 8 for k=7. While the prototypes are often aligned

Dataset	K	Percentage Evaluated	Omitted Words	Omitted pairs
WordSim353_sim	8	99.51	{aluminum}	{aluminum-metal}
	9	99.01	{aluminum, kilometer}	{mile-kilometer, aluminum-metal}
	10	99.01	{aluminum, kilometer}	{mile-kilometer, aluminum-metal}
WordSim353_rel	9	99.6	{kilometer}	{territory-kilometer}
	10	99.6	{kilometer}	{territory-kilometer}
WordSim353	8	99.72	{aluminum}	{aluminum-metal}
	9	99.15	{kilometer, aluminum}	{mile-kilometer, territory-kilometer, aluminum-metal}
	10	99.15	{kilometer, aluminum}	{mile-kilometer, territory-kilometer, aluminum-metal}
SimLex999	4	99.8	{orthodontist}	{orthodontist-dentist, doctor-orthodontist}
	5	99.7	{disorganize, orthodontist}	{orthodontist-dentist, doctor-orthodontist, disorganize-organize}
	6	99.7	{disorganize, orthodontist}	{orthodontist-dentist, doctor-orthodontist, disorganize-organize}
	7	99.7	{disorganize, orthodontist}	{orthodontist-dentist, doctor-orthodontist, disorganize-organize}
	8	99.5	{aluminum, disorganize, orthodontist}	{metal-aluminum, tin-aluminum, orthodontist-dentist, doctor-orthodontist, disorganize-organize}
	9	99.5	{aluminum, disorganize, orthodontist}	{metal-aluminum, tin-aluminum, orthodontist-dentist, doctor-orthodontist, disorganize-organize}
	10	99.5	{aluminum, disorganize, orthodontist}	{metal-aluminum, tin-aluminum, orthodontist-dentist, doctor-orthodontist, disorganize-organize}

Table 2: Words from each dataset for which fewer than K tokens were collected, along with word pairs that were consequently omitted from evaluation. Where K is not listed, the number of tokens collected for each word was sufficient to construct K-prototype vectors. Each entry reports the percentage of word-pairs in the dataset evaluated for that K.

Dataset	K	Percentage Evaluated	Omitted Words	Omitted pairs
YP-130	3	99.23	{commercialize}	{distribute-commercialize}
	4	99.23	{commercialize}	{distribute-commercialize}
	5	99.23	{commercialize}	{distribute-commercialize}
	6	99.23	{commercialize}	{distribute-commercialize}
	7	99.23	{commercialize}	{distribute-commercialize}
	8	99.23	{commercialize}	{distribute-commercialize}
	9	99.23	{commercialize}	{distribute-commercialize}
	10	99.23	{commercialize}	{distribute-commercialize}
MEN	1	99.93	{ipod}	{chair-ipod, ipod-rope}
	2	99.93	{ipod}	{chair-ipod, ipod-rope}
	3	99.93	{ipod}	{chair-ipod, ipod-rope}
	4	99.87	{donut, ipod}	{cafe-donut, chair-ipod, ipod-rope, donut-panda}
	5	99.87	{donut, ipod}	{cafe-donut, chair-ipod, ipod-rope, donut-panda}
	6	99.87	{donut, ipod}	{cafe-donut, chair-ipod, ipod-rope, donut-panda}
	7	99.67	{donut, colorful, ipod}	{colorful-outfit, cafe-donut, colorful-toy, colorful-frame, colorful-duck, colorful-wood, colorful-lab, chair-ipod, ipod-rope, donut-panda}
	8	99.67	{donut, colorful, ipod}	{colorful-outfit, cafe-donut, colorful-toy, colorful-frame, colorful-duck, colorful-wood, colorful-lab, chair-ipod, ipod-rope, donut-panda}
	9	99.67	{donut, colorful, ipod}	{colorful-outfit, cafe-donut, colorful-toy, colorful-frame, colorful-duck, colorful-wood, colorful-lab, chair-ipod, ipod-rope, donut-panda}
	10	99.67	{donut, colorful, ipod}	{colorful-outfit, cafe-donut, colorful-toy, colorful-frame, colorful-duck, colorful-wood, colorful-lab, chair-ipod, ipod-rope, donut-panda}
SimVerb3500	1	99.94	{misspend}	{misspend-pass, pass-misspend}
	2	99.94	{misspend}	{misspend-pass, pass-misspend}
	3	99.8	{broil, misspend}	{bake-broil, broil-cook, broil-burn, broil-fry, broil-boil, misspend-pass, pass-misspend}
	4	99.8	{broil, misspend}	{bake-broil, broil-cook, broil-burn, broil-fry, broil-boil, misspend-pass, pass-misspend}
	5	99.69	{broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, mow-plow, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, misspend-pass, pass-misspend}
	6	99.6	{intoxicate, broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, drink-intoxicate, mow-plow, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, dislike-intoxicate, belong-intoxicate, misspend-pass, pass-misspend}
	7	99.6	{intoxicate, broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, drink-intoxicate, mow-plow, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, dislike-intoxicate, belong-intoxicate, misspend-pass, pass-misspend}
	8	99.6	{intoxicate, broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, drink-intoxicate, mow-plow, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, dislike-intoxicate, belong-intoxicate, misspend-pass, pass-misspend}
	9	99.6	{intoxicate, broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, drink-intoxicate, mow-plow, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, dislike-intoxicate, belong-intoxicate, misspend-pass, pass-misspend}
	10	99.51	{intoxicate, hypnotize, broil, plow, misspend}	{plow-dig, bake-broil, sow-plow, drink-intoxicate, mow-plow, spell-hypnotize, hypnotize-control, broil-cook, broil-burn, broil-fry, broil-boil, plow-hit, hypnotize-remember, dislike-intoxicate, belong-intoxicate, misspend-pass, pass-misspend}

Table 3: Continued from last page. Words from each dataset for which fewer than K tokens were collected, along with word pairs that were consequently omitted from evaluation. Where K is not listed, the number of tokens collected for each word was sufficient to generate K clusters. Each entry reports the percentage of word-pairs in the dataset evaluated for that K.

in part by grammatical constructions, there is an interesting concordance between form and semantic subsense or conceptual affordance. Cluster 2 for *stream* mostly contains examples of the phrase *stream of X*. This construction has the effect of focusing attention on the substance constituting the stream and emphasizing its movement. Interestingly, the examples in this cluster which do not contain this construction also emphasize the movement of a substance. Compare with *X stream*, where the stream is construed as a medium or channel through which other entities pass.

Cluster ID	Sentences
0	Care must be taken to select the correct neutralizing agent for the specific odorous gas to be treated and there are obvious difficulties when both acidic and alkaline compounds are present in the gas **stream**. Alcohol is absorbed into the blood **stream** via the stomach and takes effect within 5-10 minutes. Every infected **stream** has a 'parent' **stream**, and it may have more than one 'daughter' **stream**. That is to say, the infant must convert stimulation from light rays, sound waves, from the speech **stream** into the appropriate representational grist if it is to get the kind of information that it requires from the world; but this gleaning of information does not constitute thought.
1	Without waiting for the others he plunged down the bank into the **stream**, slipping and slithering heedlessly over the protruding roots and rocks. The horses were quietly cropping the rich grass by the **stream**. Using cement of their own manufacture, they skilfully build tubular houses for themselves out of materials that they pick up from the bed of the **stream**. A small lagoon is formed by the **stream** between a sandbank and the rock wall.
2	Thus a **stream** of pulses lasting 1 second each and given at 10 second intervals could be the 'background' (they could be sound pulses or pulses on a screen, for example); the 'signal' being sought could be the absence of a pulse, one that was shorter or longer than the standard value or one that appeared too soon or too late. Without stopping, the combine disgorged a **stream** of grain into the trailer. Burning straw was the best fun — it was poked through the grill at the front of the grate and, when it caught fire, smoke would **stream** out of the other end. As he ate and drank she found herself chattering away to him out of nervousness, a **stream** of things that went through her head, the small happenings of a day.

Table 4: Example sentences from each *stream* cluster with k=3 and l=8. Cluster 0 is a bit of a catchall cluster, encompassing idioms like 'came on stream', but it contains all mentions of a stream as a medium through which other media pass. Cluster 1 represents the natural feature. Cluster 2 captures, but is not limited to, the usage 'stream of *X*'. Other usages in this cluster share with the construction a focus on the movement of the substance constituting the stream rather than the substance as a medium to move through. Consequently, cluster 2 contains more abstract streams, in that many of its arguments (often count nouns) are not typically thought of as fluids. Note the relationship between grammatical form and semantic subsense / conceptual affordance.

Cluster ID	Sentences
0	In this tale, two weeds grew on a **river** bank; one of them conserved its energy, and grew low and small and brown, with its sights set on a long life, while the other put forth all its strength into growing tall and into colouring itself a beautiful green. The lights dazzled, but on the broad face of the water there were innumerable V shaped eddies, showing the exact position of whatever the **river** had not been able to hide. Across the **river** and through the streets of Cliffe men fought in close combat before the royalists scattered. How can he get all three safely over the **river**?
1	The **River** Doon flows north-west from Dalmellington, past Patna and Dalrymple, under the Auld Brig o' Doon at Alloway, where Tam o' Shanter escaped from pursuing witches in Burns's magnificent poem : The Malá Strana or Lesser Town spreads beneath the castle to the banks of the **River** Vltava. Turkey's Prime Minister Suleyman Demirel arrived in Nakhichevan on May 28 to attend the opening of a bridge between Azerbaijan and Turkey over the **river** Arax. The **river** Sol is the southernmost of the Empire's rivers.
2	A few low hovels that had once been homes to **river** people were now derelict, and an empty building which was once a sailmaker's and then a barge-builder's premises now stood empty after its last owner, a steam-traction engineer, foundered in the changing times. In 1972 the Government of Sind Province declared the **river** dolphin protected by law and prohibited its killing and trapping. As in India, a **river**, a hill, mountain or lake, in Celtic legend, is personified by a god-like person. The same accountants apparently proposed getting rid of **river** wardens and people in pollution control.

Table 5: Four occurrences of *river* in the BNC belonging to each cluster with k=3 and l=8. Cluster 0 corresponds to a river as a natural feature, cluster 1 is captures the construction *river X* where *X* is the name of the river, and cluster 2 contains river used as an adjective to describe things associated with rivers.

Cluster ID	Sentences
0	Homer described it as a monster with the body of a goat, tail of a DRAGON and head of a **lion**, belching flames. With one sister slightly older and another two years younger there were real female ” spats’ at times with the sisters fighting like young **lion** cubs. A creature appeared, a **lion**, red and huge, bounding up the narrow winding streets of Edinburgh, splashing through rivers of blood which poured from the castle. The vertebrates found from that period include mammoth and other extinct elephants, extinct rhino, hippopotamus, giant deer, **lion**, spotted hyaena, tortoise and macaque (from the ‘monkey gravel’ of West Runton, Norfolk — where else?).
1	Finally, let us rekindle that vision in Isaiah 11 where the **lion** does not eat the lamb but lies down in a symbiotic relationship with it. Looking down from a height of ten or twelve feet, she saw an old friend, the MGM **lion**. He was like a man fearing his moment had come, he said, covering his eyes in silent prayer — yet astonished to find the **lion** in the same pose. The peace-keepers successfully tame the roaring **lion**.
2	The Blue Lagoon was the old Red **Lion** renamed, no one knew why, on the corner of Bankside and Trinity Street. Mrs Johnston, 35, was found dead at The **Lion** public house on Moorfields in Liverpool city centre last Thursday lunchtime. It was further established that Bacon had purchased some arsenic from a shop in Red **Lion** Square only days before, allegedly to kill rats. Glaxo sold its factory, as did Gresham **Lion**, and its successor, Dowty, which could not make the business succeed, sold to the Taiwanese.
3	The company, which claims the **lion**’s share of the object database market, has yet to record a profit. If this is done, care must be taken to ensure that each slice receives its proper priority in order of payment, otherwise one party may take the **lion**’s share of the income at the expense of the other. But he gave the **lion**’s share of the credit for the victory to Snodin, playing his first full 90 minutes for two and a half years after a series of hamstring and knee injuries put his career in doubt. Yet last week he had married Magda Tannenbaum, daughter of Sigmund Tannenbaum of Bradford and Hamburg, a wool merchant of legendary wealth, enormous possessions, and no son to inherit the **lion**’s share of them.
4	Three days after Fraser’s departure a large new flag bearing the arms of Dunbar and March, a white **lion** on red, flew from the castle’s topmost tower, indicative that the Earl had arrived. The arms granted to his chosen foundation were the fleur-de-lis of France and the royal **lion** of England, above the three lilies of the Virgin Mary. The sun shone through an elaborate crest of arms in coloured glass, with the **lion** of Venice rampant above a flurry of plumes and a Latin motto, the glass throwing dark Harlequin patterns on to his expressionless face. The green left sleeve brassard carries a red-on-yellow rampant **Lion** of Scotland patch, which we are told is special to the CO and his crew.
5	Henry the **Lion** At Acre, the ramparts of Richard Coeur de **Lion**’s massive fortress stretch down to a tideless Mediterranean while tiny Arab figures promenade in the dusk past the serail. The play was the true story of bachelor Mr Lewis, author of The **Lion**, The Witch And The Wardrobe, and his meeting at the age of 50 with writer Joy Davidman. They sent ambassadors to England to encourage marriage arrangements between two of Henry’s daughters and Henry the **Lion** and one of Barbarossa’s sons.
6	He was one of the first eminent European scientists to make a career in the USA, and rapidly became a **lion** : his lectures and books were popular, and he built up a school and museum at Harvard. She was one fine **lion** and I do n’t blame Raja, only it wasn’t me. He was only too well aware of the Talmudic dictum that a handful does not satisfy a **lion**, but he was neither apologetic nor guilty over it. He looked frightening and she had a momentary sensation of having caged herself in with an angry **lion**.

Table 6: Four tokens from the BNC from each cluster for *lion* with k=7 and l=8. Cluster 0 corresponds to a lion as a wild animal, with respect to other animals and the features which distinguish lions from them. Cluster 1 corresponds to lions interacting with humans and especially acting in ways that are unstereotypically docile. Cluster 2 corresponds to place names containing *Lion*. Cluster 3 corresponds to the idiom *lion’s share*. Cluster 4 corresponds to lions on heraldic coats of arms. Cluster 5 corresponds to human or character names containing *Lion*. Cluster 6 corresponds to metaphorical senses of *lion* to describe a human.

Cluster ID	Sentences
0	Did you believe that, Thomas is part of an eight **cat** routine at Circus International at, it's been at Sutton Coldfield, it's at Northfield over the weekend, but these are just domestic cats, ordinary domestic cats.
1	'**Cat** got yer' tongue?' Paul says he saw a little **cat** swallowing a big dog. The list consisted of an assortment of well known mammals, birds, reptiles, fish and invertebrates, and also included three domestic species : dog, horse and **cat**. I, I know that one hey diddle diddle the **cat** and the fiddle the cow jumped over the moon, the little dog laughed to see such fun and the dish ran away with the spoon
2	He sat looking at the fire with lowered eyelids, a contented expression on his face, looking like a big overfed **cat**. Ecstatic, the boy tried on one mask after another, roaring like a lion or mewing like a **cat**. He took off his clothes, and Isobel curled up on the bed watching him, like a little **cat**. Lacuna was looking like a **cat** that had seen its prey.
3	Obviously, if you are eating more fibre-rich food you are likely to **cat** more grams of fibre.
4	A FAMILY gave up their holiday to pay £700 for a life-saving operation on their **cat** Tilly. She lives in a terraced house in Lancashire with her mum and dad and her **cat**, Arthur. He sat up and stroked the **cat** gently, scratching between the backs of its ears, making Bonaventure purr with pleasure. She took out her pen and paper and wrote a very angry letter to the doctor about the death of her valuable **cat**.
5	A **cat** does not want to die with the smell of humanity in his nostrils and the noise of humanity in his delicate peaked ears. 'All right,' said the **Cat**. After this, he became a nicer **cat**, + was n't so proud of himself all the time. 'There you are !' she said, and the **cat** lifted its tail up with pleasure and rubbed its head against the branch.
6	**Cat**. **Cat**. **Cat**.

Table 7: Four tokens from the BNC from each cluster for *cat* with k=7 and l=8. Cluster 0 corresponds to a unique circus context. Cluster 1 corresponds (loosely) to a cat in relation to other animals. Cluster 2 corresponds to similes likening people to cats by way of stereotypical catlike behavior. Cluster 3 contains a colloquial use of the word to mean 'obtain'. Cluster 4 corresponds to domestic pets. Cluster 5 is not very cohesive, but includes many examples of cats as grammatical subjects or in more agentive roles. Cluster 6 is the word *Cat* in isolation.