Semantic Geometry of Sentence Embeddings

Matthieu Tehenan
Department of Computer Science
University of Cambridge
mm2833@cam.ac.uk

Abstract

Sentence embeddings are central to modern natural language processing, powering tasks such as clustering, semantic search, and retrieval-augmented generation. Yet, they remain largely opaque: their internal features are not directly interpretable, and users lack fine-grained control for downstream tasks. To address this issue, we introduce a formal framework to characterize the organization of features in sentence embeddings through information-theoretic means. Building on this foundation, we develop a method to identify interpretable feature directions and show how they can be composed to capture richer semantic structures. Experiments on both synthetic and real-world datasets confirm the presence of this semantic geometry and highlight the utility of our approach for enhancing interpretability and fine-grained control in sentence embeddings.

matthieu-perso/semantic-geometry
matthieunlp/semantic-geometry

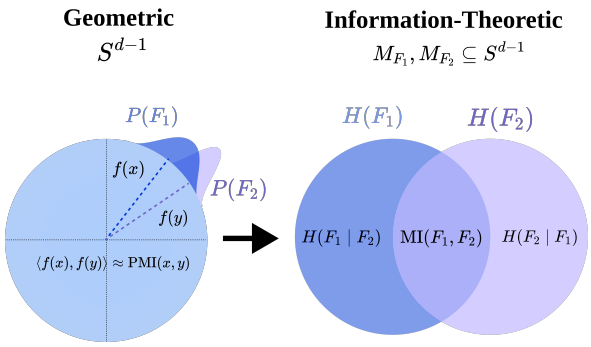

Figure 1: **Geometric–information correspondence.** (Left) On the hypersphere S^{d-1}, sentence embeddings $f(x), f(y)$ encode semantic relations through their inner product $\langle f(x), f(y) \rangle \approx \mathrm{PMI}(x, y)$. These embeddings encode latent features. These latent features F_1, F_2 correspond to regions $M_{F_1}, M_{F_2} \subseteq S^{d-1}$, with probability mass $P(F_i)$. (Right) These regions induce random variables F_1, F_2, whose entropies decompose as $H(F_1), H(F_2)$, with conditional parts $H(F_1 \mid F_2), H(F_2 \mid F_1)$ and overlap $\mathrm{MI}(F_1, F_2)$.

1 Introduction

Sentence embeddings have become a cornerstone of modern AI applications and form the foundation for tasks involving sentence similarity and retrieval (Han et al., 2023). They are essential to semantic search engines, power retrieval-augmented generation, and are core to a wide range of traditional natural language processing tasks, ranging from clustering to paraphrasing (Lewis et al., 2020; Gao et al., 2023). Yet, despite their ability to capture rich semantic relationships, tasks on sentence embeddings are typically reduced to a single scalar similarity score (Chandrasekaran and Mago, 2021; Opitz et al., 2025a). This practice obscures the internal structure of embeddings: the features and components that drive similarity remain hidden. This limits both performance and interpretability. Users lack fine-grained control over operations and cannot interpret why two sentences are considered similar or dissimilar. Understanding this internal structure is not only key to improve performance on downstream tasks, but also to enhance transparency, safety, and alignment with human values (May et al., 2019; Zou et al., 2023).

The internal structure of sentences has long been a central topic in semantics, with a focus on how meaning arises from constituent parts. At the core of this tradition is the principle of compositionality, which holds that the meaning of a sentence is a function of the meanings of its components and their syntactic combination (Partee et al., 1984; Mitchell and Lapata, 2010). Early approaches to sentence representations reflected this idea through additive or compositional models of features (Liu et al., 2023). More recently, distributional semantics has shifted this inquiry into vector spaces, where sentence embedding models capture emergent semantic structures consistent with hypotheses such as distributional inclusion (Geffet and Dagan,

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11993–12004
November 4-9, 2025

2005). Modern sentence embedding models have been based on transformer models, whose strength been the learning of intermediate representations, which encode deeper features (Bengio et al., 2014; Goodfellow, 2016). The complexity of these spaces however, as led to difficulties in identifying these features at scale (Opitz et al., 2025b). It is also not sure if these features are interpretable or map to compositional units which emerge as by-products of training. Although a semantic geometry and organization of features has been clearly identified with word vectors, such as not yet been the case for sentences(Mikolov et al., 2013; Pennington et al., 2014; Ethayarajh et al., 2019).

In this paper, we propose a framework to characterize the semantic organisation of sentence representations and its interpretability. On this basis, we make the following contributions:

1. First, we provide a formal framework to describe the semantic organization of features in sentence embeddings.
2. Based on the above, we provide a tractable method to identify features and their compositions.
3. We empirically validate our approach across both synthetic and real-world datasets.

2 Background

In this section, we provide background on modern sentence embedding models, their training and interpretability.

2.1 Representing Sentence Similarity

Modern sentence embedding models aim to generate representations that encapsulate the semantic meaning of sentences (Wieting et al., 2015; Li et al., 2020). Within computational linguistics, this is typically grounded in distributional semantics, where the meaning of a sentence is defined by the sentences with which it is similar. In this context, learning the meaning of a sentence can be reframed as modeling sentence similarity across a representative corpus. Importantly, similarity can take multiple forms (Liu et al., 2020; Kashyap et al., 2023). Sentences may exhibit similarity through paraphrase, translation, or entailment relations. Contemporary models capture these diverse forms of similarity within a unified representation space, enabling them to represent meaning more comprehensively and to generalize more effectively across linguistic tasks.

Despite reflecting different kinds of relationships, these similarities are typically reduced to a single similarity score. This score is most often computed using cosine similarity (Deerwester et al., 1990; Lin et al., 1998), a geometric formulation that aligns with graded human judgments of semantic relatedness, where similarity scores range from -1 (maximally dissimilar) to 1 (maximally similar) (Tversky, 1977; Chandrasekaran and Mago, 2021). In other words, higher similarity score reflects a greater degree of shared meaning.To enable consistent comparison within a space that naturally supports the cosine similarity metric, sentence embeddings are typically normalized to unit length (Kashyap et al., 2023; Gao et al., 2021). This constrains them to the surface of a unit hypersphere $S^{d-1} \subset R^d$, where d is the dimensionality of the embedding space. Under this normalization, the cosine similarity between two embeddings $\mathbf{u}$ and $\mathbf{v}$ simplifies to their dot product, allowing the similarity measure to depend solely on the direction of the vectors rather than their magnitude.

2.2 Training Sentence Embeddings

To effectively model sentence similarity, contemporary sentence embedding models rely on pretrained transformers, which are then fine-tuned on sentence-level data (Conneau et al., 2017; Reimers, 2019; Gao et al., 2021). Given an input sentence composed of N tokens, these models produce contextualized token embeddings $h_1, h_2, \ldots, h_N$, where each $h_i \in R^d$ and d is the dimensionality of the embedding space (Kashyap et al., 2023). To derive a fixed-size sentence embedding, a pooling operation such as averaging, or attention-based weighting is applied over these token embeddings (Reimers, 2019; Gao et al., 2021). The contextual nature of the embeddings, combined with token positional information, already gives strong baseline results on sentence similarity tasks (Arora et al., 2017). While this captures an approximate meaning of a sentence, it does not inherently reflect syntactic structure, logical form, or semantic implications. To address these limitations, a fine-tuning step is commonly applied to the pooled sentence embeddings (Gao et al., 2021; Chuang et al., 2022; Kashyap et al., 2023). Most modern frameworks take a variant of Information Noise Contrastive Estimation (InfoNCE). The objective minimizes a variant of the following loss over batches of size N, where x_i is an anchor, x_i^+ its positive pair, x_j a negative sample, and τ the temperature parameter:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(x_i, x_i^+)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i, x_j)/\tau)} \quad (1)$$

The goal is to refine the sentence embedding space such that semantically similar sentences form tight clusters, while unrelated or contradictory sentences are mapped further apart (Kashyap et al., 2023). To account for the various senses of similarity mentioned above, these models are trained on various sentence level datasets such as paraphrase, similarity or NLI (Conneau and Kiela, 2018; Bowman et al., 2015; Williams et al., 2017).

2.3 Interpretability

Once trained, sentence embedding models acquire a rich representational capacity, yet the internal structure of these representations remains opaque (Opitz et al., 2025a). Early approaches constructed sentence embeddings by summing or averaging word vectors (Mitchell and Lapata, 2010), with extensions such as Doc2Vec introducing paragraph-level vectors learned alongside words (Le and Mikolov, 2014). While more directly interpretable, these methods lacked expressive power.

The use of modern transformer architectures proved a considerable step forward, but at the cost of opaque representations (Reimers, 2019). Early approaches at interpretability have relied on probing, testing whether specific features (e.g., syntax, part-of-speech, entailment) can be linearly recovered from embeddings (Conneau et al., 2017; Hewitt and Manning, 2019; Nikolaev and Padó, 2023b,a). While probing demonstrates the presence of information, it does not explain how these features are organized within the embedding space. Other lines of work attempt to disentangle embeddings into more interpretable components, using techniques such as variational methods or feature decomposition (Opitz and Frank, 2022; Sun et al., 2024) . These approaches can recover partial structure but are often limited to small feature sets and do not scale reliably to the full sentence meaning (Huang et al., 2021). Despite these efforts, there is still no clear account of the representational geometry of sentence embeddings and how features are organized and compose.

3 The Structure of Sentence Embeddings

In the previous section, we saw how sentence representation models aimed at capturing semantic similarities. In the following section, we will study the emergent spaces that emerge from this objective.

3.1 Overview

We begin by formalizing the link between contrastive learning and mutual information. Sentence embedding models trained with the contrastive loss $\mathcal{L}$ optimize similarities that approximate a PMI kernel.

Proposition 3.1 Let $\mathcal{L}$ be the contrastive objective and let $f : \mathcal{S} \to S^{d-1}$ denote the embedding function of a model trained with $\mathcal{L}$. For any two sentences $x, y \in \mathcal{S}$, the inner product between their embeddings satisfies $\langle f(x), f(y)\rangle \propto \text{PMI}(x, y) + c$ for some additive constant c.

This result shows that inner products in the embedding space can be rewritten in terms of information content. When negatives are sampled correctly, the embedding space provides a low-rank approximation of the true PMI distribution, constrained to lie on the unit hypersphere. Thus, the inner product between two embeddings corresponds to the strength of their statistical association in the data distribution. Rewriting the inner product allows us to understand the representational space of sentence embeddings.

Definition 3.2 We define the **sentence representation space** as the kernel space $(S^{d-1}, \langle\cdot,\cdot\rangle)$ induced by the mapping $f : \mathcal{S} \to S^{d-1}$, where sentence embeddings satisfy $\langle f(x), f(y)\rangle \approx \text{PMI}(x, y)$.

Proposition 1 and Definition 1 together imply that the sentence representation space is, up to scaling and shift, an inner product space whose kernel approximates pointwise mutual information. Formally, the embeddings $f(x) \in S^{d-1}$ inhabit a vector space $(R^d, \langle\cdot,\cdot\rangle)$, so all algebraic operations are inherited from the ambient space. Since inner products correspond to PMI values, these vector space operations (e.g. scalar multiplication or addition) can be interpreted as acting on statistical quantities. addition of embeddings aggregates. This in turn helps us reinterpret a single embedding x as below.

Observation 3.3 Fix $x \in \mathcal{S}$. The PMI kernel satisfies $\text{PMI}(x, y) \approx \langle f(x), f(y)\rangle \quad \forall y \in \mathcal{S}$.

Thus the embedding $f(x)$ is a finite-dimensional code for the function $\text{PMI}(x, \cdot)$, i.e. the entire row of PMI values indexed by x. In this sense, a single

embedding can be interpreted as a compressed representation of all information-theoretic associations of x with the rest of the space.

3.2 Decomposition

We now move from the PMI kernel to its internal decomposition into features. Intuitively, pointwise mutual information is additive under the chain rule of mutual information, and this additive structure carries over to the kernel formulation.

Definition 3.4 A **feature** F is a latent factor that recurs across sentences and captures a recognizable semantic or structural aspect (e.g., concepts, tenses, or higher level semantic structures).

Proposition 3.5. Let $x \in \mathcal{S}$ be a sentence with embedding $f(x)$, and let $(F_1, \ldots, F_n)$ denote the latent features that compose $f(x)$. Then for any $y \in \mathcal{S}$, $\mathrm{PMI}(x, y) = \sum_{i=1}^{n} \mathrm{PMI}(F_i, y \mid F_{<i})$ with $F_{<i} = (F_1, \ldots, F_{i-1})$.

Being shared across sentences, latent features can be understood as conditional contributions to pointwise mutual information. In particular, the PMI between two sentences x and y decomposes additively into the contributions of the features $\{F_i\}$ that compose x. Each term $\mathrm{PMI}(F_i, y \mid F_{<i})$ quantifies the incremental information carried by feature F_i about y, given the previous features. Thus, the PMI kernel admits a structured additive factorization in terms of latent feature contributions.

Observation 3.6 The embedding $f(x)$ admits a structured decomposition into latent features. Specifically, each coordinate $f_i(x)$ reflects the strength of feature F_i in sentence x, and the similarity with another sentence y decomposes as $\langle f(x), f(y) \rangle = \sum_{i=1}^{d} f_i(x)\, f_i(y)$. Thus, a single embedding $f(x)$ can be interpreted as a weighted combination of latent features $\{F_i\}$, whose alignment with another embedding determines their PMI.

3.3 Points and Regions

So far, we have treated embeddings as points: each embedding $f(x)$ encodes a weighted decomposition of PMI values with all other sentences. We now extend this view from individual points to *regions* of the embedding hypersphere S^{d-1}, corresponding to sets of sentences that share a common latent feature. We thereby transition from individual point measurements to mutual information between regions.

Definition 3.8 (semantic region). Let $p(x)$ denote the global data distribution over sentences on the hypersphere S^{d-1}. For a latent feature F_i, let $p_{F_i}(x)$ denote the conditional distribution of embeddings given that feature F_i is active. For a threshold $\tau \in (0, 1)$, the associated **semantic region** is defined as

$$M_{F_i}(\tau) = \{\, x \in S^{d-1} : p_{F_i}(x) \geq \tau \,\}.$$

Thus $M_{F_i}(\tau)$ is the subset of the hypersphere where the feature F_i is active with probability at least τ. The probability mass of this region under the global distribution is $P(M_{F_i}(\tau)) = \int_{M_{F_i}(\tau)} p(x)\, dx$,i.e. the probability that a random sentence sampled from $p(x)$ instantiates feature F_i above the chosen threshold. This construction allows us to quantify mutual information between two latent features F_i, F_j in terms of their joint and marginal regions $M_{F_i}(\tau), M_{F_j}(\tau)$.

This allows us to quantify the mutual information between two latent features F_i, F_j, which is determined by the joint and marginal regions M_{F_i}, M_{F_j}. As a consequence, the collection of semantic regions $\{M_{F_i}\}$ admits set-theoretic relations encoding information-theoretic dependencies. Amongst these, we can identify *inclusion:* $M_{F_i} \subseteq M_{F_j}$ when the presence of feature F_i entails F_j (e.g., "corgi" $\subseteq$ "dog"); *exclusion:* $M_{F_i} \cap M_{F_j} =$ when features never co-occur; *overlap:* $M_{F_i} \cap M_{F_j} \neq$ when sentences instantiate both features simultaneously.

In practice, estimating these probability masses is challenging: the regions are high-dimensional, sample sizes are limited, and many features appear entangled rather than cleanly separated. In this sense, the learned PMI kernel defines not only pairwise similarities but also an organized geometry of semantic sets based on mutual information. The challenge is to reliably identify these regions and disentangle overlapping features in high-dimensional space.

4 Identifying Features

The previous section established that the PMI kernel induces a structured decomposition of sentence embeddings into latent features. We now turn to the problem of identifying these features in a tractable and interpretable way.

4.1 Identifying Features

If we collect a set $S_F \subset \mathcal{S}$ of sentences that instantiate a semantic feature F (e.g., all sentences containing the feature "dog as subject"), their embeddings $\{f(x) : x \in S_F\}$ form a semantic region $M_F \subseteq S^{d-1}$, representing the distribution of that feature on the hypersphere. By Proposition 3.1, embeddings satisfy $\langle f(x), f(y) \rangle \approx \mathrm{PMI}(x, y)$. If two sentences $x, y \in S_F$ share feature F, their PMI is high, which forces their embeddings to lie close together on the hypersphere. Even if each sentence contains other features, the set $\{f(x) : x \in S_F\}$ is still expected to cluster around a common direction.

A natural idealization is therefore that embeddings associated with F are distributed according to a von Mises–Fisher (vMF) distribution, the analogue of a Gaussian distribution for directional data on the hypersphere. Under this assumption, the maximum-likelihood estimator of the feature direction is simply the normalized sample mean (Mardia and Jupp, 2009).

Proposition 4.1 Let F be a latent feature and let $S_F \subset \mathcal{S}$ be a set of sentences instantiating F. If embeddings in S_F follow a vMF distribution, then the normalized centroid $\hat{\mu}_F$ is the maximum-likelihood estimator of the true mean direction μ_F of feature F.

The estimator $\hat{\mu}_F$ converges to the true direction μ_F as $|S_F| \to \infty$. While this provides a point estimate, the underlying distribution can also be modeled more richly.

Observation 4.2 Let p_F denote the distribution of feature F on the unit hypersphere. The distribution p_F may be further approximated by incorporating measures of concentration or local density estimates around $\hat{\mu}_F$.

Observation 4.3 Each feature corresponds to a direction μ_{F_i} on the hypersphere, and a sentence embedding can be decomposed with respect to these directions. Geometrically, the projection $\langle f(x), \mu_{F_i} \rangle$ reflects the degree to which sentence x expresses feature F_i. Thus, embeddings $f(x)$ can be interpreted as structured combinations of latent features $\{F_i\}$, and similarity between sentences arises from the alignment of their feature projections.

4.2 Compositionality of Features

The identification of feature directions (Proposition 4.1) provides a basis for identifying desired features in the representation space. Beyond identification, the inner product structure of the PMI kernel endows the space with algebraic operations that naturally support compositionality. That is, new features can be constructed from existing ones through vector-space operations, which can reflect the way linguistic meaning arises from the composition of semantic parts.

Proposition 4.4 Let $\{\hat{\mu}_{F_i}\}_{i=1}^{k}$ be the maximum-likelihood feature directions estimated for a set of features. Because the representation space is an inner product space, any linear combination $\hat{\mu}_G = \sum_{i=1}^{k} \alpha_i \hat{\mu}_{F_i}$ defines a new direction corresponding to a composed feature G.

Geometrically, this operation corresponds to the additive structure of the PMI kernel, where the contribution of a composite feature is expressed as the sum of its constituent PMI terms. In practice, this allows, for instance, the combination of feature directions corresponding to "South America" and "North America" into a more general feature direction "America."

Observation 4.5 The additive structure of PMI, well established in word embeddings, then extends to sentence embeddings (Pennington et al., 2014; Arora et al., 2016; Allen and Hospedales, 2019; Ethayarajh et al., 2019). This provides a foundation for the linguistic regularities observed in sentence-level representations (Zhu and de Melo, 2020). In particular, analogical reasoning could arise directly from the linear geometry of feature directions. If F_1 and F_2 are related features (e.g., *dog* and *puppy*), their difference $\hat{\mu}_{F_1} - \hat{\mu}_{F_2}$ encodes the information difference between them, which can be transferred to other contexts.

4.3 Composition of Feature Distributions

Feature identification characterizes individual features as directions on the hypersphere. In practice, however, we are often interested not only in single directions but in the semantic regions of the embedding space where such features are expressed with high likelihood. We will then focus on compositional feature distributions.

Observation 4.6 As discussed above, each feature F induces a probability distribution p_F on the hypersphere, with an associated region $M_F(\tau) =$

$\{ x \in S^{d-1} : p_F(x) \geq \tau \}$ where τ is a threshold of activation. Composed features can then be modeled by combining distributions; mixtures approximate unions of regions ($M_G \approx \bigcup_i M_{F_i}$) and products approximate intersections ($M_G \approx \bigcap_i M_{F_i}$).

These operations on distributions extend the linear operations on feature directions to full semantic regions. Thus, addition, subtraction, or scaling of feature directions corresponds to unions, differences, or rescalings of their semantic regions, capturing relations such as inclusion (*corgi* $\subset$ *dog* $\subset$ *animal*) or contrast (*South America* vs. *North America*). This allows the construction of complex features as sequences of operations on more basic ones. This allows one to localize subsets of the hypersphere where specified combinations of features are present (or absent) with the desired probability. As a semantic region on a subspace is by nature of the ambient space constrained on a subspace, we can also represent these regions as a subspace if easier for downstream tasks.

5 Experiments

We now present experiments to empirically test the properties introduced above. Our full code can be found on GitHub, and the dataset can be accessed on Hugging Face. Additional implementation details are provided in Appendix B and further results in Appendix C .

5.1 Setup

We evaluate our approach using a selection of high-performing sentence embedding models and two datasets.

Models To ensure robust evaluation, we experiment with four widely used sentence embedding models available on Hugging Face: `gte-large-en-v1.5` (Alibaba-NLP), `all-mpnet-base-v2` and `all-MiniLM-L6-v2` (Sentence-Transformers), and `multilingual-e5-large-instruct` (IntFloat). These models were chosen for their strong performance on retrieval tasks (Muennighoff et al., 2022), as well as their widespread adoption in both academic research and industry applications.

Datasets We evaluate our method on two datasets: a controlled synthetic dataset based on WordNet, and a real-world dataset, TREC. **(1) WordNet**: We construct a dataset by generating sentence variants in which a target word is systematically replaced with its hypernyms and hyponyms in Wordnet (Miller, 1995). This controlled manipulation allows us to isolate and assess the impact of semantic abstraction on the representation space. **(2) TREC**: For evaluation on real-world data, we use the TREC question classification dataset (Dietz et al., 2017), which organizes questions into coarse and fine-grained conceptual categories. This hierarchical labeling enables us to analyze how semantic structure and granularity are reflected in the embedding space.

5.2 Existence of Semantic Regions

Experiments. We begin with the hypothesis that concepts are organized within semantic regions. To empirically assess this statement, we test whether the prototype representation of a feature F is closest to its peripherals than the peripherals amongst themselves. In other words, an order relation holds on the set, where a hyperonym A embedding is more similar to its hyponyms B_i than the hyponyms are to each other. We perform this test across both of our datasets. In the WordNet hierarchy, we would expect that the sentence "There is a corgi" should be more similar to the prototype sentence "There is a dog" than to the more specific "There is a German Shepherd".

Results Table 2 summarizes these findings for our WordNet dataset, showing consistency across different models and sentence embeddings. The prototype embedding is consistently closer to its hypernyms than the hyponyms are to each other and is also positioned nearer to the centroid of the data. Figure 2 and Figure 3 illustrates that the prototype sentence embedding for the synset "dogs" is typically centrally located and exhibits the lowest pairwise similarity variance within the synset. Despite variations in cosine similarity with the hyponyms, the hypernym consistently maintains a higher position in the hierarchy. Similar to the trends observed in Table 2, we find that sentence clusters exhibit meaningful structure: the category centroid consistently ranks among the top 5% closest embeddings, but rarely amongst the top 0.5%. This is expected given the presence of a large variety of hypernyms in each sentence. The noise could then be explained by the intersection of overlapping semantic sets. We also note that the smallest models, such as `all-MiniLM-L6-v2`, exhibit lower representational capacity compared to the

	gte-large-en-v1.5			all-mpnet-base-v2		
Feature	N+1	N+2	N+3	N+1	N+2	N+3
mammal.n.01	0.7050	0.6800	0.6500	0.6100	0.5900	0.5700
food.n.01	0.7150	0.6700	0.6450	0.6050	0.5800	0.5500
plant.n.02	0.7200	0.7000	0.6700	0.6000	0.5900	0.5700
institution.n.01	0.7350	0.6900	0.6750	0.6200	0.6000	0.5800
cognition.n.01	0.7100	0.6650	0.6400	0.5900	0.5700	0.5400
	all-MiniLM-L6-v2			multilingual-e5-large-instruct		
Feature	N+1	N+2	N+3	N+1	N+2	N+3
mammal.n.01	0.6050	0.5800	0.5600	0.6150	0.5900	0.5700
food.n.01	0.5950	0.5700	0.5400	0.6100	0.5800	0.5500
plant.n.02	0.6000	0.5800	0.5600	0.6200	0.6000	0.5700
institution.n.01	0.6150	0.5900	0.5800	0.6050	0.5900	0.5650
cognition.n.01	0.6000	0.5700	0.5500	0.6100	0.5800	0.5600

Table 1: Comparison of hierarchical levels $N+1$ to $N+3$ for our selected sentence embedding models. The order relation is broadly preserved across the model.

	gte-large-en-v1.5		all-mpnet-base-v2		all-MiniLM-L6-v2		multilingual-e5-large-instruct	
Feature	Closest	Centroid	Closest	Centroid	Closest	Centroid	Closest	Centroid
mammal.n.01	80.84	75.00	66.93	62.50	68.98	75.00	71.00	75.00
food.n.01	72.56	100.00	69.32	87.50	67.79	87.50	69.38	87.50
plant.n.02	72.84	62.50	60.49	87.50	69.69	75.00	74.49	75.00
cognition.n.01	73.21	37.50	49.23	25.00	51.73	25.00	43.81	25.00

Table 2: Comparison of closest-match and centroid-based similarity scores across sentence embedding models for various synsets. The "Closest" column reports the percentage of cases in which a hyponym sentence (e.g., The corgi is running) is most similar to one of its hypernyms (e.g., The dog is running). The "Centroid" column reports the percentage of cases in which the hypernym sentence is closest to the centroid of all sentences in its semantic category. Note that the hypernym sentence is not necessarily the centroid. Higher percentages indicate stronger alignment with the expected hierarchical semantic structure.

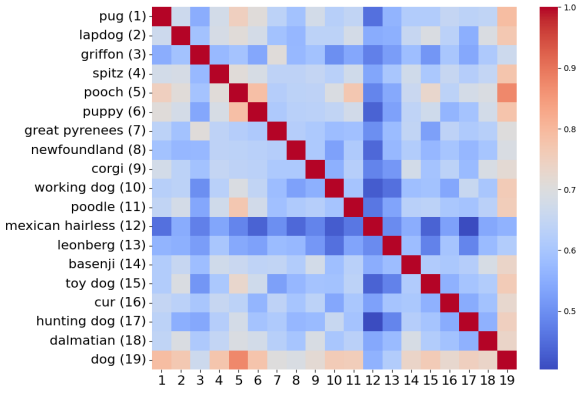

Figure 2: Heatmap of cosine similarity between the WordNet hypernym sentence in row 19 ("dog") and its corresponding hyponyms in the synset "dog", computed using the gte-large-en-v1.5 model. "Dog" is not a semantic region for this set of embeddings.

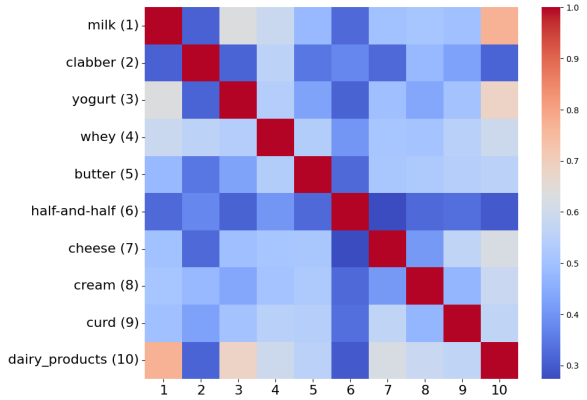

Figure 3: Heatmap of cosine similarity between the WordNet hypernym sentence in row 10 ("dairy products") and its corresponding hyponyms, computed using the all-mpnet-base-v2 model.

`gte-large-en-v1.5`. This disparity likely reflects differences in their ability to encode and structure information, with smaller models constrained by a more compressed geometric representation.

5.3 Hierarchy of Semantic Regions

Experimental Setup. In this section, we empirically evaluate the set-theoretic operations introduced earlier. When an embedding belongs to multiple semantic regions, it participates in several independent ordering structures defined by the centroids of these regions. A notable example is

TREC Coarse Class	gte-large-en-v1.5	all-mpnet-base-v2	all-MiniLM-L6-v2	multilingual-e5-large-instruct
ABBR	0.74	0.72	0.68	0.76
DESC	0.72	0.70	0.66	0.74
ENTY	0.68	0.66	0.61	0.71
HUM	0.78	0.75	0.70	0.80
LOC	0.77	0.74	0.69	0.79
NUM	0.69	0.67	0.62	0.72

Table 3: TREC hierarchy preservation across our four sentence-embedding models. Each value is the mean accuracy with which the expected order holds, i.e. that is, a sentence is more similar to its own fine-grained centroid than to its coarse parent.

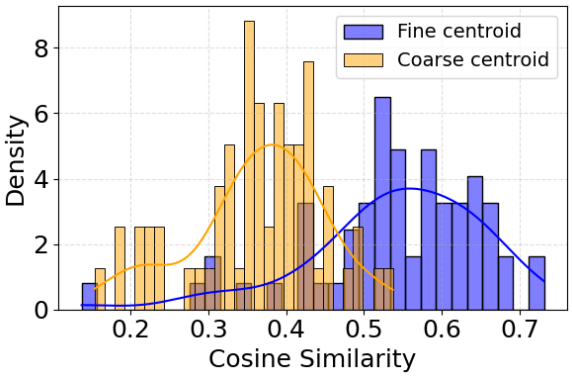

Figure 4: Cosine similarity to fine and coarse centroids for the Human class in the TREC dataset.

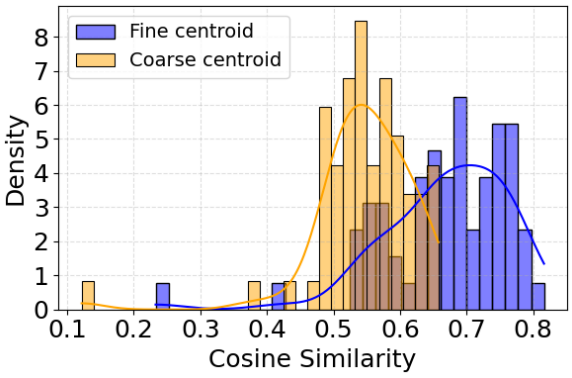

Figure 5: Cosine similarity to fine and coarse centroids for the Entity class in the TREC dataset.

lexical hierarchy, where a hyperonym (e.g., A, A') encompasses a set of hyponyms (B_i). In such cases, for hyponyms A and its associated hyponyms B_i, we expect: $B_i \preceq A \preceq A' \preceq A''$ where each hyperonym A is positioned above its hyponyms and below more abstract categories in the hierarchy. To test this hypothesis, we analyze three levels within the WordNet hierarchy and calculate the percentage of decreasing pairs. We do the same on the TREC dataset between coarse and fine categories.

Results. Table 1 presents the proportion of cases where this decreasing trend holds. While the expected pattern is observed, there are exceptions. For instance, "animal" is often more similar to related words than "mammal," reflecting the impact of co-occurrence frequencies in the dataset. Additionally, words with low co-occurrence probabilities tend to deviate from the expected order, suggesting that embeddings capture intersecting semantic structures rather than a strict hierarchy. Nevertheless, the transitive ordering among hyperonyms remains present across models. In our real world dataset, the structure is partially preserved, as seen in Table 3. As shown in Figure 4 and Figure 5, sentence embeddings tend to cluster more tightly around their fine-level centroids than their coarse-level counterparts, as evidenced by a rightward shift in the blue distribution (fine centroid) compared to the orange (coarse centroid). This indicates that fine classes are embedded more precisely, likely due to their narrower semantic scope. However, we also observe substantial overlap between the distributions, especially in regions where fine and coarse labels are densely populated. This suggests interference effects from semantically adjacent classes.

5.4 Method Comparison and Baseline

Experimental Setup. To assess the effectiveness of our method, we compare it against a widely used baseline -linear probing (Belinkov and Glass, 2019; Hewitt and Manning, 2019). Specifically, we implement a multi-prototype extension of our centroid-based interpretability framework. For each class (coarse and fine), we compute $k = 3$ cluster centroids over the normalized sentence embeddings using KMeans. If our hypothesis holds, these centroids represent a more general, common representation of the set. Each centroid is ℓ_2-normalized and stored in a prototype bank. Zero-shot classification is then performed by assigning each sample

to the label of its nearest prototype based on cosine similarity. In parallel, we train a multinomial logistic regression classifier on the same embeddings using an 80/20 stratified train-test split. Accuracy is evaluated on the held-out set for both coarse- and fine-grained classification.

Results. The multi-prototype approach achieves 0.792 accuracy on coarse labels and 0.714 on fine labels. In comparison, the linear probe yields 0.885 accuracy on coarse labels and 0.712 on fine labels. While the probe performs slightly better on coarse classification, our method matches it on fine-grained prediction despite requiring no parameter training. Importantly, the entire pipeline executes in under 5% of the time required to train the probing classifier. Appendix D discusses the computational requirements of our method.

6 Discussion and Conclusion

A central insight of this work is that even simple operations, i.e. identifying a prototype of a feautre F, can reveal meaningful structure in sentence embeddings. Our results suggest that sentence embeddings are not uniformly dispersed but rather organize into latent semantic regions, with meaningful ordering and overlap. If such structure can be identified reliably, it opens the door to scalable, per-model interpretability: semantic clusters could be extracted automatically, visualized on the hypersphere, and used to explain similarity, entailment, or category membership in embedding-based systems. Understanding the internal geometry of sentence embeddings is critical for modern information retrieval and NLP pipelines, particularly as AI-augmented applications, such as semantic search, retrieval-augmented generation, and recommendation, where safety and ethical applications. are paramount. The uncovered representational geometry could also be relevant across other disciplines. It notably bears a striking resemblance to cognitive theories of similarity, reflecting Tversky's work on feature-based similarity (Tversky, 1977) or conceptual spaces (Gardenfors, 2004; Osta-Vélez and Gärdenfors, 2020; Douven et al., 2023; Douven and Verheyen, 2024).

Limitations

While our study studies the foundations of embedding representational geometry, it is not without limitations. The number of semantic sets within the dataset poses a challenge. A truly comprehensive analysis would require identifying a core set of dominant clusters that encapsulate the majority of the structural variance in the embeddings. Future work should aim to map these principal sets, as they are likely to drive both the overall structure and the interpretability of the embedding space. This implies that a simple model may be insufficient to capture the nuances of the embedding space. We suggest for future research to explore models that can more accurately represent the variability and uncertainty intrinsic in these regions.

References

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2014. Representation Learning: A Review and New Perspectives. *arXiv preprint*. ArXiv:1206.5538 [cs].

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised

learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Igor Douven and Steven Verheyen. 2024. Concept learning: Convexity versus connectedness. *Erkenntnis*, pages 1–18.

Igor Douven, Steven Verheyen, Shira Elqayam, Peter Gärdenfors, and Matías Osta-Vélez. 2023. Similarity-based reasoning in conceptual spaces. *Frontiers in Psychology*, 14:1234483.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards Understanding Linear Word Analogies. ArXiv:1810.04882 [cs].

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Peter Gardenfors. 2004. Conceptual spaces as a framework for knowledge representation. *Mind and matter*, 2(2):9–27. Publisher: Imprint Academic.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114.

Ian Goodfellow. 2016. Deep learning.

Yikun Han, Chunjiang Liu, and Pengfei Wang. 2023. A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*.

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.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling Semantics and Syntax in Sentence Embeddings with Pre-trained Language Models. ArXiv:2104.05115 [cs].

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2023. A comprehensive survey of sentence representations: From the bert epoch to the chatgpt era and beyond. *arXiv preprint arXiv:2305.12641*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Dekang Lin et al. 1998. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2023. *Representation learning for natural language processing*. Springer Nature.

Kanti V Mardia and Peter E Jupp. 2009. *Directional statistics*. John Wiley & Sons.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Dmitry Nikolaev and Sebastian Padó. 2023a. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. *arXiv preprint arXiv:2310.11923*.

Dmitry Nikolaev and Sebastian Padó. 2023b. Representation biases in sentence transformers. *arXiv preprint arXiv:2301.13039*.

Juri Opitz and Anette Frank. 2022. SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Juri Opitz, Lucas Möller, Andrianos Michail, and Simon Clematide. 2025a. Interpretable text embeddings and text similarity explanation: A primer. *arXiv preprint arXiv:2502.14862*.

Juri Opitz, Lucas Möller, Andrianos Michail, and Simon Clematide. 2025b. Interpretable Text Embeddings and Text Similarity Explanation: A Primer. ArXiv:2502.14862 [cs].

Matías Osta-Vélez and Peter Gärdenfors. 2020. Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*, 96:102357.

Barbara Partee et al. 1984. Compositionality. *Varieties of formal semantics*, 3:281–311.

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.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony KH Tung, and Jun Yu. 2024. A general framework for producing interpretable semantic text embeddings. *arXiv preprint arXiv:2410.03435*.

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

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Xunjie Zhu and Gerard de Melo. 2020. Sentence Analogies: Linguistic Regularities in Sentence Embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Model Information and License

We utilized the following pretrained sentence embedding models, all available on Hugging Face:

- **GTE-large-en-v1.5**: https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5
  (Alibaba-NLP)
  License: Apache 2.0
- **all-mpnet-base-v2**: https://huggingface.co/sentence-transformers/all-mpnet-base-v2
  (Sentence-Transformers)
  License: Apache 2.0
- **all-MiniLM-L6-v2**: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
  (Sentence-Transformers)
  License: Apache 2.0
- **multilingual-e5-large-instruct**: https://huggingface.co/intfloat/multilingual-e5-large-instruct
  (IntFloat)
  License: MIT

B Experimental Details

B.1 Hardware Setup

All experiments were conducted on NVIDIA A100 GPUs. The training and evaluation were carried out using PyTorch on a machine hosted on RunPod.

B.2 Preprocessing

Text data used in this study were preprocessed as follows:

- Tokenization was performed using the Hugging Face tokenizer for each pretrained model.
- Sentences were embedded using the sentence transformers model for clustering and similarity tasks.

B.3 Evaluation Metrics

For our evaluation, we used the following metrics:

- **Cosine Similarity**: To measure the closeness between embeddings generated by different models.

- **Clustering Accuracy**: For evaluating the quality of semantic grouping using K-means clustering on sentence embeddings.

B.4 Confidence Intervals.

To ensure robustness of our results, we compute 95% confidence intervals using bootstrapping with 1,000 resamples. This allows us to quantify the variability of the probing performance and determine whether observed differences across models or configurations are statistically significant.

C Datasets

WordNet: To investigate the existence of semantic regions, we construct a structured dataset derived from WordNet, a large-scale lexical database that encodes hierarchical relationships between concepts (Miller, 1995). Specifically, we define a set of four root concepts— mammal, food, plant and cognition—which serve as central semantic anchors, unifying their respective hyponym sets. To generate meaningful linguistic contexts, we construct a diverse set of sentences incorporating these terms in varying syntactic and semantic configurations. Following the criteria outlined in Section 3.3, we select a subset of hyperonyms where the frequency of the hyperonym exceeds that of its hyponyms. We then systematically create sentence pairs using these terms. Our data can be found on our Huggingface link `https://huggingface.co`.

D Computational Complexity

Let N be the number of samples, d the embedding dimension, and C the number of classes.

Centroid Method. Class centroids are computed as:

$$\mu_c = \frac{1}{|S_c|} \sum_{i \in S_c} E_i$$

with a cost of $O(Nd)$. Label assignment is performed by computing cosine similarity to each class centroid, costing $O(NCd)$. **Total complexity:** $O(Nd + NCd)$.

Linear Probing. A weight matrix $W \in R^{C \times d}$ is trained using gradient descent over T epochs and N samples, resulting in a training cost of $O(T \cdot NCd)$. Inference similarly requires $O(NCd)$. **Total complexity:** $O(T \cdot NCd)$.

Summary. Centroid-based inference is 1–2 orders of magnitude more efficient than linear probing methods (Belinkov and Glass, 2019; Hewitt and Manning, 2019). It requires no parameter training and only a single forward pass through the data, making it viable for zero-shot interpretability with minimal computational cost.