Understanding Contextual Meaning in Transformer Embeddings

BIONB 3500 Project Proposal

LAWRENCE GRANDA ZARZUELA*, Cornell University. BIONB 3500, USA

OVERVIEW

Transformer-based language models such as BERT and DistilBERT produce contextualized word embeddings. Given an input sentence, these models output a sequence of high-dimensional vectors (one per token) representing how each word's meaning is shaped by its surrounding context. For example, the word *bank* appears in very different contexts: *river bank*, *finance bank*, or *seating bank*. Although the surface form is identical, we expect the model to internally represent these occurrences differently depending on their meaning.

Our goal is to investigate how these embeddings reflect context-dependent meaning and whether transformer models implicitly separate different senses of the same word. Specifically, we will analyze how the embeddings of ambiguous words distribute in vector space across multiple contexts. We expect embeddings of the same sense (e.g., *river bank*, *muddy bank*) to cluster closely together, while embeddings from different senses (e.g., *bank account* vs. *river bank*) should be far apart.

HYPOTHESES

(1) **H1: Unsupervised Sense Induction.** If a model truly learns to represent meaning, then the embeddings of the same word used in different contexts should naturally form clusters corresponding to distinct senses, even without explicit supervision.

(2) **H2: Layer-wise Semantic Specialization.** Lower layers of the transformer encode surface and syntactic information, while deeper layers capture more abstract, semantic representations. We hypothesize that sense differentiation emerges and peaks in the middle-to-upper layers, before the final layers become too specialized or task-tuned.

METHODS

We will extract contextual embeddings for polysemous words (e.g., *bank*, *pitch*, *bat*) from BERT and DistilBERT across all layers. For each target token occurrence, we will collect the hidden state vectors at each layer. We will then:

- Visualize structure via dimensionality reduction (e.g., t-SNE, UMAP) to qualitatively inspect clustering by sense.
- Apply unsupervised clustering (e.g., k-means, HDBSCAN) and compare discovered clusters with sense labels derived from contextual heuristics or small annotated subsets.
- Quantify separation with cluster metrics (e.g., silhouette score, ARI/NMI when labels available) as a function of layer depth and model.

EVALUATION PLAN

We will evaluate whether sense-specific groupings emerge without supervision (H1) and where they are strongest across layers (H2). For H1, we expect clear multi-modal structure corresponding to distinct senses. For H2, we expect

*GitHub: https://github.com/lawrencegranda.

Author's address: Lawrence Granda Zarzuela, lg626@cornell.edu, Cornell University. BIONB 3500, USA.

clustering quality to increase from lower layers, peak in middle-to-upper layers, and possibly taper near the output layers.

FEASIBILITY

To evaluate feasibility, we obtained an automated research report from Google Gemini's Research mode (https://gemini.google.com/share/8ef7ab3edcd1). While informative, the report may be biased toward conservative (null) conclusions; therefore, our analysis will emphasize rigorous quantitative tests and clear visualizations.

RELATED WORK

Our study is related to work on clustering with large-language-model embeddings [2], analyses of polysemy in BERT [3], and investigations of layer-wise representational dynamics in transformers [1].

REFERENCES

[1] Nadipalli, S. Layer-wise evolution of representations in fine-tuned transformers: Insights from sparse autoencoders, 2025.

[2] Petukhova, A., Matos-Carvalho, J. P., and Fachada, N. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering 6* (Dec. 2025), 100–108.

[3] Yenicelik, D., Schmidt, F., and Kilcher, Y. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (Online, Nov. 2020), A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad, Eds., Association for Computational Linguistics, pp. 156–162.