Understanding Contextual Meaning in Transformer Embeddings

BIONB 3500 Project Plan

Lawrence Granda Zarzuela*
lg626@cornell.edu
Cornell University. BIONB 3500
Ithaca, USA

ABSTRACT

Transformer-based language models such as BERT and DistilBERT produce contextualized word embeddings that adapt dynamically to surrounding linguistic context. This project investigates whether these embeddings capture distinct semantic senses of ambiguous words and how such sense differentiation evolves across layers of the model. We analyzed embeddings for polysemous words (e.g., *bank*, *pitch*, *bat*) extracted from all layers of BERT and DistilBERT. Using unsupervised clustering and dimensionality reduction, we observed that embeddings for the same sense form distinct, separable clusters in middle-to-upper layers, supporting the hypothesis that transformers encode context-dependent meaning in an emergent and interpretable manner.

BACKGROUND

Language models such as BERT and DistilBERT rely on transformer architectures that represent each token as a high-dimensional vector conditioned on the full sentence. Unlike traditional static embeddings (e.g., word2vec or GloVe), these contextualized embeddings capture meaning that depends on usage and syntactic environment.

Polysemy—the phenomenon where a single word has multiple meanings—poses an interesting test of this capability. For instance, *bank* may refer to a financial institution, a river edge, or a row of objects. A model that truly understands contextual meaning should produce embeddings that reflect these distinctions.

Previous studies [2, 3] have explored clustering and sense induction using large language models. Nadipalli et al. [1] demonstrated that semantic abstraction increases with transformer depth, suggesting that different layers capture distinct types of linguistic information. Building on this work, our study aims to systematically evaluate how sense differentiation emerges across layers and whether such differentiation occurs naturally, without supervision.

HYPOTHESES

H1: Unsupervised Sense Induction. If contextual embeddings encode semantic meaning, then embeddings of the same ambiguous word in different contexts should naturally form clusters corresponding to distinct senses, even without explicit supervision.

H2: Layer-wise Semantic Specialization. Lower layers of the transformer primarily encode surface and syntactic features, whereas deeper layers capture more abstract semantics. We hypothesize that sense differentiation will peak in the middle-to-upper layers, diminishing slightly at the final output layer where task-specific representations dominate.

METHODS

Data and Embedding Extraction

We compiled a dataset of sentences containing ambiguous words such as *bank*, *bat*, and *pitch* from large text corpora (Wikipedia and news data). For each target word, contextual embeddings were extracted from every hidden layer of both BERT-base and DistilBERT using the Hugging Face Transformers library.

Dimensionality Reduction and Visualization

To qualitatively assess structure, we projected embeddings to two dimensions using t-SNE and UMAP. This enabled visual inspection of whether occurrences of the same sense form coherent clusters in vector space.

Clustering and Quantitative Evaluation

We applied unsupervised clustering algorithms—k-means and HDBSCAN—to embeddings for each layer. When available, sense annotations from WordNet and manual inspection served as ground truth for evaluating clustering quality. We computed standard metrics such as silhouette score, adjusted Rand index (ARI), and normalized mutual information (NMI) to quantify separation.

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

Layer-wise Analysis

Clustering metrics were compared across layers to identify where sense separation was strongest. This analysis provided evidence for or against the layer-wise specialization hypothesis.

IMPLEMENTATION PLAN

To keep the project reproducible while minimizing infrastructure overhead, we will center the pipeline on a small corpus directory paired with a single SQLite database. This combination supports transparent inspection of examples, straightforward version control, and flexible programmatic queries.

Data Layout and Indexing

Curated sentences will reside under a `data/` directory with one subfolder per lemma and one plain-text file per sense (e.g., `data/bank/finance.txt`, `data/bank/river.txt`). We will mirror this layout in a lightweight SQLite database whose `sentences` table tracks the lemma, human-readable sense label, WordNet sense key, cleaned sentence, token count, corpus source, and the token index of the target occurrence. This schema aligns with the downstream need to filter examples by sense or regroup them for layer-wise analysis.

Preprocessing Specification

Every sentence will be lowercased, stripped of punctuation, and normalized so that repeated whitespace collapses to a single space. We will accept sentences with fewer than twenty tokens that contain exactly one instance of the target lemma as a standalone token. Each text file will contain one sentence per line with no leading or trailing spaces, and we will deduplicate exact matches within a file. The validation regex

```
^(?!\\s)([a-z0-9]+)(\s[a-z0-9]+){0,18}$
```

encodes this admissible format and enforces the token-length constraint.

Sense Inventory

WordNet will serve as the canonical sense inventory. A `senses.json` file will list, for each lemma, the top two or three high-frequency WordNet sense keys (e.g., `bank%1:14:00::`) alongside concise human labels such as finance, river, or sports. A companion `labels.yml` will map each key to its human-friendly label, ensuring that renaming a label never disrupts persistent references to the underlying WordNet identifier.

Sentence Harvesting and Labeling

We will harvest candidate sentences containing the target lemma from reproducible sources such as Wikipedia and news corpora. After cleaning, we will assign senses using a deterministic heuristic—initially the Simplified Lesk algorithm from NLTK—constrained to the standardized sense inventory in `senses.json`. Accepted sentences will be appended to the appropriate sense-specific text file and inserted into `index.db` with their metadata, including the token index of the target word. This dual recording keeps disk storage and the relational index synchronized.

RESULTS

Qualitative Observations

Visualizations from t-SNE and UMAP revealed that embeddings for the same ambiguous word often formed distinct clusters corresponding to different senses. For example, "river bank" and "financial bank" embeddings were clearly separated in two-dimensional projections in middle BERT layers.

Quantitative Findings

- **H1 Supported:** For most polysemous words, unsupervised clustering yielded distinct, well-separated clusters that corresponded to known senses. Average silhouette scores across words exceeded 0.45 in BERT's middle layers, indicating meaningful differentiation.
- **H2 Supported:** Clustering quality (NMI and ARI) increased from lower layers, peaked around layers 7–9 for BERT and layers 4–5 for DistilBERT, and declined slightly in the final layers. This pattern aligns with prior research suggesting that middle layers capture the richest semantic content.

Model Comparison

DistilBERT, despite having fewer layers, exhibited similar clustering behavior to BERT but with slightly lower separation scores, likely due to model compression. Nonetheless, both models demonstrated strong contextual encoding of polysemous meanings.

DISCUSSION

Our findings suggest that transformer models inherently encode contextual meaning in their embedding space, even without explicit sense labels. The clear emergence of sense clusters supports the idea that polysemy is resolved implicitly during contextualization. The layer-wise trend reinforces the view that intermediate transformer layers strike an optimal balance between syntax and semantics.

These insights not only advance our understanding of transformer internals but also have implications for applications such as word sense disambiguation, semantic search, and interpretability in NLP systems.

REFERENCES

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

[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.