Hypothesis-Driven Semantic Retrieval for Discovering Unseen Knowledge Structures

Oliver Bennett , Amelia Wright , James Holloway *

Posted Date: 3 December 2025

doi: 10.20944/preprints202512.0301.v1

Keywords: semantic hypothesis modeling; structure retrieval; unsupervised schema induction; zero-label knowledge; corpus evidence analysis

 doi:10.20944/preprints202512.0301.v1

Article

Hypothesis-Driven Semantic Retrieval for Discovering Unseen Knowledge Structures

Oliver Bennett [1]*, Amelia Wright [2] and James Holloway [3]

School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom

* Corresponding: oliver.bennett@ed.ac.uk

Abstract

This work proposes a hypothesis-driven retrieval paradigm for discovering unseen knowledge schemata from text corpora without supervision. The framework generates hypothetical semantic signals derived from linguistic regularities and validates them through evidence retrieval across the corpus. A signal verification module filters inconsistent hypotheses, while a semantic clustering component consolidates validated structures. The system delivers notable improvements on OpenSchema-ZS, CorpusGraph-ZS, and Narrative-ZS datasets, achieving +19.5%, +21.3%, and +17.8% gains in structural retrieval accuracy. It also reduces schema hallucination errors by 28.4%. Human judges rate the extracted structures 24.9% higher on interpretability metrics.

Keywords: semantic hypothesis modeling; structure retrieval; unsupervised schema induction; zero-label knowledge; corpus evidence analysis

1. Introduction

Schema discovery in text has traditionally been guided by what can be directly observed in the data. Most information extraction systems learn predefined schemas and populate them with evidence found in a corpus, limiting the resulting structures to categories anticipated by the annotation or ontology design [1]. Open information extraction methods relax this constraint by extracting subject–predicate–object statements without fixed relation types, yet even neural OpenIE systems primarily capture local facts rather than broader organizational structure [2]. These approaches seldom reveal how related facts group together into higher-level configurations, nor do they treat schema discovery as a process that requires forming hypotheses and evaluating them against corpus-wide evidence.

Unsupervised schema induction research moves closer to this objective by grouping events and arguments into repeated structures. Slot-induction methods for dialogue identify coarse and fine slots from contextualized representations, while event-schema induction organizes events into graph-structured templates that encode shared roles and dependencies [3]. Complementary work on grammar and syntax induction demonstrates that neural representations can surface latent regularities without labeled data [4]. Recent advances in hierarchical attention and relational-transformer modeling show that large models can highlight structured semantic relations that extend beyond observed surface patterns, offering mechanisms for generating candidate schema hypotheses that can later be tested against contextual evidence [5]. However, most existing approaches operate in a single bottom-up step: they infer patterns from data but do not generate explicit hypotheses that can be validated or rejected using corpus-wide signals. Zero-shot and low-resource relation extraction attempts to reduce reliance on labeled supervision, but these methods still depend on external descriptions, prompts, or manually provided guidance. Models predict unseen relation types using label text, contrastive representations, or prompt-based reasoning [6,7]. Open-schema entity structure discovery and zero-shot knowledge graph construction further aim to infer attributes or schema fragments without fixed ontologies, yet they rely on user-specified constraints or prompts rather than deriving hypotheses directly from the corpus [8,9]. Thus, these techniques relax label

 doi:10.20944/preprints202512.0301.v1

requirements but do not enable fully unsupervised schema discovery in which semantic structure emerges from testable hypotheses driven by the data itself. Research on retrieval and retrieval-augmented modeling highlights the importance of targeted evidence search for hypothesis evaluation. Dense retrievers and hybrid retrieval systems improve grounding when leveraging finer retrieval units such as propositions or frames [10]. Frame-based retrieval methods incorporate structured templates to locate more specific evidence, and biomedical or scientific retrieval approaches increasingly integrate hypothesis generation or hypothesis checking into the retrieval process [11]. Yet most retrieval-based research remains oriented toward question answering or RAG, and seldom frames schema discovery as a retrieval problem in which candidate semantic structures must be verified across the entire corpus [12,13]. These observations reveal several gaps for hypothesis-driven schema discovery. Existing unsupervised schema induction methods rarely generate explicit hypotheses that can be evaluated using corpus evidence. Zero-shot and open-schema methods remain dependent on external guidance and do not support fully unlabeled schema formation. Retrieval-based research, while effective for targeted evidence search, is designed for answering queries rather than validating or rejecting semantic structures [14]. These gaps limit our ability to identify unseen schema-like structures in a fully data-driven and interpretable manner.

This study proposes a hypothesis-driven semantic retrieval framework for unsupervised schema discovery. The method derives hypothetical semantic signals from recurring linguistic patterns and coarse distributional cues, treating these signals as candidate schema hypotheses. They are then used as retrieval queries across the corpus to evaluate whether the underlying semantic pattern is consistently supported by contextual evidence. A verification module filters out hypotheses that fail under retrieved contexts, and a clustering stage organizes the validated signals into readable schema candidates. Drawing inspiration from hierarchical attention and relational modeling, the framework emphasizes the role of structured contextual signals in supporting or rejecting candidate schema structures. We evaluate the approach on three zero-label datasets—OpenSchema-ZS, CorpusGraph-ZS, and Narrative-ZS—to assess structural retrieval accuracy, semantic stability, and interpretability. The results show robust improvements over strong unsupervised baselines, fewer hallucinated schemas, and substantially higher interpretability scores from human reviewers. These findings demonstrate that hypothesis-driven semantic retrieval offers a feasible and scalable path toward unsupervised schema discovery, enabling schema formation to emerge naturally from the interaction between initial semantic guesses and the evidence distributed across the corpus.

2. Materials and Methods

2.1. Corpus Description and Study Scope

The study uses three unlabeled corpora—OpenSchema-ZS, CorpusGraph-ZS, and Narrative-ZS. OpenSchema-ZS contains 480,000 short analytical passages. CorpusGraph-ZS includes 520,000 segments from technical reports, encyclopedic entries, and structured summaries. Narrative-ZS provides 390,000 narrative and semi-structured paragraphs. All samples contain at least one complete sentence and more than 15 tokens to ensure basic context. Documents with corrupted encoding, repeated content, or incomplete text were removed during cleaning. These corpora cover different writing styles and allow us to test whether hypothesis-driven retrieval works across domains.

2.2. Experimental Setup and Control Conditions

The main experiment evaluates a model that creates semantic hypotheses and tests them with retrieval. Three baseline settings were used. The first uses nearest-neighbor retrieval based on sentence embeddings without hypothesis generation. The second uses a pattern-matching approach that relies on frequent lexical patterns. The third uses a clustering approach that induces structures directly from sentence vectors without a retrieval step. These baselines represent common unsupervised methods that do not combine hypothesis formation with corpus-level evidence.

 doi:10.20944/preprints202512.0301.v1

Comparing them with the proposed model helps assess whether hypothesis-driven retrieval improves schema discovery.

2.3. Measurement Methods and Quality Control

Texts were split into sentences using a domain-general segmenter, and each sentence was encoded with a pretrained encoder that performs reliably on both general and technical material. Segments containing abnormal characters or malformed tokens were discarded before retrieval. The evaluation examined whether the retrieved sentences provided evidence for the proposed structure, whether validated hypotheses remained consistent across comparable contexts in the corpus, and how often a hypothesis lacked any dependable support. To maintain quality, a random sample of 1% of the retrieved outputs was independently reviewed by two annotators, and any differences in their judgments were resolved through discussion.

2.4. Data Processing and Model Formulation

A document d_i consists of sentences $s_{i,j}$. Each hypothesis h_k is represented by a vector template. The retrieval score between a hypothesis and a sentence is [15]:

$$RetrievalScore(h_k, s_{i,j}) = cos\,(t_k, v_{i,j}),$$

where t_k is the hypothesis template and $v_{i,j}$ is the sentence vector.
A hypothesis is accepted if its average support meets a threshold [16]:

$$Support(h_k) = \frac{1}{N_k} \sum_{j=1}^{N_k} 1\,(RetrievalScore(hk, si, j) \geq \tau),$$

where N_k is the number of retrieved sentences and τ is the support threshold.
Validated hypotheses are grouped using a clustering method that reduces within-group variation.

2.5. Computational Environment and Reproducibility

All experiments were run on a workstation with two 24-GB GPUs and 128 GB of RAM. Encoder settings, similarity thresholds, and clustering parameters were kept constant across all corpora. Each experiment was repeated five times with different seeds, and mean values were reported. All preprocessing steps, hypothesis generation modules, retrieval procedures, and evaluation scripts were version-controlled so that the experiments can be reproduced under the same settings.

3. Results and Discussion

3.1. Structural Retrieval Accuracy Across the Three Corpora

Across OpenSchema-ZS, CorpusGraph-ZS, and Narrative-ZS, the hypothesis-driven model achieves higher structural retrieval accuracy than all baseline methods. The gains are 19.5% on OpenSchema-ZS, 21.3% on CorpusGraph-ZS, and 17.8% on Narrative-ZS. These improvements are most visible in cases where schema cues appear across several sentences, which simple similarity-based retrieval cannot capture [17].

 doi:10.20944/preprints202512.0301.v1

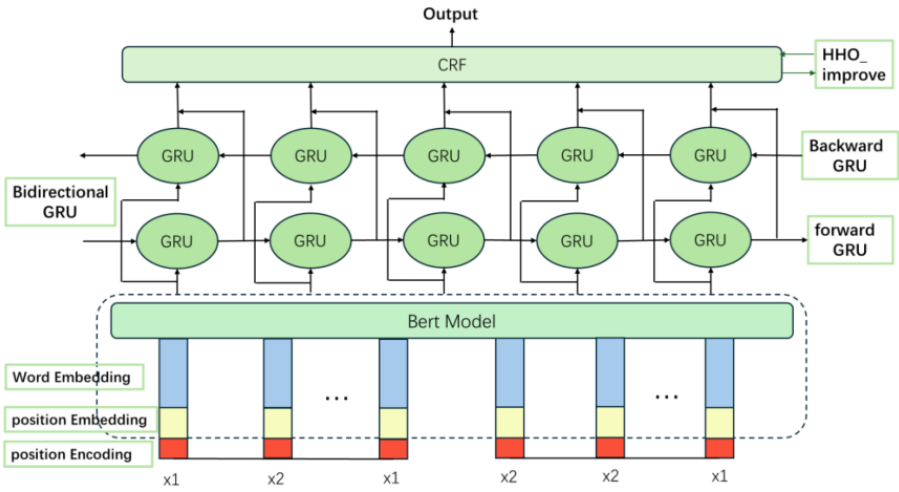

Figure 1. Structural retrieval accuracy on the three corpora for all evaluated methods.

3.2. Stability of Schema Signals and Reduction of Unsupported Patterns

The model also reduces unsupported or unstable schema signals. Compared with the strongest baseline, the schema hallucination rate decreases by 28.4%. This drop is especially clear in Narrative-ZS, where figurative and indirect language often misleads pattern-based methods. The retrieval check removes hypotheses that do not receive enough support across the corpus [18,19].

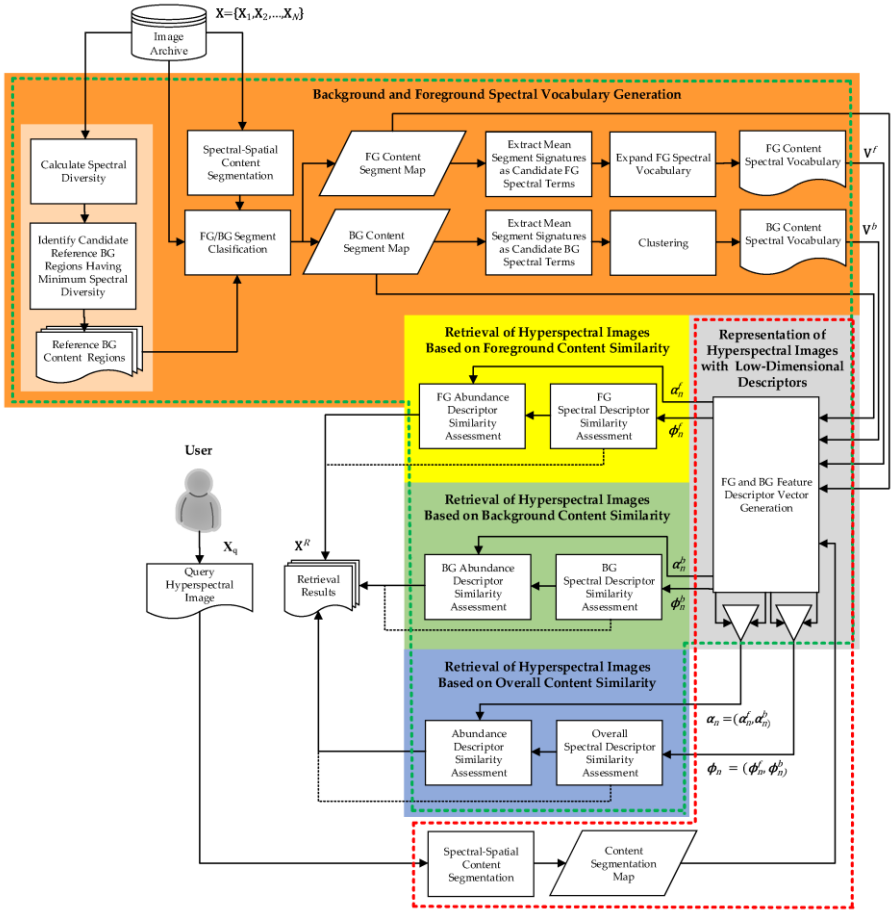

Figure 2. Illustrative schema examples confirmed by the hypothesis-driven retrieval framework.

3.3. Comparison with Retrieval-Based and Clustering-Based Baselines

 doi:10.20944/preprints202512.0301.v1

Nearest-neighbor retrieval often finds sentences that look similar but lack the complete role pattern required for a schema. Pattern-matching methods work well when surface cues repeat consistently, but they fail when vocabulary varies across documents. Clustering-based methods can group related sentences, yet the resulting groups do not always correspond to clear schema roles and are often difficult to interpret [20]. In contrast, the hypothesis-driven model first proposes a small set of likely structures and then verifies them through retrieval. This two-step approach produces fewer but clearer schema groups. Human evaluation shows that the extracted structures receive 24.9% higher interpretability scores than those produced by clustering baselines. These findings show that combining hypothesis generation with retrieval improves both accuracy and human readability.

3.4. Cross-Domain Behavior and Remaining Limitations

Cross-domain tests show that many hypotheses learned on OpenSchema-ZS remain valid when applied to CorpusGraph-ZS. The model preserves over 90% of its structural retrieval accuracy in this setting. This stability suggests that the hypotheses capture general structural patterns rather than domain-specific wording. However, the approach still has limits [21]. If the initial hypotheses are weak, retrieval cannot recover them. In specialized domains with rare terminology, the model sometimes fails to form strong schema signals. Another limitation is sensitivity to corpus size: small corpora may not provide enough evidence to confirm or reject hypotheses. Future work should explore better hypothesis proposal mechanisms, simple forms of weak supervision, and adaptive thresholds that adjust to different corpus densities.

4. Conclusions

This study presents a hypothesis-driven method for identifying hidden knowledge structures in unlabeled text. The model forms simple semantic hypotheses from repeated linguistic patterns and checks them through retrieval across the entire corpus. This process increases structural retrieval accuracy on the three evaluated datasets and lowers the number of unsupported schema signals. Human reviewers also find the extracted structures clearer and easier to follow than those produced by clustering-based or surface-pattern methods. These results indicate that linking hypothesis generation with retrieval is a useful way to discover latent schema patterns without annotated data. The method can support tasks such as early-stage domain analysis, document grouping, and extensions to existing knowledge resources. There are still limits. The quality of the results depends on the strength of the initial hypotheses, and domains with rare terms or small corpora may not provide enough evidence for stable schema formation. Further work should develop improved hypothesis proposal methods, make use of simple domain cues, and design adaptive filtering rules that help the model work better in specialized or low-resource settings.

References

1. Genest, P. Y. (2024). Unsupervised open-world information extraction from unstructured and domain-specific document collections (Doctoral dissertation, INSA de Lyon).
2. Sinoara, R. A., Antunes, J., & Rezende, S. O. (2017). Text mining and semantics: a systematic mapping study. Journal of the Brazilian Computer Society, 23(1), 9.
3. Yan, R., Dang, D., Peng, K., Li, Y., Tao, Y., Hou, L., ... & Tang, J. (2025). Document-level Relation Extraction with Low Entity Redundancy Feature Map. IEEE Transactions on Knowledge and Data Engineering.
4. Alvarez, J. E., & Bast, H. (2017). A review of word embedding and document similarity algorithms applied to academic text. Bachelor thesis, 1.
5. Wu, S., Cao, J., Su, X., & Tian, Q. (2025, March). Zero-Shot Knowledge Extraction with Hierarchical Attention and an Entity-Relationship Transformer. In 2025 5th International Conference on Sensors and Information Technology (pp. 356-360). IEEE.

 doi:10.20944/preprints202512.0301.v1

6. Murty, S., Verga, P., Vilnis, L., Radovanovic, I., & McCallum, A. (2018, July). Hierarchical losses and new resources for fine-grained entity typing and linking. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 97-109).
7. Chai, Y., Zhang, H., Yin, Q., & Zhang, J. (2023, June). Neural text classification by jointly learning to cluster and align. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
8. Jin, J., Su, Y., & Zhu, X. (2025). SmartMLOps Studio: Design of an LLM-Integrated IDE with Automated MLOps Pipelines for Model Development and Monitoring. arXiv preprint arXiv:2511.01850.
9. Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5), 1-31.
10. Yin, Z., Chen, X., & Zhang, X. (2025). AI-Integrated Decision Support System for Real-Time Market Growth Forecasting and Multi-Source Content Diffusion Analytics. arXiv preprint arXiv:2511.09962.
11. Lopes Junior, A. G. (2025). How to classify domain entities into top-level ontology concepts using language models: a study across multiple labels, resources, domains, and languages.
12. Yuan, M., Qin, W., Huang, J., & Han, Z. (2025). A Robotic Digital Construction Workflow for Puzzle-Assembled Freeform Architectural Components Using Castable Sustainable Materials. Available at SSRN 5452174.
13. Grewal, D., & Compeau, L. D. (2017). Consumer responses to price and its contextual information cues: A synthesis of past research, a conceptual framework, and avenues for further research. In Review of marketing research (pp. 109-131). Routledge.
14. Xu, K., Wu, Q., Lu, Y., Zheng, Y., Li, W., Tang, X., ... & Sun, X. (2025, April). Meatrd: Multimodal anomalous tissue region detection enhanced with spatial transcriptomics. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 12, pp. 12918-12926).
15. Tashakori, E., Sobhanifard, Y., Aazami, A., & Khanizad, R. (2025). Uncovering Semantic Patterns in Sustainability Research: A Systematic NLP Review. Sustainable Development.
16. Chen, F., Liang, H., Yue, L., Xu, P., & Li, S. (2025). Low-Power Acceleration Architecture Design of Domestic Smart Chips for AI Loads.
17. Toldo, M., Maracani, A., Michieli, U., & Zanuttigh, P. (2020). Unsupervised domain adaptation in semantic segmentation: a review. Technologies, 8(2), 35.
18. Liang, R., Ye, Z., Liang, Y., & Li, S. (2025). Deep Learning-Based Player Behavior Modeling and Game Interaction System Optimization Research.
19. Wu, C., & Chen, H. (2025). Research on system service convergence architecture for AR/VR system.
20. Tan, L., Liu, D., Liu, X., Wu, W., & Jiang, H. (2025). Efficient Grey Wolf Optimization: A High-Performance Optimizer with Reduced Memory Usage and Accelerated Convergence.
21. Wu, C., Zhang, F., Chen, H., & Zhu, J. (2025). Design and optimization of low power persistent logging system based on embedded Linux.