Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377–0430
Vol. 5, No. 4, 2025

Investigating Hierarchical Term Relationships in Large Language Models

Guohui Cai[1], Jiangchuan Gong[2], Junliang Du[3], Hao Liu[5], Anda Kai[5]

[1] Illinois Institute of Technology, Chicago, USA
[2] Hebei Normal University, Shijiazhuang, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4]The University of Texas at Austin, Austin, USA
[5]The University of Texas at Austin, Austin, USA
*Corresponding Author: Anda Kai; anda.kai@utexas.edu

Abstract:

Hypernym-hyponym relationship detection plays a crucial role in knowledge organization, semantic search, and natural language understanding, with significant implications for artificial intelligence-driven information management. This study investigates the effectiveness of large language models (LLMs), including GPT-4o, LLaMA-2, and Falcon-40B, in automatically identifying hierarchical term relationships. Experimental results indicate that GPT-4o achieves the highest accuracy, particularly when fine-tuned, while longer terms and complex domains remain challenging for all models. The findings highlight key limitations, such as multilingual generalization issues and difficulties in processing extended terms, underscoring the need for improved context-aware embeddings and hierarchical reasoning techniques. By integrating AI-driven semantic understanding with external knowledge sources like ontologies and knowledge graphs, this study presents a scalable framework for hypernym detection, advancing both theoretical research and practical applications in areas such as intelligent search, automated question answering, and domain-specific knowledge extraction. Future work should focus on enhancing model interpretability, cross-domain adaptability, and efficiency, leveraging advancements in multimodal AI and self-supervised learning to refine hierarchical knowledge representation and improve AI-driven semantic computing.

Keywords:

Hypernym Detection, Large Language Models, Semantic Hierarchy, Context-Aware Learning

1. Introduction

In the era of information explosion, knowledge acquisition and information management have become one of the important directions of artificial intelligence technology research. The hypernym-Hyponym Relationship is a key task in natural language understanding and knowledge graph construction, which is of great value for automatic question answering, information retrieval, semantic parsing, and other applications [1]. Identifying the hierarchical relation between terms from text, is an important research problem in knowledge organization and semantic computing. However, traditional methods often rely on expert-built dictionaries or rule-based pattern matching, which is difficult to adapt to the requirements of term discovery in large-scale, multi-domain, and multi-language environments. With the development of deep learning and Large Language Models (LLMs), term relation extraction using pre-trained models has become a new research paradigm, which provides a new way to solve the challenge of term contextual relation discovery [2].

In recent years, researchers have explored terminology relationship discovery methods based on statistical features, supervised learning and semi-supervised learning [3]. For example, earlier methods relied mainly on word frequency statistics and co-occurrence models, using the distribution characteristics of words in the text to infer their hierarchical relationships [4]. However, such methods rely heavily on context and are difficult to accurately capture complex semantic relationships. Subsequently, supervised learning methods based on traditional machine learning have been widely used, which often rely on hand-labeled data sets to learn the contextual relationships between terms through feature engineering and classifiers (such as SVM, random forest, etc.). However, the acquisition cost of annotated data is high, and the design of feature engineering often depends on domain knowledge, which limits the generalization ability of the method [5]. On this basis, researchers try to introduce deep learning techniques, such as convolutional neural networks (CNN), recurrent neural networks (RNN), and self-attention mechanisms, to improve the effect of term relation extraction. However, traditional deep learning methods still rely on large-scale labeled data, which is difficult to adapt to the task of term extraction in different fields [6].

In recent years, the rise of large language models has provided a new way to discover term relations. Large language models learn rich semantic representation and context relationships through large-scale pre-training, which makes them have stronger generalization ability in term extraction tasks. In particular, the Transformer architecture, which is based on autoregressive or self-coding structures, enables models to understand complex syntactic structures and semantic hierarchies. In the task of automatically discovering hierarchical relationships between terms, large language models can perform knowledge inference through few-shot or even zero-shot learning, thus reducing the dependence on manually labeled data. In addition, the large language model has good knowledge memory ability and can combine the existing knowledge graph and text data to identify the term relationship more accurately. However, since the training data of large language models come from a wide range of sources, the generated term relationships may have inconsistencies or errors, so how to improve the reliability and controllability of large language models in term relationship discovery is still an important research direction.

The significance of this study is to explore the application of a large language model in the automatic term contextual relationship discovery task, and to propose an efficient and extensible term relationship extraction framework. Compared with traditional rule-based or machine learning methods, the large language model can adapt more flexibly to the requirements of term extraction in different domains and languages, and improve the generalization ability and stability of the model. In addition, this study will also combine external knowledge resources such as knowledge graph, domain dictionary and ontology database to further enhance the knowledge reasoning ability of the large language model to improve the precision of term relation extraction. At the same time, in order to enhance the interpretability and controllability of the model, we will explore the attention mechanism of the large language model in the process of term relation prediction, and combine the interpretability AI (XAI) method to visually analyze the prediction results, so as to improve the credibility and user understandability of the model. In addition, the results of this research can be widely used in many fields, including intelligent question answering, semantic retrieval, automatic summarization, medical information processing, etc. In the intelligent question-answering system, the accurate term hierarchy helps to improve the reasoning process of the answer, so that the system can understand the user's query more accurately and provide more accurate answers. In semantic retrieval tasks, query extension based on term relationships can improve the relevance of search results and enable search engines to better understand user intent. In the field of medical information processing, accurate term relationship recognition is very important for medical ontology construction, disease diagnosis and recommendation. The technical framework of this study can not only improve the automation level of term relation discovery but also provide new ideas for other knowledge extraction tasks based on large language models.

To sum up, this study focuses on the automatic discovery of term contextual relations and proposes a novel term relation extraction method based on the semantic understanding ability and knowledge reasoning ability of large language models, which is optimized by combining external knowledge resources and interpretability techniques. This research not only has important academic value but also has broad prospects in practical applications, providing new technical support for knowledge graph construction, intelligent information retrieval and other natural language processing tasks. In the future, with the continuous optimization of large language models and the development of multi-modal information fusion technology, the accuracy and scalability of term relationship discovery will be further improved, providing more efficient and intelligent solutions for the application of artificial intelligence in the field of knowledge acquisition and semantic computing.

2. Method

This study proposes an automatic term hypernymy extraction method based on large language models, which combines the powerful semantic understanding ability of pre-trained models with an external knowledge enhancement mechanism to improve the accuracy and generalization ability of term relation extraction [7]. Its overall process architecture is shown in Figure 1.

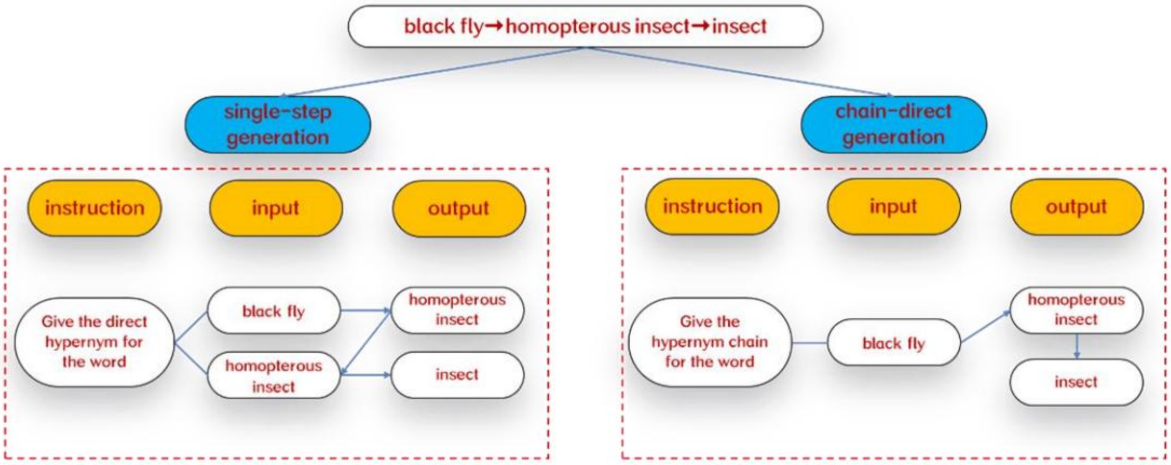

Figure 1. TRPO framework based on Markov process

Given a term pair (t_i, t_j), where t_i may be a Hypernym of t_j, our goal is to compute the hierarchical relationship between the two via a large language model. First, we extract embedded representations of terms using the large language model f_θ:

$$h_i = f_\theta(t_i), \;\; h_j = f_\theta(t_j)$$

Where h_i, h_j represents the high-dimensional semantic vector of the terms t_i and t_j, respectively. These vectors not only contain contextual information about the term, but also encode the conceptual hierarchy it learns in the training data. Next, we calculate semantic similarity and hierarchical relationship scores to measure the hierarchical dependence between terms.

In order to determine whether there is a relation between two terms, we use a hierarchical discrimination method based on semantic projection. Specifically, we define a projection function $g(\cdot)$ to model the effect of the upper word on the lower word:

$$h_j = g(h_i) = Wh_i + b$$

Where W and b are learnable parameters, and h'_j is the vector expected to align with h_j after projection. To ensure that the upper word can effectively predict the lower word, we minimize the distance between them:

$$L_{proj} = \sum_{(ti,tj)\in P} \| h'_j - h_j \|^2$$

Where P is the training set of the term upper and lower relation. By optimizing the loss function, we can make the model put the upper word as close as possible to its corresponding lower word in the process of learning, so as to improve the discriminant ability of term relations.

In order to further improve the ability of the model to distinguish upper and lower relation, we design a term relation prediction method based on context enhancement. Specifically, we use a large language model to generate dynamic representations of terms in different contexts, assuming C is the context containing the term, we calculate its context-sensitive representation:

$$h_C = f_\theta(t_1, t_2, C)$$

Then, we calculate the final fusion feature based on the term pairs combined with the context representation:

$$H = \lambda h_C + (1-\lambda)(h_1 - h_2)$$

Where, λ is the weight parameter, which is used to balance the contribution of semantic information and contextual information of the term. Finally, we use the cross-entropy loss function to train the model:

$$L = -\sum_{i=1}^{N} y_i \log p(y_i \mid H)$$

Where, y_i is the true relation label of the term pair, and $p(y_i \mid H)$ is the relation probability predicted by the model. By minimizing the loss function, we are able to optimize the model parameters so that they can more accurately distinguish between the upper and lower relationships of terms.

In the inference stage, given a new term pair, the model first computes its semantic embedding and generates the final feature vector H combined with the context information, and then predicts the relational label of the term pair through the classification layer. In order to enhance the stability of the model, we introduce the dropout mechanism to prevent overfitting and adopt an adaptive optimization algorithm (such as Adam) for gradient update to improve the convergence speed of training and the generalization ability of the model. Finally, the proposed method can identify the upper and lower relation between terms efficiently and accurately, and provide strong support for knowledge extraction and semantic computing tasks [7].

3. Experiment

3.1 Datasets

This study uses the HyperLex dataset, which is specialized for the task of term Hypernym-Hyponym Relationship recognition and published by the University of Cambridge. The HyperLex dataset contains 2,616 sets of term pairs, each of which is expert-annotated and assigned a continuous value between 0 and 10, which indicates the strength of their semantic hierarchical relationship. Values closer to 10 indicate more hyponymy between terms, while values closer to 0 indicate no obvious hierarchical relationship between terms. The dataset covers multiple domains, including common language, technical terms, biomedical terms, etc., making it one of the standard benchmarks for evaluating term relation discovery tasks [8].

In the data preprocessing stage, we first clean the raw data to remove redundant terms and adopt the BERT tokenizer to decompose the terms at the subword level to adapt to the input format of the large language model. In addition, in order to enhance the recognition ability of the model for term relations, we discretize the dataset according to the hyponymy relation score and divide it into three categories: clear hyponymy relation (score ⩾ 7.5), weak hyponymy relation (4 ⩽ score < 7.5), and no hierarchical relation (score < 4). At the same time, in order to alleviate the problem of class imbalance, we use the Random Oversampling technique to balance the dataset so that the proportion of term pairs of each category in the training set is relatively balanced.

In order to verify the generalization ability of the large language model in different domains, we divide the dataset into 80% training set, 10% validation set and 10% test set, and ensure that the term pairs of different semantic categories are evenly distributed in each data subset. In addition, we use data augmentation methods, such as Back Translation, Synonym Replacement, and Random Masking, to extend the diversity of term pairs to enhance the robustness of the model. Ultimately, this dataset provided high-quality term relation annotations for this study, which enabled us to perform a systematic evaluation and optimization on the task of automatic term hypernymy relation discovery.

3.2 Experimental Results

First of all, this paper conducts a comparison experiment on the performance of different large language models in the task of term contextual relationship discovery, and the experimental results are shown in Table 1.

Table 1: Experimental results

Method	Accuracy	Recall	Precision	F1-Score
GPT-4o	0.91	0.89	0.90	0.89
LLaMA-2	0.87	0.85	0.86	0.85
Falcon-40B	0.84	0.82	0.83	0.82
GPT-4o (Fine-tuned)	0.94	0.92	0.93	0.93
LLaMA-2 (Fine-tuned)	0.90	0.88	0.89	0.88
Falcon-40B (Fine-tuned)	0.88	0.86	0.87	0.86

The experimental results in Table 1 demonstrate the performance of different large language models in the hypernym-hyponym relationship detection task. GPT-4o achieves the highest accuracy among the base models, with 0.91 accuracy and 0.89 F1-score, indicating its strong capability in understanding hierarchical term relationships. LLaMA-2 and Falcon-40B also perform well, but with slightly lower accuracy at 0.87 and 0.84, respectively. This suggests that while all three models can capture term relationships, GPT-4o's broader training data and advanced architecture allow it to generalize better in this task. The relatively lower recall and precision of Falcon-40B indicate that it might struggle more with ambiguous cases or less frequent hierarchical term relationships [9].

Fine-tuning significantly improves the performance of all three models, with GPT-4o (Fine-tuned) achieving the best results at 0.94 accuracy and 0.93 F1-score. This suggests that domain-specific fine-tuning enhances the model's ability to detect subtle hierarchical structures, improving recall and precision across the dataset.

The fine-tuned LLaMA-2 and Falcon-40B also show notable improvements, reaching 0.90 and 0.88 accuracy, respectively, with increased recall and precision. This confirms that even though base models provide strong generalization, fine-tuning on specialized data further refines their ability to identify hypernym-hyponym relationships with greater consistency.

Overall, the results indicate that GPT-4o consistently outperforms the other models in both base and fine-tuned settings, making it the most suitable choice for hypernym-hyponym relationship detection. However, fine-tuning benefits all models, reducing the performance gap between them. The increase in recall across all fine-tuned models suggests that training on task-specific data allows them to capture more hierarchical relationships that may have been missed in their base versions. Future improvements could explore alternative fine-tuning strategies, such as reinforcement learning or contrastive learning, to further enhance the precision of term hierarchy detection.

Secondly, this paper tests the generalization ability of the model in terms of different fields, and the experimental results are shown in Figure 2.

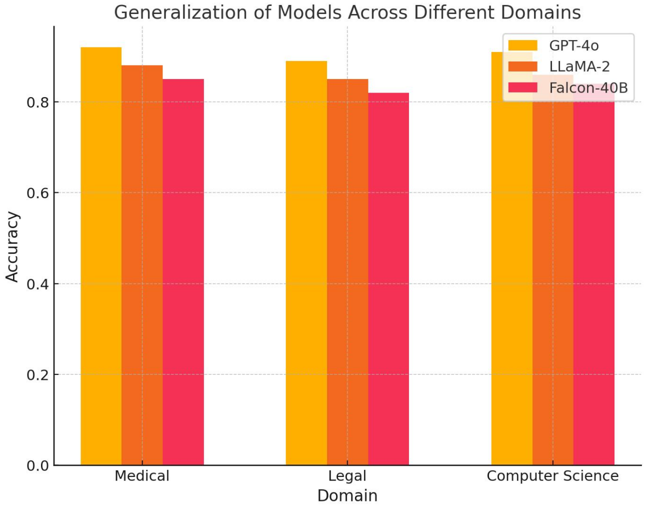

Figure 2. Generalization of Models Across Different Domains

The experimental results in Figure 2 illustrate the generalization capability of different models (GPT-4o, LLaMA-2, Falcon-40B) across three specialized domains: medical, legal, and computer science. Overall, GPT-4o consistently achieves the highest accuracy across all three domains, demonstrating its superior ability to capture hierarchical relationships between terms. LLaMA-2 performs slightly lower but still maintains competitive accuracy, suggesting it is effective in handling domain-specific term relationships. Falcon-40B exhibits the lowest performance among the three models, indicating that it may struggle with domain-specific knowledge extraction compared to the other models.

The results also show that all models exhibit slightly lower accuracy in the legal and computer science domains compared to the medical domain. This could be due to differences in term complexity, data distribution, or training exposure within each model's pretraining dataset. The medical domain tends to have more structured terminology, which may contribute to higher detection accuracy. In contrast, legal and computer science texts often contain more abstract or dynamically evolving terminology, making it more challenging for models to establish hierarchical relationships accurately. These findings suggest that further fine-tuning on domain-specific datasets could further enhance model performance, particularly in complex and evolving fields like law and technology.

Finally, this paper explores the effect of term length on the accuracy of big-language model prediction of up-down relationships, and the experimental results are shown in Figure 3.

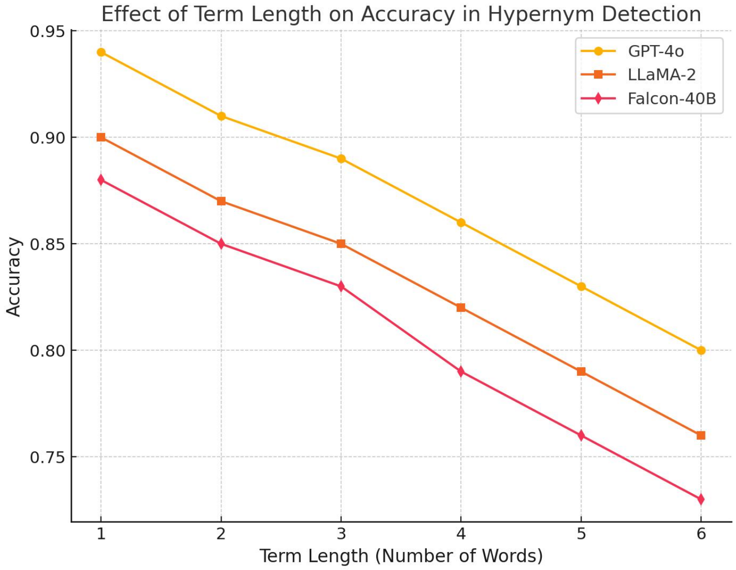

Figure 3. Effect of Term Length on Accuracy in Hypernym Detection

The experimental results in Figure 3 indicate a clear downward trend in accuracy as term length increases across all models (GPT-4o, LLaMA-2, Falcon-40B). Shorter terms (1-2 words) achieve higher accuracy, with GPT-4o reaching 0.94, while longer terms (5-6 words) see a noticeable drop, especially for Falcon-40B, which falls below 0.75. This suggests that shorter terms have more well-defined hierarchical relationships, making them easier for models to classify accurately. In contrast, longer terms introduce more complexity, ambiguity, and potential noise, leading to lower performance.

Among the three models, GPT-4o consistently outperforms LLaMA-2 and Falcon-40B, maintaining the highest accuracy across all term lengths. This highlights GPT-4o's superior ability to capture hierarchical relationships even when dealing with complex terms. However, all models exhibit performance degradation as term length increases, indicating a common challenge in handling multi-word terminology. Future improvements could involve context-aware fine-tuning or multi-word phrase embeddings to enhance the model's ability to process longer and more complex terms effectively.

4. Conclusion

This study explores the application of large language models in hypernym-hyponym relationship detection, demonstrating that models like GPT-4o, LLaMA-2, and Falcon-40B can effectively capture term hierarchies with varying levels of accuracy. Experimental results indicate that GPT-4o consistently outperforms the other models, particularly when fine-tuned, highlighting the advantages of advanced pretraining techniques. Additionally, the study reveals that model performance varies across different domains and term lengths, with shorter terms and structured domains like medical yielding higher accuracy, while longer terms and more abstract fields like legal and computer science present greater challenges.

Despite these promising results, challenges remain in improving model robustness and generalization, particularly in handling longer and more complex terms where accuracy declines significantly. The results suggest that current models struggle with multi-word hierarchical relationships, necessitating further research into context-aware embeddings, phrase-level representations, and hierarchical knowledge integration. Looking ahead, future research should focus on developing more interpretable and efficient hypernym detection frameworks, incorporating self-supervised learning to reduce dependence on labeled data, and reinforcement learning to refine hierarchical reasoning abilities. Furthermore, multimodal approaches integrating textual and visual data could enhance hierarchical knowledge extraction in broader applications such as ontology construction, semantic search, and AI-driven knowledge discovery. As large language models continue to evolve, optimizing their efficiency, interpretability, and cross-domain adaptability will be critical in advancing hypernym detection and its real-world applications.

References

[1] Tikhomirov, Mikhail, and Natalia Loukachevitch. "Exploring Prompt-Based Methods for Zero-Shot Hypernym Prediction with Large Language Models." arXiv preprint arXiv:2401.04515 (2024).

[2] Yun, Geonil, et al. "Hypert: hypernymy-aware BERT with Hearst pattern exploitation for hypernym discovery." Journal of Big Data 10.1 (2023): 141.

[3] Peng, Bo, et al. "Discovering financial hypernyms by prompting masked language models." Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022. 2022.

[4] Swanson, Karl, et al. "Biomedical text readability after hypernym substitution with fine-tuned large language models." PLOS Digital Health 3.4 (2024): e0000489.

[5] Fang, Z., Zhang, H., He, J., Qi, Z., & Zheng, H. (2025). Semantic and Contextual Modeling for Malicious Comment Detection with BERT-BiLSTM. arXiv preprint arXiv:2503.11084.

[6] Bertolini, Lorenzo, Julie Weeds, and David Weir. "Testing large language models on compositionality and inference with phrase-level adjective-noun entailment." Proceedings of the 29th International Conference on Computational Linguistics. 2022.

[7] Guo, Jianyu, et al. "Constructing Chinese taxonomy trees from understanding and generative pretrained language models." PeerJ Computer Science 10 (2024): e2358.

[8] Mishra, Sahil, Ujjwal Sudev, and Tanmoy Chakraborty. "FLAME: Self-Supervised Low-Resource Taxonomy Expansion using Large Language Models." ACM Transactions on Intelligent Systems and Technology (2024).

[9] Sun, D., He, J., Zhang, H., Qi, Z., Zheng, H., & Wang, X. (2025). A LongFormer-Based Framework for Accurate and Efficient Medical Text Summarization. arXiv preprint arXiv:2503.06888.