

Evolutionary Trends and Application Research of Semantic Representation in Natural Language Processing

Shaoguai Liu^{1*}

¹ School of Electronic Information, Wuhan University, Wuhan, China

Email Address

liushaoguai@foxmail.com (Shaoguai Liu)

*Correspondence: liushaoguai@foxmail.com

Received: 2 October 2025; **Accepted:** 14 November 2025; **Published:** 25 November 2025

Abstract:

The evolution of semantic representation in natural language processing has undergone a paradigm shift from static word embeddings to dynamic contextual encoding and then to generative language models. This article systematically analyzes the internal logic and technical turning points of this evolutionary path: static distributed representations laid the foundation for semantic vectorization, but were limited by the discreteness of word meaning and context independence; Transformer-based dynamic encoding models achieved deep contextual awareness through the pre-training and fine-tuning paradigm, but lacked generative capabilities; Large language models break through the boundaries of understanding and generation with a self-regressive generation architecture, demonstrating emergent abilities such as in-context learning and chain-of-thought reasoning, while also raising new challenges such as interpretability and knowledge credibility. The research points out that the future development of semantic representation will rely on innovative paths such as neuro-symbolic fusion and multi-modal unified representation, ensuring trustworthiness and controllability while improving representation capabilities.

Keywords:

Semantic Representation, Natural Language Processing, Large Language Model, Context Awareness, Multimodal Learning

1. Introduction

The core challenge in natural language understanding lies in the abstract and dynamic nature of semantics. Semantics not only reside in the vocabulary itself but are also deeply embedded in complex contextual relationships, domain knowledge, and conversational settings [1,2]. This inherent ambiguity, compositionality, and context-dependence make the task of transforming human language into a machine-computable form extremely challenging. Semantic representation learning, as the cornerstone of addressing this challenge, aims fundamentally to construct mathematical models that can capture and quantify the meaning of language, thereby providing a computable foundation for advanced natural language processing tasks [3].

In the history of technological evolution, semantic representation has primarily undergone three paradigm shifts. Early symbolic-based discrete representations (such as One-hot encoding and Bag-of-Words models), while intuitive and simple, suffered from inherent data sparsity and the curse of dimensionality, making it difficult to effectively capture semantic relationships. Connectionism gave rise to distributed representations (represented by Word2Vec and GloVe), which characterize words in continuous space using low-dimensional dense vectors [4]. Based on the “distributional hypothesis,” they utilized co-occurrence statistics to achieve the numerical representation of semantics, marking a significant breakthrough in semantic representation. However, their static nature could not solve the problem of polysemy. The current mainstream paradigm is contextualized representation, exemplified by ELMo, BERT, and large language models. The core idea is that the representation of a word is dynamically generated by its context, realizing a qualitative leap from “word type” representation to “word token” representation [5].

The evaluation system for semantic representation models has also continuously evolved. Early research focused on intrinsic evaluation, such as directly assessing the quality of vector spaces through word similarity calculations and semantic analogy tasks. With the establishment of the pre-training and fine-tuning paradigm, extrinsic evaluation has become mainstream, indirectly measuring the effectiveness of semantic representations through the performance of downstream tasks (such as question answering, sentiment analysis, and natural language inference). Entering the era of large language models, the evaluation dimensions have been further expanded to include the model's zero-shot/few-shot generalization ability, complex reasoning ability, and the ability to understand and follow instructions [6]. (Figure 1)

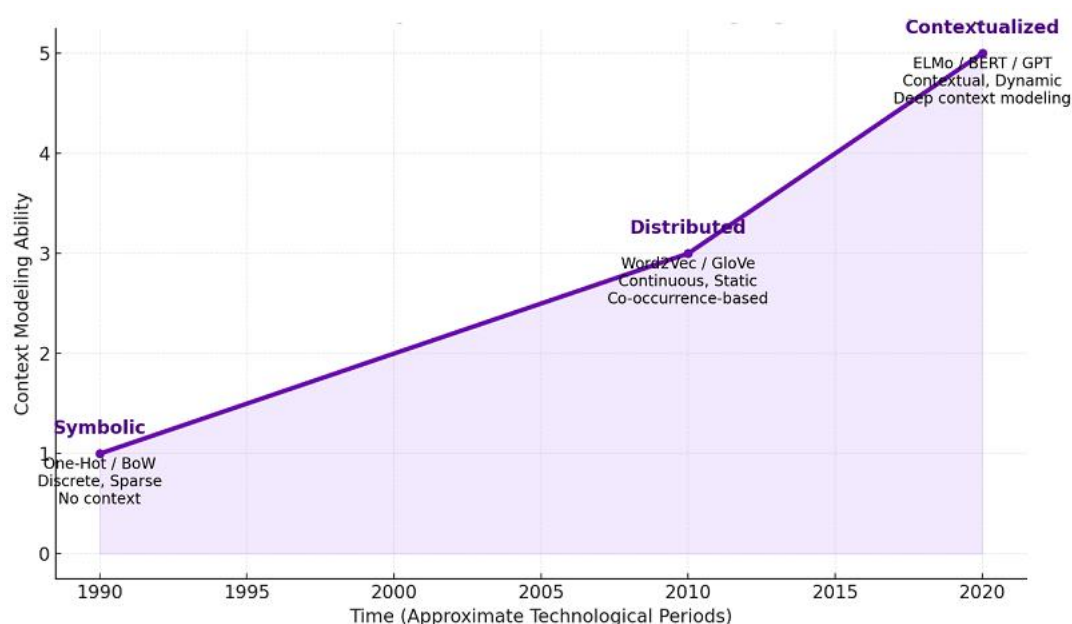


Figure 1. Evolution of Semantic Representation in Natural Language Understanding.

This article aims to systematically review the evolution path of semantic representation, with “context modeling ability” as the main thread. The discussion will revolve around the lineage from static word embeddings without context modeling, to early neural networks with shallow/static context modeling, to Transformer encoders with deep/dynamic context awareness, and finally to generative context modeling with large language models. By dissecting the technological innovations made in each stage to overcome the limitations of previous generations,

this article seeks to reveal the intrinsic development logic of semantic representation research and propose forward-looking reflections on its future trends.

2. From Static to Dynamic: A Paradigm Shift in Context-Aware Representation

Research on semantic representation in the field of natural language processing has consistently revolved around the central question of how to more accurately capture and quantify the meaning of language. The introduction of static word embedding models marked a significant transition from symbolism to connectionism in semantic representation. However, inherent theoretical limitations gave rise to more powerful dynamic context-aware models, ultimately triggering a profound paradigm shift.

2.1. Foundations and Limitations of Static Word Embeddings

The distributional hypothesis provides the theoretical basis for static word embeddings, with its core idea being that the semantics of a word can be defined by its context distribution. The Word2Vec series of models employs shallow neural network architectures to learn word vector representations based on prediction tasks within local context windows. The Skip-gram model optimizes the objective of predicting context words given a center word, while the CBOW model performs the opposite prediction task. Both models optimize training efficiency through techniques such as negative sampling or hierarchical softmax, thereby generating high-quality semantic representations in a low-dimensional continuous space. (Figure 2)

The GloVe model takes a global matrix factorization perspective, combining global word-word co-occurrence statistics with local context window prediction. This model explicitly captures semantic relationships between words by modeling the ratios of word co-occurrence probabilities. Experiments have shown that GloVe performs well in systematic semantic relationship evaluations such as word analogy tasks, proving that it can effectively encode the linear semantic regularities of language.

The FastText model further advanced the development of static word embeddings by introducing subword (n-gram) information into representation learning. By representing words as a collection of character n-grams, this model can generate reasonable vector representations for out-of-vocabulary words and better handle morphological variations in morphologically rich languages.

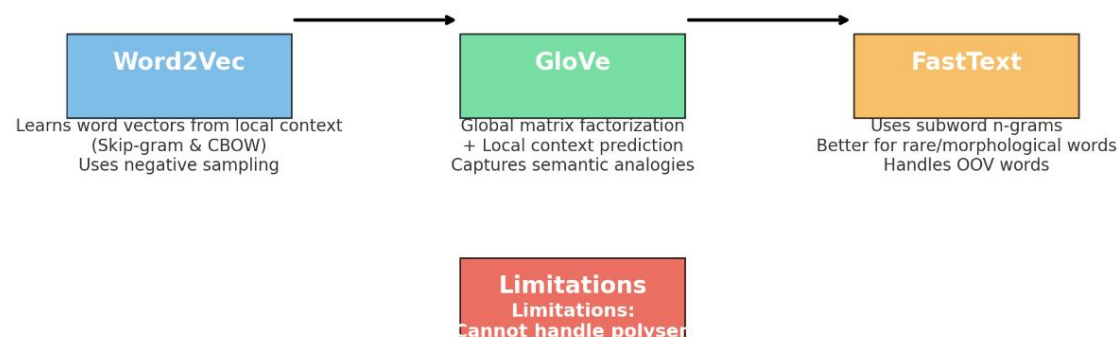


Figure 2. Foundations and Limitations of Static Word Embeddings.

However, static word embeddings have fundamental theoretical limitations. The most significant problem is the inability to resolve polysemy. Since each word form corresponds to only a single vector representation, the model cannot distinguish

between different meanings of a word based on the specific context. Static representations also lack compositionality, making it difficult to effectively capture the overall semantics of phrases and sentences. Simple vector addition or concatenation cannot accurately express complex semantic combination rules. Furthermore, these models utilize contextual information in a relatively shallow manner, relying only on limited local windows, and are unable to model long-range semantic dependencies.

2.2. The Paradigm Shift: From Word Type Vectors to Word Instance Representations

To overcome the limitations of static representations, the research paradigm began to shift from “word type” representations to “word instance” representations. The core of this shift lies in recognizing that the meaning of a word should be a function of its complete context, rather than a pre-defined, fixed mapping in a vocabulary.

ELMo (Embeddings from Language Models) pioneered this shift by employing a deep, bidirectional LSTM architecture to generate context-dependent word representations. The model captures context information from left to right and right to left through the stacking of forward and backward language models, respectively. ELMo's innovation lies in the linear combination of LSTM states from different layers, where lower layers capture syntactic features and higher layers capture semantic features, thereby generating deeply context-aware word representations. This representation method enables the same word to obtain different vector representations in different contexts, initially addressing the polysemy problem. (Figure 3)

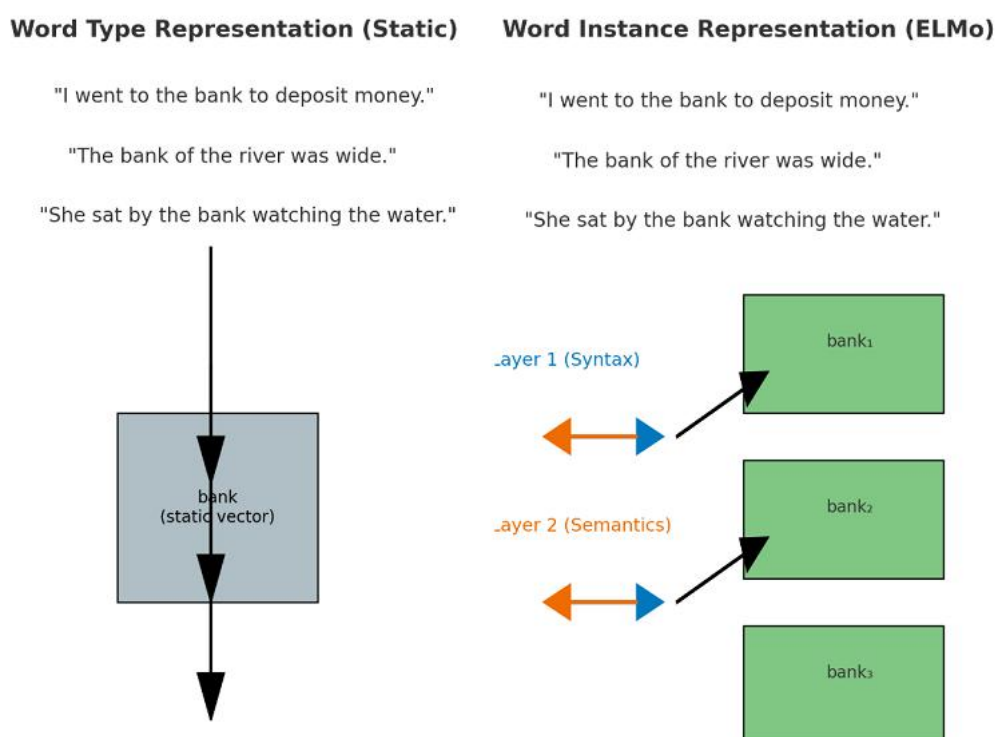


Figure 3. Paradigm Shift: From Word Type to Word Instance Representations.

2.3. Transformer and BERT: The Pinnacle of Dynamic Representation and the Establishment of the Pre-training Paradigm

The introduction of the Transformer architecture completely changed the technological path of sequence modeling. Its core self-attention mechanism allows the model to directly calculate the strength of association between any two positions in the sequence, thereby achieving effective modeling of long-range dependencies. Self-attention uses a query-key-value triplet calculation method, enabling each position to directly attend to all positions in the input sequence. This global perception ability is significantly superior to architectures based on recurrence or convolution.

BERT (Bidirectional Encoder Representations from Transformers) achieved groundbreaking progress based on the Transformer encoder architecture through two key pre-training tasks. The masked language model task randomly masks some words in the input sequence, requiring the model to predict the masked original words based on the bidirectional context. This design enables BERT to utilize context information from both the left and right sides simultaneously, achieving true deep bidirectional encoding. The next sentence prediction task, on the other hand, explicitly models the relationships between sentences by judging whether two sentences are consecutive in the original text, enhancing the model's understanding of discourse-level semantics.

BERT's success established the new NLP paradigm of “pre-training-fine-tuning.” Pre-training on large-scale, unlabeled corpora to learn general language representations, and then performing lightweight fine-tuning on labeled data for specific tasks, significantly reduces the dependence on task-specific architecture design and large amounts of labeled data. From the basic 12-layer Transformer version to the large 24-layer architecture, BERT demonstrated a significant improvement in representation learning ability by increasing model depth and the number of parameters. (Figure 4)

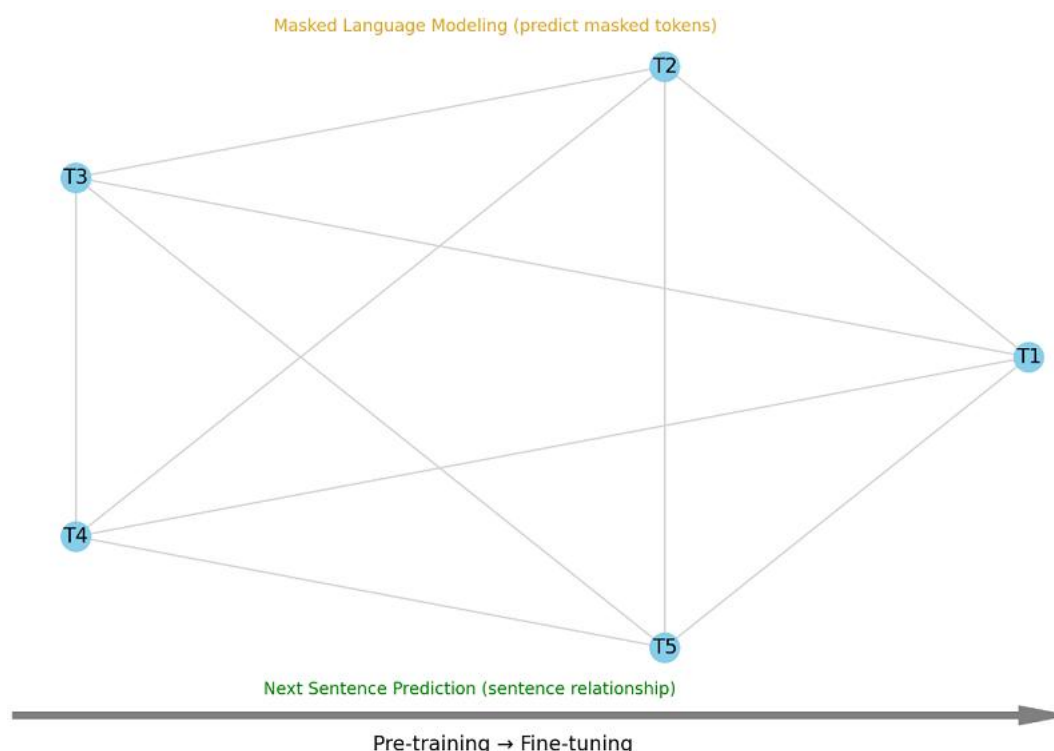


Figure 4. Transformer and BERT: Dynamic Representation & Pre-training Paradigm.

2.4. Technical Comparison and Application Reflections

Static word embeddings and dynamic contextual encoding exhibit systematic differences in technical characteristics. At the architectural level, static models are based on shallow neural networks or matrix factorization, while dynamic models employ deep Transformer encoders. Regarding context utilization, static models rely on limited local windows, whereas dynamic models can process complete input sequences. In terms of representation properties, static models generate context-independent word-type representations, while dynamic models produce word-token representations highly dependent on context.

In terms of application effectiveness, dynamic contextual encoding demonstrates a significant advantage in tasks requiring deep semantic understanding. In question answering systems, BERT can accurately locate answer ranges based on the context of the question. In natural language inference tasks, the model can accurately determine the logical relationships between sentences. In named entity recognition, dynamic representations can effectively resolve the ambiguity of entity boundaries. These advances all benefit from the model's deep understanding of context.

However, dynamic encoders also have clear limitations. Computational complexity increases quadratically with sequence length, which limits the model's ability to process long documents. The encoder architecture lacks generative capabilities, making it unsuitable for direct application in text generation tasks. Furthermore, these models are highly dependent on large-scale, high-quality pre-training data, and the fine-tuning process still requires a considerable amount of labeled data.

Overall, the shift from static word embeddings to dynamic contextual encoding represents a profound paradigm revolution in the field of semantic representation research. This shift not only addressed the core limitations of static representations but, more importantly, established the pre-training and fine-tuning paradigm, laying the theoretical foundation and technical path for the development of subsequent large language models. This evolution embodies the inevitable development trend of semantic representation research from shallow statistics to deep understanding and from isolated representation to context-awareness.

3. The Era of Large Language Models: Generative Architectures and Implicit Semantic Understanding

Following breakthroughs in dynamic context encoding, the field of Natural Language Processing (NLP) has entered a more profound paradigm shift. This shift is not only reflected in the transformation of model architectures from encoders to generative models, but also signifies the expansion of semantic representation from explicit vector encoding to implicit model capabilities. Large language models, through autoregressive generation tasks and massive parameter scales, demonstrate an unprecedented mastery of language semantics, redefining the boundaries of natural language understanding [7,8].

3.1. Architectural Shift: From Encoders to Generative Autoregressive Models

The establishment of the generative pre-trained Transformer architecture represents a distinct technological path from encoder models such as BERT. The GPT series of models strictly employs the decoder component of the Transformer, utilizing a masked self-attention mechanism to ensure that each position can only focus on the left-side contextual information. Although this unidirectional architecture appears limited in the comprehensiveness of contextual awareness, it forces the model to

develop strong language modeling and prediction capabilities through strict autoregressive generation tasks.

The growth of model scale plays a crucial role in this evolution. From 117 million parameters in GPT-1 to 175 billion parameters in GPT-3, quantitative changes have led to qualitative leaps. With the exponential increase in parameter scale, models exhibit emergent abilities such as in-context learning and chain-of-thought reasoning, which were not observed in smaller models. This phenomenon indicates that when the model scale exceeds a certain critical point, large language models are not just performing pattern matching, but are developing some form of semantic internal representation and reasoning mechanism.

The comparison with encoder architectures reveals two different design philosophies. Encoder models like BERT optimize the representation quality of words through bidirectional context encoding, focusing on understanding tasks; while generative models, through the next word prediction task, must simultaneously possess language understanding, knowledge storage, and content generation capabilities. This inherent difference in training objectives enables generative models to maintain strong understanding capabilities while gaining the creativity and flexibility that encoder models lack.

3.2. Deepening the Connotation of Semantic Representation: From Vectors to Capabilities

In the era of large language models, the traditional concept of semantic representation has undergone a fundamental reconstruction. Semantics are no longer merely embodied in a static vector or context-dependent embedding. Instead, they are distributed and stored within the entire model's parameter space, and dynamically manifested as various complex capabilities through the forward computation process.

In-context learning has become one of the most remarkable capabilities of large language models. By using only a few examples in the prompt, the model can understand the task requirements and generate expected responses. This capability indicates that the model has formed a highly flexible task representation mechanism internally. Unlike traditional supervised learning, which requires massive amounts of labeled data and parameter updates, in-context learning demonstrates the model's ability to quickly adapt to new tasks during the reasoning process.

Chain-of-thought prompting further reveals the potential of large language models to handle complex semantic relationships. By requiring the model to generate intermediate steps of reasoning, researchers have discovered that the model can perform complex tasks such as multi-step logical reasoning, mathematical calculations, and common-sense reasoning. This capability suggests that large language models may have mastered symbolic manipulation and logical deduction skills to some extent, although their implementation differs fundamentally from traditional symbolic artificial intelligence.

Instruction tuning and reinforcement learning from human feedback mark important progress in aligning semantic representation with human values. By fine-tuning pre-trained models for instruction-following capabilities and using reinforcement learning algorithms to optimize generated content based on human preferences, large language models can better understand user intentions and generate more helpful, harmless, and honest responses. This process essentially involves adjusting the model's internal semantic space to better align with human communication habits and value judgments.

3.3. Disruption and Challenges of Application Paradigms

The emergence of large language models (LLMs) has fundamentally changed the application paradigm of natural language processing. The traditional approach of “designing a dedicated model for each task” is being replaced by a new paradigm of “a single general-purpose model solving multiple tasks.” Developers no longer need to collect large amounts of labeled data and train dedicated models for each specific task. Instead, they can guide LLMs to complete various tasks by designing appropriate prompts.

In open-domain dialogue systems, LLMs demonstrate the ability to maintain coherent and in-depth conversations, understand contextual references, manage dialogue states, and generate natural and fluent responses. In code generation tasks, models can not only generate functionally correct code based on natural language descriptions but also perform code debugging, explanation, and translation between different programming languages. In the field of knowledge reasoning, LLMs can integrate multi-source information, perform complex logical reasoning and common-sense judgments, and even approach human levels in certain tasks.

However, these breakthrough capabilities are accompanied by serious challenges. The “knowledge hallucination” phenomenon has become a major threat to the reliability of LLMs. Models may generate statements that seem reasonable but are actually incorrect, and often present them with high confidence. This hallucination stems from the fact that models essentially generate text based on statistical regularities, rather than accessing external knowledge bases for fact-checking.

The problem of factual inconsistency is particularly prominent in long text generation. Models may contradict themselves in the generated text, or give inconsistent answers to the same question at different times. This phenomenon reflects the inherent limitations of the autoregressive generation mechanism, where the model only considers the left-side context when generating each word and cannot perform global planning.

The challenge of model interpretability intensifies as the model size increases. The decision-making process of LLMs is like a black box, making it difficult for researchers to trace the reasons for the generation of a specific output or to accurately understand how the model internally organizes and uses knowledge. This lack of interpretability poses safety and ethical hazards to the deployment of critical applications.

The development of LLMs marks a new stage in semantic representation research. Semantics are no longer simplified as geometric relationships in vector space but are embodied in the understanding and reasoning abilities exhibited by models in complex tasks. This shift requires us to rethink the evaluation criteria and methodologies of semantic representation and opens up new possibilities for building intelligent systems that truly understand human language. As research deepens, how to strike a balance between these powerful capabilities and reliability and interpretability will become an important topic for future semantic representation research.

4. Conclusions

The evolution of semantic representation in natural language processing presents a clear technical path: “static word embedding - dynamic context encoding - generative

language model.” This evolution is not simply a performance improvement, but a paradigm shift in understanding the essence of language: from isolated lexical representation to contextualized semantic understanding, from explicit vector geometric relationships to implicit model reasoning capabilities. Static word embedding established the feasibility of the distributional hypothesis but was limited by the discreteness of word meaning. Dynamic context encoding achieved deep bidirectional representation through the Transformer architecture but was limited by the separation of understanding and generation. Large language models, on the other hand, unify understanding and generation with autoregressive generation, demonstrating emergent semantic control, but face new challenges such as trustworthiness and interpretability.

Currently, semantic representation research is at a new crossroads. The diminishing marginal benefits of scale expansion indicate that simply increasing parameters is no longer the optimal solution. Future breakthroughs will depend on the interpretability and reasoning abilities brought by neuro-symbolic integration, the perceptual basis provided by multimodal unified representation, the computational efficiency introduced by cognitively inspired architectures, and the context adaptability conferred by embodied interaction. These directions collectively point to a core goal: to build a semantic representation system that combines the depth of human language understanding with the precision of machine computation. Achieving this goal requires establishing a new balance between representation ability and verifiability, statistical learning and symbolic reasoning, data-driven and knowledge guidance, and promoting the development of natural language processing in a more reliable, interpretable, and efficient direction.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Qu, J.B.; Wang, M.Y. Research on Topic Evolution Based on BERTopic Model and Semantic Function. *Science and Technology Information Research*, 2025, 7(04), 13-23.
- [2] Hu, P. Research on Semantic Communication Knowledge Co-evolution Based on Federated Learning. University of Electronic Science and Technology of China, 2025.
- [3] He, C.B.; Cheng, Q.W.; Cheng, J.W. et al. Method for Semantic Community Discovery and Evolution Analysis of Dynamic Attribute Network. *Journal of Electronics*, 2024, 52(11), 3757-3768.
- [4] Chen, B.T.; Kang, Y.J. Exploration of Cross-Domain Evolution of Academic Innovation Based on Distributed Semantic Analysis. *Library and Information Work*, 2024, 68(12), 95-108.

- [5] Lian, Z.P. Research and Practice Strategies on the Application of Natural Language Processing Technology in Fast Culture Translation and Communication. *East China Science and Technology*, 2025, 09, 117-119.
- [6] Wang, L.Z. Application and Optimization of Natural Language Processing Technology in the Intelligent Customer Service Platform of All Channels. *China Broadband*, 2025, 21(09), 22-24.
- [7] Zhang, J.Y.; Li, X.Q. Review of the Application of Deep Learning in Natural Language Processing. *Computer Knowledge and Technology*, 2025, 21(19), 23-25.
- [8] Lin, J.; Zhou, Y.Z.; Song, Y.F. et al. Technical Application and Implementation of Large Models in Semanticization of Standard Data Structures. *Science and Innovation*, 2025, 08, 222-224+228.



© 2025 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)