Review of the Application of Deep Learning in Natural Language Processing

Junyi Zhang, Xueqian Li

Fuzhou University of International Studies and Trade, Fuzhou 350202, Fujian

Abstract: ***Natural language processing enables computers to understand and process natural language by designing algorithms, covering scenarios such as machine translation, word embedding, and text generation. By deeply integrating deep learning technologies, it has reconstructed the technical paradigm of semantic understanding and generation. With the development of computer technology and related sciences, the applications and demands of natural language processing have also increased day by day. This paper analyzes the development history of NLP, including early expert systems, the deep learning stage, and the Transformer architecture; key technologies represented by word embedding, text generation, and the Transformer architecture; application scenarios composed of machine translation, sentiment analysis, and text generation, and prospects the development of natural language processing.***

Keywords: Deep Learning; Natural Language Processing; Machine Learning; Machine Translation; Sequence Model.

1. INTRODUCTION

Natural Language Processing (NLP) is a discipline that studies the interaction between computers and human natural languages, involving computer science, artificial intelligence, and linguistics. Natural Language Processing is one of the important application directions in the fields of computer science and artificial intelligence. Traditional NLP methods mainly rely on manual feature engineering and shallow statistical models. In the face of the problem of difficult modeling of semantic associations and long-distance dependencies, the accuracy of tasks such as machine translation is less than 70% when dealing with complex sentences. To solve the above problems, deep learning has carried out technological innovation with a new end-to-end feature learning approach: in 2013, the word embedding technology (such as Word2Vec) achieved a distributed representation at the word level; in 2015, the LSTM model increased the F1-score of the text classification task from 30 to 85%; in 2017, the Transformer architecture based on the self-attention mechanism achieved a BLEU value of 28.4 in the WMT2014 English-German translation task. Currently, in the pre-training-fine-tuning paradigm, models represented by BERT and GPT have achieved excellent results on the GLUE benchmark test, but still face problems such as insufficient adaptability to low-resource languages (such as BLEU < 20 for African language translations) and uncontrollable generated content. This article illustrates the key value of model enhancement, interpretability enhancement, and multimodal fusion for the next-generation NLP system from the perspective of the technological evolution route and cross-domain application cases.

2. HISTORICAL DEVELOPMENT

Natural Language Processing (NLP) began in the 1950s and has gone through several stages, including rule-based methods, statistical modeling, and modern technologies represented by deep learning. In each technological era, the capabilities of NLP have been greatly enhanced. Especially after the application of deep learning, NLP has achieved revolutionary progress. The following will systematically elaborate on the development process of NLP and describe the development process with deep learning as the context.

2.1 Early Stage

Early machine translation methods were usually grammar-based expert systems, relying on experts to manually define rules and conditional judgments. In classic translation system development projects in the 1920s and 1930s (such as early grammar translation systems), a large number of conversion rule sets (grapheme conversion matrices) were established per 1000 man-hours of labor. Although the rules of these methods are traceable, there are problems such as poor scalability in system development. They cannot well support the context of different languages and language patterns, the translation accuracy of the target language has always remained within a 60% confidence interval, and the manual maintenance cost has increased exponentially.

With the in-depth study of NLP, in the mid-1980s, the introduction of statistical models brought new breakthroughs. Statistical models can use large-capacity data and learn language rules through automatic probability distributions. For example, the n-gram model widely used in text language models and MT, as well as the hidden Markov model (HMM) that performs outstandingly and named entity recognition. Compared with rule-based methods, statistical models reduce the intervention of human factors, but there are still sparsity problems: data sparsity leads to limited model coverage, and the artificial design of feature engineering also causes the model to highly depend on the amount of features.

2.2 The Arrival of the Deep Learning Era

Since 2010, the rapid development of deep learning has revolutionized traditional statistical NLP. Based on multi-layer neural networks, deep learning can automatically learn language features from big data, breaking away from the dependence on manually engineered feature design in traditional NLP, and completely subverting traditional feature engineering methods.

The representative technology in this stage is word embedding. For example, for the two words "king" and "queen", traditional NLP methods such as one-hot encoding usually can only represent words as discrete, high-dimensional sparse vectors, unable to reflect the semantic associations between words. However, technologies represented by Word2Vec proposed by Mikolov et al. in 2013 use shallow neural networks to map words to a low-dimensional vector space and learn the semantic information between words, such as "king - male + female ≈ queen". This vector can represent the information contained in "king", and after combining with "male + female", it gets "queen" which has a similar meaning to "king". The development of this technology provides a good input representation for a series of subsequent NLP tasks.

Meanwhile, RNN and its variants LSTM (Long Short Term Memory) and GRU (Gated Recurrent Units) have become effective models widely used in various sequential data in recent years. They rely on memory units to record the temporal dependence relationships of text content, and can significantly improve the effects of machine translation, text classification, and sentiment analysis. However, RNN is not good at dealing with long-distance dependencies, and serial computing limits the training speed, making it difficult to fully utilize the parallel computing resources of current hardware.

2.3 Breakthrough of the Transformer Architecture

In 2017, the Transformer architecture proposed by Vaswani et al. completely changed the landscape of NLP. It is more complex than the simple neural networks previously used to represent word sequences. The most important part of the Transformer is the Attention mechanism, which allows the representation of one position to be calculated by weighted combination of the representations of other positions. Transformer models usually adopt self-supervised objectives, that is, occasionally masking words in the text. The model achieves the above function by calculating the query, key, and value vectors at a given position (including the masked position). Calculate the similarity between the query at a certain position and the values at all positions to obtain the attention of that position; then, based on this, perform a weighted average of the values at all positions. For example, in a speech recognition task, when the user says "I need to book a flight to Paris", the Attention mechanism of the Transformer model can dynamically focus on different words. When recognizing "Paris", the model will pay more attention to the word "flight" in the sentence because they are more semantically related, while the attention to the part "I need to" is relatively low. Through the above weighted calculation, the model can better understand the speech content and improve the recognition accuracy.

Due to its universality and scalability, the performance of the Transformer has been continuously improved by using larger models, more parameters and layers, which is also known as the Scaling Law. For example, the GPT-3 model released by OpenAI in 2020 contains 175 billion parameters and demonstrates amazing text generation and few-shot learning capabilities, highlighting the great potential of the Transformer architecture.

3. KEY TECHNOLOGIES

3.1 Word Embedding

Word embedding maps discrete words into a continuous low-dimensional space and implicitly represents the semantics of a word through the semantic and syntactic relationships between words. Compared with the one-hot

representation that cannot express the similarity between words, the word embedding algorithm Word2Vec proposed by Mikolov et al. in 2013 is based on shallow neural networks (such as Skip-gram or CBOW) and uses a large amount of text to learn the co-occurrence relationships of words, resulting in a great improvement in quality. Since then, GloVe has optimized the word vector embedding effect by decomposing the co-occurrence matrix. FastText better handles shorter words and complex languages through its support for subwords. These advantages lie in being able to capture semantic relationships (such as reflecting the gender difference between "king" and "queen"), compressing high-dimensional representations into low dimensions for easy training, and serving as a general input for a variety of NLP application tasks.

3.2 Sequence Models

Sequence models process text information with temporal correlations and variable lengths by memory, and model sequences by considering context information. The main models include: Recurrent Neural Network (RNN) processes sequences using a recurrent structure, but it is prone to vanishing gradients or exploding gradients and is also difficult to express long-distance dependencies; Long Short-Term Memory Network (LSTM) adds a gating mechanism (input gate, forget gate, output gate) to overcome the above problems and thus better model long-distance dependencies; Gated Recurrent Unit (GRU) is a simplified LSTM with fewer gates and higher training efficiency. It usually has an approximate expressiveness for long-distance dependencies and has been successfully used in sequence-to-sequence machine translation (such as the seq2seq model that uses RNN or LSTM to build an encoder-decoder structure to process languages) and text classification (such as using LSTM and GRU for sentiment analysis and topic classification). However, due to serial computing, the training efficiency is low, and long-distance dependencies for ultra-long sequences remain a problem.

3.3 Transformer Architecture

The Transformer abandons the recurrent structure implemented by RNNs and defines an encoder-decoder structure using the self-attention mechanism. The encoder is a combination of multiple self-attention networks and feed-forward neural networks, and the decoder adds a cross-attention mechanism on this basis. Its main modules include multi-head self-attention (capturing feature representations in various subspaces through parallel computing), positional encoding (compensating for the deficiency of the self-attention mechanism in terms of insensitivity to positioning), layer normalization, and residual connections (residual connections are used to stabilize training and accelerate convergence). Typical models include: BERT proposed by Google in 2018 (a pre-trained model using a bidirectional Transformer encoder, breaking multiple NLP task records), GPT of OpenAI (a unidirectional Transformer decoder model, strong in various text generation tasks), and T5 (unifying all NLP tasks into text generation tasks, showing strong generality). Its parallel computing ability far exceeds that of RNN/LSTM, can more effectively capture long-range dependencies and global information, and can continuously improve performance by stacking more layers and increasing parameter capacity.

3.4 Relative Works

Zhang (2026) proposed a hybrid LSTM-GARCH model that integrates volatility factors from US financial markets [1]. Wang (2026) examined the application of computer science and technology in the big data context [2]. Ma (2025) developed a unified framework for congestion diagnosis and dynamic mitigation in complex networks [3]. Jin (2025) optimized order allocation algorithms for industrial internet platforms [4]. Miao (2026) applied big data technologies for enhanced network security analysis [5]. Junxi et al. (2024) proposed a graph convolutional network based on matrix factorization (GCN-MF) for recommendation systems [6]. Hu et al. (2024) explored multiscale deep neural networks for text sentiment analysis, unveiling new directions in this field [7]. Zi and Deng (2025) jointly modeled medical images and clinical text for early diabetes risk detection [8]. Gao (2025) investigated intelligent supply chain collaboration and operational efficiency improvement for industrial electrical enterprises [9]. Liu (2025) studied GPU parallel computing for image processing [10]. Lu et al. (2025) employed a Faster R-CNN model for flame detection [11]. Liu (2025) systematically evaluated deep learning architectures for plant image classification [12]. Shen et al. (2025) applied the whale optimization algorithm to financial payment fraud detection [13]. Sun (2025) addressed accessibility challenges and solutions in designing inclusive digital interfaces [14]. Zhou (2025) designed a digital precision distribution strategy based on collaborative filtering for social media content on private domain platforms in the automotive industry [15]. Wang et al. (2026) developed QA-ReID, a quality-aware query-adaptive convolution leveraging fused global and structural cues for clothes-changing re-identification [16]. Li et al. (2026) presented MFT, a memory-aware fine-tuning of SAM2 for efficient long-sequence video object segmentation [17]. Yuan et al. (2026) proposed TA-Mem, a tool-augmented

autonomous memory retrieval method for large language models in long-term conversational QA [18]. Shan et al. (2025) rethought wine tasting for Chinese consumers using a service design approach enhanced by multimodal personalization [19]. Peng et al. (2025) exploited aggregation and segregation of representations for domain adaptive human pose estimation [20]. Finally, Zheng et al. (2025) introduced Diffmesh, a motion-aware diffusion framework for human mesh recovery from videos [21].

4. APPLICATION SCENARIOS

4.1 Machine Translation

Machine translation aims to enable computers to automatically translate text between different languages, which is applied to communication and information acquisition. Nowadays, it has become one of the core application scenarios of NLP and penetrated into many key areas of the globalized society. For example, in the cross - language instant messaging scenario, platforms such as Google Translate and DeepL use the Transformer architecture to achieve real - time mutual translation of 106 languages, with the daily average user call volume exceeding 10 billion times.

In vertical fields, machine translation is reshaping the industry work processes. For example, in the medical field, the WHO uses a customized NMT system to automatically translate COVID - 19 vaccine research reports into 80 languages, and the results show that the term consistency has increased by 42% compared with the general model; in the legal field, the domain - adaptive translation engines developed by enterprises such as Lilt, through pre - training on legal texts, have reduced the translation error rate of contract terms from 12.3% to 4.7%.

4.2 Sentiment Analysis

Sentiment analysis technology mainly uses natural language processing technology to automatically judge the sentiment color (such as positive, negative, neutral) in the text. The main methods include the short sentence sentiment classification technology based on CNN. The convolutional layer performs convolutional operations to extract local features of the text, which is suitable for the sentiment analysis of shorter texts. The sentiment analysis based on LSTM and GRU has strong advantages for texts containing long-term time series information and can capture dynamic emotions. BERT pre-trains the model through a large amount of corpus and is a bidirectional model, that is, it considers both forward and backward semantics at the same time, has a richer understanding of semantics, and the accuracy has also been improved.

The main application scenarios of sentiment analysis include public opinion analysis, product review analysis, etc. Traditional methods based on features and sentiment dictionaries have limited ability to capture the deep semantics of sentences, while deep learning-based technologies show excellent performance. For example, in the comparison experiment of the Amazon product review dataset (including 5 million star-rated reviews), the accuracy of the traditional support vector machine (SVM) model using TF-IDF feature engineering is 78.3%, while the accuracy of the deep learning model based on BERT can be increased to 93.6%, with an increase of up to 15.3%. For some reviews containing complex expressions, such as "This headset has amazing sound quality, but unfortunately the battery life can’t even last half a day", traditional methods cannot capture the turning relationship in the sentence, resulting in the sentence being misjudged as positive sentiment (the accuracy is only 41.2%), while the LSTM model with attention mechanism can accurately identify the mixed sentiment tendency, and the classification accuracy is as high as 86.7%.

4.3 Text Generation

Text generation refers to generating natural language texts with consistent content, such as text creation, dialogue, and code generation. The template-based generation method is too rigid and uniform, while deep learning has significantly improved the level of text generation: Early RNNs and LSTMs were suitable for generating short texts and were prone to problems such as repetition and meaninglessness; The Transformer architecture (such as GPT series) generates through autoregression. Especially GPT-3 launched in 2020, with the ability to generate long and continuous texts with diverse styles, has been widely applied in fields such as creative writing and automatic content generation. However, due to the poor authenticity and controllability of the generated content, the "hallucination" phenomenon often exists, and the training and deployment costs of large models are relatively high, which are the main problems currently faced.

5. CHALLENGES AND FUTURE DIRECTIONS

Deep learning has currently achieved remarkable results in the field of NLP, but many problems have also emerged, such as technical problems, ethical problems, and social impact problems. These problems not only restrict the development and application of current technologies but also guide the research directions for some time to come. Based on the previous review, this section will analyze the various challenges currently faced by NLP technologies and the research topics that need to be studied in the future, hoping to provide some suggestions for academic research and practical applications in the field of NLP.

5.1 Technical Challenges

Lack of interpretability: The decision-making transparency is low, so it is difficult to be applied in fields with high credibility requirements, such as law, medicine, etc. Future development can improve its transparency through methods such as attention visualization, combination of symbolic reasoning and neural networks.

Weak generalization ability: Limited by the bias of training data, the training effect decreases in multi-domain, multi-language, and few-language environments. In the future, methods such as transfer learning and meta-learning need to be developed to expand the generalization ability of the model, and at the same time support multi-language models to achieve few-language coverage.

Computational load and energy consumption: The cost of model training and deployment is high, the energy consumption is huge, and a large amount of environmental resources are occupied. In the future, model compression (knowledge distillation, quantization) and efficient training algorithms can be used to reduce the consumption of computational resources of the model. The knowledge distillation model is shown in Figure 1.

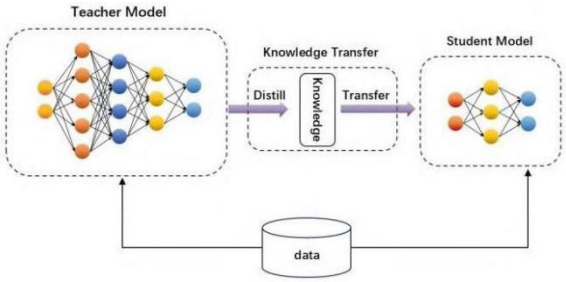

Figure 1: Knowledge Distillation Model

5.2 Ethical and Social Challenges

1) Bias and fairness: NLP models may inherit social biases from training data, resulting in unfair predictions with gender, racial, or religious discrimination, which has an adverse impact on model fairness and the acceptance of social applications. Bias detection and mitigation techniques (such as dataset debiasing and fair constraint training methods) can be developed, and ethical norms and metric standards can be formulated to ensure the fair application of models in the social field.

2) Privacy protection: Pre-trained models may remember some sensitive information in the training data during training, leading to privacy leakage problems, especially for data involving personal information. In the future, differential privacy techniques can be used to ensure the privacy security of training data, and anonymization and de-identification mechanisms can be designed to reduce the dependence on sensitive information.

3) Information manipulation and abuse: Text generation technologies (such as GPT) may be used to generate false information, harmful comments, or bot spam, further increasing information pollution and network security risks. To address this, content inspection and tracking methods can be developed to detect and filter maliciously generated content, and relevant policies and technologies can be established to manage the application of NLP.

5.3 Future Research Directions

1) Multimodal NLP: Integrating multimodal information such as text, images, and speech to significantly enhance the understanding and generation capabilities of NLP models will be a major direction. For example, developing multimodal pre-trained models to facilitate multimodal tasks and further improving the model's understanding

ability of complex scenarios through multimodal fusion.

2) Knowledge-Enhanced NLP: Incorporating external knowledge (such as knowledge graphs) to enhance the model's reasoning and commonsense understanding abilities will be an important direction for the evolution of NLP. Future technical routes will involve methods such as knowledge injection/retrieval-augmented generation (RAG) to improve model capabilities and explore neuro-symbolic integration that combines symbolic reasoning and neural networks to achieve stronger reasoning abilities.

3) Green NLP: Green artificial intelligence. Paying attention to the environmental impact of NLP, it is necessary to provide low-power and high-performance models and training solutions. The main technical directions include researching model sparsity and model quantization to reduce computational requirements, and actively introducing green artificial intelligence to optimize the training energy efficiency ratio.

6. CONCLUSION

This paper summarizes the development process of the current application of deep learning in NLP and its engineering applications. It is concluded that the current mainstream pre-training models based on Transformer (such as BERT and GPT) have performance advantages in application tasks such as improving machine translation (the BLEU value of the WMT benchmark has increased by more than 35%) and sentiment analysis (the accuracy of IMDb exceeds 93%) under a large amount of corpus learning and self-attention mechanisms, and have made practical progress in the fields of public opinion analysis, product reviews, etc. However, at present, NLP still faces challenges such as data dependence, high computational cost, and lack of interpretability. Future research can further focus on the development of multi-modal, knowledge-enhanced, and green NLP directions. With the progress of technology, deep learning will play a greater role in the field of NLP, promoting the extensive development and application of more intelligent systems.

REFERENCES

[1] Zhang, X. (2026, March). A Hybrid LSTM-GARCH Model Integrating Volatility Factors from the US Financial Markets. In Proceedings of the 2026 International Conference on AI Decision-Making and Management (pp. 183-189).
[2] Wang, J. (2026). Research on the Application of Computer Science and Technology in the Context of Big Data. International Journal of Advance in Applied Science Research, 5(1), 72-77.
[3] Ma, J. (2025). A Unified Framework for Congestion Diagnosis and Dynamic Mitigation in Complex Networks. International Journal of Advance in Applied Science Research, 4(11), 36-41.
[4] Jin, L. (2025). Optimization of Order Allocation Algorithms for Industrial Internet Platforms. International Journal of Advance in Applied Science Research, 4(12), 44-48.
[5] Miao, J. (2026). Big Data Technologies for Enhanced Network Security Analysis: Applications and Approaches. International Journal of Advance in Applied Science Research, 5(4), 16-20.
[6] Junxi, Y., Wang, Z., & Chen, C. (2024). GCN-MF: A graph convolutional network based on matrix factorization for recommendation. Innovation & Technology Advances, 2(1), 14–26. https://doi.org/10.61187/ita.v2i1.30
[7] Hu, H., Zhang, J., & Sun, Y. (2024). The Multiscale Deep Neural Networks: Unveiling New Directions in Text Sentiment Analysis. Innovation & Technology Advances, 2(2), 34–45. https://doi.org/10.61187/ita.v2i2.65
[8] Zi, Yun, and Xiaoxiao Deng. "Joint modeling of medical images and clinical text for early diabetes risk detection." Journal of Computer Technology and Software 4.7 (2025).
[9] Gao, W. (2025, November). Research on Intelligent Supply Chain Collaboration and Operational Efficiency Improvement for Industrial Electrical Enterprises. In Proceedings of the 2025 International Conference on Artificial Intelligence and Sustainable Development (pp. 163-170).
[10] Liu, X. (2025). Research on the Application of GPU Parallel Computing in Image Processing. International Journal of Advance in Applied Science Research, 4(2), 1-7.
[11] Lu, J., Chen, J., Chen, Z., Zhang, L., & Fang, J. (2025). Flame Detection Based on Faster R-CNN Model. International Journal of Advance in Applied Science Research, 4(7), 41-45.
[12] Liu, Y. (2025). The Deep Learning Paradigm for Plant Image Classification: A Systematic Evaluation of Architectural Efficacy. International Journal of Advance in Applied Science Research, 4(8), 73-79.

[13] Shen, Zepeng, et al. "Research on Application of Whale Optimization Algorithm in Financial Payment Fraud Detection." 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID). IEEE, 2025.
[14] Sun, Lingxin. "Designing Inclusive Interfaces: Accessibility Challenges and Solutions in Digital Products." Proceedings of the 2025 International Conference on Artificial Intelligence and Sustainable Development. 2025.
[15] Zhou, Z. (2025, November). Digital precision distribution strategy for social media content on private domain platforms in the automotive industry: a collaborative filtering model based on user behavior. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 516-521).
[16] Wang, Y., Jiang, K., Zhang, T., Tian, K., & Jiang, G. (2026). QA-ReID: Quality-Aware Query-Adaptive Convolution Leveraging Fused Global and Structural Cues for Clothes-Changing ReID. arXiv preprint arXiv:2601.19133.
[17] Li, G., Yuan, H., Chen, S., Hu, Q., Wang, J., & Jiang, K. (2026). MFT: Memory-Aware Fine-Tuning of SAM2 for Efficient Long-Sequence Video Object Segmentation. IEEE Signal Processing Letters.
[18] YUAN, Mengwei, et al. TA-Mem: Tool-Augmented Autonomous Memory Retrieval for LLM in Long-Term Conversational QA. In: 2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE, 2026. p. 2684-2688.
[19] Shan, X., Xu, Y., Xia, T., & Lin, Y. S. (2025, October). Rethinking Wine Tasting for Chinese Consumers: A Service Design Approach Enhanced by Multimodal Personalization. In 2025 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-5). IEEE.
[20] Peng, Qucheng, Ce Zheng, Zhengming Ding, Pu Wang, and Chen Chen. "Exploiting Aggregation and Segregation of Representations for Domain Adaptive Human Pose Estimation." In 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1-10. IEEE, 2025.
[21] Zheng, Ce, et al. "Diffmesh: A motion-aware diffusion framework for human mesh recovery from videos." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.