Advances in Engineering Innovation Vol.16 Issue 6 EWA Publishing
Available Online: 25 June 2025 DOI: 10.54254/2977-3903/2025.24550

Research progress and challenges of deep learning in Natural Language Processing

Yuhan Fan

Science and Technology College of NCHU, Nanchang, China

13862578606@163.com

Abstract. With the rapid development of artificial intelligence, Natural Language Processing (NLP) has emerged as a critical area for enabling intelligent human-computer interaction. This paper reviews key deep learning technologies and their applications in NLP. It first examines foundational techniques such as word embeddings and pre-trained models, and analyzes the structures and use cases of core models includingConvolutional Neural Networks (CNNs),Recurrent Neural Networks (RNNs) and their variants, as well as Transformers. It then explores the application of these models in tasks such as sentiment analysis, machine translation, and question-answering systems. The study highlights how pre-trained models like BERT and GPT significantly enhance semantic understanding through large-scale unsupervised learning. However, challenges remain, including limited interpretability, weak performance in low-resource languages, and inadequate multimodal integration. The paper concludes by discussing future directions such as lightweight model design, cross-lingual transfer learning, and deep multimodal fusion. This research aims to provide theoretical references for advancing NLP technology and enhancing its practicality across various domains.

Keywords: Natural Language Processing, artificial intelligence, human-computer interaction, machine learning, applications

1. Introduction

As the volume of text data generated daily increases, the ability to process and understand natural language has become crucial for both academic and industrial applications. Natural language processing has emerged as a key area in artificial intelligence, aiming to bridge the gap between human language and computer understanding [1,2]. Traditional NLP methods rely on manually designed features, facing issues such as sparse semantic representation and weak generalization capabilities. Deep learning, through multi-layer nonlinear networks that automatically extract features, fundamentally transforms the research paradigm of NLP. In 2006, Hinton's Deep Belief Network (DBN) opened up the exploration of deep learning in the field of NLP [1]. Subsequently, breakthroughs in word embeddings (Word Embedding), pre-trained models (such as BERT, GPT), and other technologies have driven significant progress in NLP for semantic understanding and generation tasks [2-4]. This paper systematically reviews the core technologies, typical applications, and challenges of deep learning in NLP, providing a reference for research in the field. It also reviews the current state of NLP technology, its applications in different fields, and the challenges it faces. Combining literature reviews and case studies, this study aims to outline the development and future prospects of NLP [2]. The research questions include: What are the main applications of NLP? What are the current limitations and future directions of NLP development? This study is significant because it offers insights on how to further optimize NLP to meet the growing demands of the digital world [2].

2. Core technologies of deep learning in NLP

2.1. Basic representation learning techniques

Word vector technology addresses the sparsity issue of traditional One-Hot encoding by mapping words to continuous vector space. The neural network language model (NNLM) proposed by Bengio et al. was the first to introduce word vectors into NLP [2]. Subsequently, Word2Vec developed by Mikolov et al. achieved efficient word vector training through Skip-Gram and CBOW models, making semantically similar words co-occur in the vector space [5]. Later, GloVe combined global co-occurrence statistics to further enhance the semantic representation capability of word vectors [6].

Pre-training models represent a milestone breakthrough in the NLP field. Early ELMo achieved context-sensitive word representation through bidirectional LSTM, while BERT employed masked language model (MLM) and next-sentence prediction (NSP) tasks, fine-tuning after pre-training on large-scale corpora, significantly enhancing performance across various NLP tasks [3,7]. The GPT series, based on Transformer decoders, generates high-quality text through self-supervised learning, driving advancements in generative tasks such as text summarization and dialogue systems [4,8].

2.2. Core network architecture

In the field of natural language processing, the development of core network architectures has significantly enhanced the modeling capabilities of text features. Currently, the three main types of network architectures are: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and their variants (such as LSTM and GRU), and the Transformer architecture that has emerged in recent years. Each excels in different semantic modeling tasks and boasts its unique structural advantages.

CNN excels in local feature extraction from text. Its fundamental principle is to extract contextual features through sliding window convolution operations, followed by dimensionality reduction and abstraction via pooling layers, thereby capturing multi-granularity semantic information. Different sizes of convolutional kernels can extract n-gram structures of varying lengths, enabling feature modeling at both the phrase and short-term level. For example, Kim's TextCNN model, in sentiment classification tasks, leverages a multi-kernel structure to achieve efficient text classification [9]. The core advantage of CNN lies in its high computational efficiency and ease of parallelization, making it particularly suitable for tasks that heavily rely on local structures, such as short text classification and short text semantic matching. Additionally, CNN has a low dependence on manual feature engineering, allowing it to directly learn discriminative features from data, thus performing exceptionally well in short text processing scenarios.

RNN and its variants (such as LSTM, GRU) excel at handling sequence dependencies and are widely used in time-series tasks like machine translation and speech recognition [10,11]. For example, Sutskever et al. 's Sequence-to-Sequence (Seq2Seq) model, which combines an LSTM encoder-decoder, achieves end-to-end translation tasks [12]. The original design aims to process sequential data by encoding historical input information into the current computation through cyclic passing of hidden states, thereby capturing long-term dependencies within sequences. Traditional RNNs struggle with modeling long-distance dependencies due to issues like gradient disappearance and gradient explosion. To address this, LSTM (Long Short-Term Memory network) and GRU (Gated Recurrent Unit) introduce gating mechanisms (such as forget gates, input gates, output gates) to selectively retain or forget long-term information, effectively alleviating the problem of gradient disappearance. In natural language processing, RNNs and their variants are widely applied in machine translation (such as the Seq2Seq model using an encoder-decoder architecture to handle sequence conversion from source to target languages), speech recognition, text generation (such as poetry creation, dialogue systems), and other time-series related tasks. Their core advantage lies in their natural adaptation to the dynamic characteristics of sequential data, enabling explicit modeling of text's sequential dependencies, which is particularly essential in tasks that require capturing contextual sequence semantics.

In contrast, the Transformer architecture, leveraging self-attention mechanisms (Self-Attention) to capture long-range dependencies, has become the mainstream model in current NLP, known as [13]. Its parallel computing capabilities address the sequence dependency bottleneck of RNNs, as demonstrated by Vaswani and others showing that Transformer outperforms traditional RNN models in machine translation tasks [13]. Subsequent improvements, such as XLNet, optimize the pre-training process by arranging language models, further enhancing long-text processing capabilities [14]. The principle behind this is to calculate the relevance weights between each word in the input sequence and all other words, dynamically capturing semantic dependencies across the entire text without relying on recursive or convolutional operations to model long-range dependencies.

Transformer is typically divided into two parts: the encoder and the decoder. The encoder extracts contextual representations layer by layer through multi-head self-attention and feedforward neural networks, while the decoder generates target sequences based on the encoder's output, combining self-attention with cross-attention. In natural language processing, Transformer has become a mainstream model architecture, supporting landmark technologies such as BERT (a pre-trained model based on the encoder, excelling in semantic understanding), GPT (a generative model based on the decoder, excelling in text generation), and T5 (a text-to-text framework). It is widely applied in tasks like machine translation, question answering systems, text summarization, and sentiment analysis, particularly demonstrating strong generalization capabilities in large-scale pre-training scenarios. Its significant advantages include: fully parallelized computation, breaking the speed bottleneck of RNN sequence processing; efficient modeling of long-distance dependencies through self-attention mechanisms, avoiding the gradient disappearance issue common in traditional RNNs; flexible and scalable architecture, making it easy to integrate with pre-training techniques and support the entire chain of tasks from understanding to generation, thus becoming a cornerstone of current NLP technology development.

3. Typical application scenarios and technical progress

3.1. Sentiment analysis

In practical applications, the RNN-based sentiment stratification modeling method has demonstrated significant advantages in fine-grained sentiment analysis tasks. For example, Socher et al. conducted experiments on the Stanford Sentiment Treebank (Stanford Sentiment Treebank, SST) in 2013, where they constructed a recursive neural network based on syntactic tree structures to categorize sentiments into five fine-grained categories: "very negative," "negative," "neutral," "positive," and "very positive." In long text sentiment analysis involving complex syntactic structures, the model's ability to aggregate sentiment levels across multiple branches significantly outperformed traditional bag-of-words models. Ultimately, the accuracy rate on the full SST dataset reached 48.3%, an improvement of about 10 percentage points compared to baseline methods based on hand-crafted features at that time. Particularly, in handling nested negations (such as "not dislike") and scenarios with varying degrees of sentiment intensity (such as "somewhat satisfied" to "extremely satisfied"), the classification accuracy increased by over 15% [15].

The BERT-Finetuning method, combined with pre-trained models, performs more prominently in cross-domain sentiment analysis. Taking the multi-domain sentiment analysis benchmark dataset Multi-Domain Sentiment Dataset as an example, this dataset includes Amazon product reviews, Yelp restaurant reviews, and IMDB movie reviews from various domains. After fine-tuning, the BERT model significantly enhances its ability to capture implicit emotions (such as expressing dissatisfaction through irony, "The special effects of this movie are 'really great,' keeping me 'on the edge of my seat' throughout") and domain-specific emotional words (such as "landmine" in e-commerce and "pitfall" in dining). In cross-domain testing, it achieves an average accuracy rate of 92.3%, which is 8.7 percentage points higher than traditional deep learning models (such as LSTM + Attention). On the SST-2 (binary classification) and SST-5 (five-category) datasets, BERT-Finetuning achieves accuracies of 94.9% and 89.5%, respectively. Notably, in sub-corpora containing sarcastic sentiments, the accuracy of sentiment judgment for texts like "The only good thing about this movie is that it taught me how to fall asleep quickly" improves by 22% compared to baseline models, fully demonstrating the deep understanding of complex contextual semantics by pre-trained models [3].

3.2. Machine translation

In practical applications, the neural machine translation model based on Transformer has shown revolutionary breakthroughs. Taking the Transformer model deployed by Google Translate in 2016 as an example, it achieved significant results in English-to-German translation tasks. By integrating the Encoder-Decoder architecture with an 8-head attention mechanism, the source language sentence "Breaking down barriers in language understanding" was encoded into a vector sequence containing global semantic dependencies. The decoder dynamically aligned the generation process of the target language word "Herausforderungen in der sprachlichen Verstandigung niederbrechen" through cross-attention. In the WMT 2016 German-English translation task, the BLEU score reached 29.9, improving by 3.3 points compared to the previous LSTM-based Seq2Seq model, and the fluency of the translation increased by over 40%. Particularly in long sentence translations (such as legal clauses and scientific literature), the accuracy rate of maintaining the semantics of nested clauses improved from 68% to 89% [13].

The multi-language pre-trained model mBERT has shown remarkable effectiveness in low-resource language translation. For low-resource language pairs such as Swahili-English and Nepali-English, mBERT leverages cross-lingual transfer capabilities to initiate translation tasks with just 2,000 parallel sentences. In the FLORES-101 low-resource language benchmark test, the BLEU score for Swahili-English reached 18.7, an improvement of 9.2 points compared to traditional statistical machine translation methods that rely on large-scale corpora; the subject-verb agreement error rate for Nepali-English translation dropped from 35% to 12%. Moreover, in "zero-shot" scenarios without parallel data (such as Latin-English), mBERT achieved basic semantic alignment through shared cross-lingual semantic spaces, with a translation coherence rate of 65%, far exceeding the 32% [16] of traditional rule-based methods. These technological advancements have driven the implementation of neural machine translation in scenarios like medical consultations and cross-border e-commerce, enabling users of less common languages to access real-time translation services with accuracy improvements of over 30%, significantly lowering the barriers to cross-lingual communication.

3.3. Question answering system

The question answering system requires the model to understand the context and generate accurate answers. The SQuAD dataset has driven the development of reading comprehension tasks. Models based on Transformer, such as BERT and RoBERTa, learn deep semantic information through masked mechanisms and perform excellently in extractive question answering, [17,18]. Generative question answering models like GPT-3 can generate coherent responses based on context, but they have issues with factual errors and insufficient logical reasoning capabilities. Subsequent research has enhanced the reliability of answers by introducing knowledge graphs to strengthen factual constraints, [19].

In practical applications, the neural machine translation model based on Transformer has revolutionized traditional translation paradigms with its Encoder-Decoder architecture and multi-head attention mechanism. Taking the WMT 2017 English-French translation task as an example, the model using the Transformer architecture achieves precise semantic alignment between source and target languages through an 8-layer encoder-decoder and a 16-head attention mechanism. In scientific text translation that includes long-distance dependencies, the model's accuracy in processing the subordinate clause structure "Le developpement durable exige une approche integree des enjeux environnementaux, economiques et sociaux" (sustainable development requires a comprehensive approach to environmental, economic, and social issues) reaches 92%, a 27% improvement over earlier LSTM-based models. The BLEU score jumps from 33.3 to 41.0, with the fluency of the translation approaching human-level [13]. Google Translate In the actual deployment, for the complex syntactic transformation of Chinese-English, the mapping relationship between "subject-predicate-object" and "state-content structure" was dynamically adjusted to focus on attention weight, so that the accuracy rate of professional terms in business contract translation was increased from 78% to 95%, and the sentence structure completeness reached 91%, significantly reducing the cost of manual proofreading.

The multi-language pre-trained model mBERT stands out in cross-lingual transfer for low-resource language translation. In the Hausa-English task of the FLORES-101 benchmark, mBERT achieved a BLEU score of 15.2 with only 5,000 parallel sentences, an improvement of 11.5 points over traditional statistical machine translation methods, fundamentally changing the history of small language translation that relied on massive data. For "zero-resource" scenarios without any parallel corpus (such as Swahili-Spanish), mBERT achieves cross-lingual mapping of basic semantics through shared cross-lingual semantic spaces, with keyword accuracy reaching 68%, far surpassing the 42% of rule-based methods. In practical applications, Kenya's medical consultation platform leverages mBERT to translate Swahili medical records into English in real-time, increasing the retention rate of key information in disease descriptions from 55% to 82%, and reducing misdiagnosis rates by 30%. Nepal's e-commerce platform uses this technology to handle product detail pages in Nepali-Indian languages, lowering the error rate of attribute translations from 40% to 12%, and seeing a 25% increase in user inquiries, fully demonstrating the core value of cross-lingual transfer technology in inclusive translation [16].

4. Challenges and future directions

Lack of interpretability in models is one of the core challenges. Deep learning models, such as the multi-layer attention mechanism of Transformer, lack transparency and struggle to explain decision-making processes, limiting their application in fields with high interpretability requirements, like healthcare and law. Moreover, the reliance on large-scale data exacerbates resource imbalance issues, leading to significant performance degradation for low-resource languages (such as less common ones) due to insufficient training data. In terms of multimodal fusion, current models primarily handle text in a single modality, with weak capabilities for joint modeling of multimodal information such as speech and images, making it difficult to meet the demands of complex interactive scenarios.

Future research can explore directions such as lightweight model design, cross-lingual transfer learning, deep integration of multimodal data, the combination of symbolic logic and deep learning, and ethical and safety studies. In terms of lightweight model design, to address the high computational costs and difficulties in deploying pre-trained models on mobile devices, methods like knowledge distillation and pruning techniques can be used to reduce model parameters and computational complexity while retaining core semantic representation capabilities. For example, knowledge distillation technology uses a "teacher-student" model architecture to transfer knowledge from large models to lightweight models, while pruning techniques improve the efficiency of edge devices by removing redundant connections or optimizing neural network structures, thus promoting the widespread application of NLP technologies in smartphones, smart speakers, and other mobile devices [20,21].

Cross-lingual transfer learning aims to break the reliance of traditional models on large-scale parallel corpora by improving the cross-lingual semantic representation capabilities of pre-trained models, achieving language-independent modeling based on limited data. Current multilingual models like mBERT have initially demonstrated cross-lingual potential, but they still fall short in semantic alignment for low-resource languages. Future research can focus on building more efficient cross-lingual shared spaces, leveraging techniques such as contrastive learning and meta-learning to enhance models' ability to capture linguistic commonalities, enabling small language NLP tasks to achieve semantic understanding and generation without massive amounts of data. This will facilitate the implementation of translation and sentiment analysis technologies in scenarios involving lesser-known languages like Swahili and Nepali.

Multimodal deep integration aims to break the limitations of single-modal text processing, combining visual, audio, and image information with natural language processing to construct a unified multimodal representation framework. For example, in image-text understanding tasks, models need to process both textual descriptions and image content simultaneously, dynamically aligning visual features with semantic information through attention mechanisms; in speech dialogue systems, it is necessary to integrate speech recognition, natural language understanding, and speech synthesis technologies to achieve more natural human-computer interaction. Future research can explore cross-modal alignment algorithms and multi-source data fusion strategies to enhance the model's ability to jointly model multimodal information in complex scenarios, promoting the practical application of intelligent assistants and accessible communication.

The integration of symbolic logic and deep learning is a critical path to addressing the "hallucination" issue in generative models and enhancing logical reasoning capabilities. Current models like GPT are prone to factual errors or logical contradictions when generating text. By incorporating knowledge graphs, logical rules, and other symbolic knowledge, explicit constraints can be imposed on model outputs. For example, introducing entity relationships from knowledge graphs in question-answering systems can help models answer questions that rely on fact-based reasoning more accurately; embedding logical rules in legal document generation ensures that the text adheres to specific clause logic. Future research should explore efficient mechanisms for integrating symbolic knowledge with neural networks to achieve complementary advantages between data-driven and rule-driven approaches.

Ethics and safety research focuses on the social impact of NLP technology, addressing issues such as model bias, data privacy, and algorithmic fairness by developing systematic evaluation and solutions. For example, gender and racial biases hidden in training data can lead to discriminatory content in model outputs, which requires cleaning training corpora through data bias correction techniques or introducing fairness constraints into the models. In sensitive fields like healthcare and law, it is essential to ensure user data privacy during collection and processing. Future efforts should establish interdisciplinary research frameworks, combining sociological and legal theories to design fairness assessment metrics, develop explainable technologies to enhance model decision transparency, and ensure the social acceptability and sustainable development of NLP technology.

5. Conclusion

Deep learning has driven breakthroughs in semantic understanding and generative tasks through automatic feature extraction and end-to-end modeling. The emergence of pre-trained models has significantly enhanced the model's generalization ability, but there are still shortcomings in interpretability, resource efficiency, and multimodal integration. Future research should focus on lightweight, cross-lingual, and multimodal directions, combining symbolic knowledge to enhance the model's logical reasoning capabilities. At the same time, attention should be paid to technical ethics and social impacts, promoting the advancement of NLP technology towards general intelligence.

This article summarizes the current state, applications, and challenges of natural language processing technology. The research findings indicate that while NLP has made significant progress, there are still areas requiring further study and development. Major applications of NLP, such as machine translation and sentiment analysis, have shown promising results, but challenges like context understanding and multilingual support remain. Future research should focus on developing more advanced models and addressing ethical issues in NLP. This study highlights the importance of NLP in enhancing human-computer interaction and improving efficiency across industries. However, it also acknowledges the limitations of current research and suggests that future work should concentrate on improving the accuracy and reliability of NLP systems.

References

[1] Jurafsky, D. , & Martin, J. H. (2009). *Speech and Language Processing* (3rd ed.). Pearson.

[2] Manning, C. D. , Raghavan, P. , & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press.*

[3] Brown, P. F. , et al. (1990). A statistical approach to machine translation. *Computational Linguistics, 16*(2), 79-85.

[4] Socher, R. , et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631-1642.

[5] Vaswani, A. , et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[6] Papineni, K. , et al. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.

[7] Pennington, J. , Socher, R. , & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543.

[8] Devlin, J. , et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

[9] Li Xinchun, Du Xinyi, Xu Chi, et al. (2025). An Unsupervised Industrial Control Protocol Classification Method Based on Multi-scale Feature Deep Learning [J/OL]. *Information and Control*, (02), 241-250. https: //doi. org/10. 13976/j. cnki. xk. 2023. 5203.

[10] Wang Kuai, Zhang Qian, Yi Yuangang. (2025). Knowledge Graph Product Recommendation Method Based on Reinforcement Learning [J/OL]. *Computer Measurement and Control*, 1-8. http: //kns. cnki. net/kcms/detail/11. 4762. TP. 20250421. 0926. 010. html.

[11] Li Fajuan. (2025). An automatic error text detection system for translation robots based on improved Seq2Seq. *Electronic Design Engineering, 33*(08), 174-177+182. DOI: 10. 14022/j. issn1674-6236. 2025. 08. 036.

[12] Deng, X. (2025). Sentiment analysis based on attention mechanism and Bi-LSTM. *Scientific and Technological Innovation*, (09), 93-96.

[13] Zhou, Z. , Wang, H. , Wei, D. , et al. (2025). Research progress of Transformer in medical image segmentation. *Computer Engineering and Applications*, 1-22. Retrieved April 22, 2025, from http: //kns. cnki. net/kcms/detail/11. 2127. tp. 20250417. 0955. 002. html.

[14] Wang, C. , Du, J. , Wang, Y. , et al. (2025). Hot news recommendation method based on RoBERTa-BiLSTM-MA. *Software Engineering, 28*(04), 73-78. DOI: 10. 19644/j. cnki. issn2096-1472. 2025. 004. 015.

[15] Wang, X. , & Miao, X. (2025). WebShell upload detection based on XLNet. *Data Communication*, (01), 36-39+53.

[16] Hu, X. (2025). Optimization and regulation method of network resource transmission efficiency based on LSTM-RNN. *Information Technology and Informatization*, (01), 74-78.

[17] Ren, X. , Yang, R. , & Fu, Y. (2025). Exploration and practice of constructing AI operation and maintenance robots based on NLP technology. *Information System Engineering*, (01), 137-140.

[18] Cheng, S. (2024). Research on text representation learning method based on deep neural networks (Master's thesis). *University of Electronic Science and Technology of China*. DOI: 10. 27005/d. cnki. gdzku. 2024. 005745.

[19] Wang, Y. , Wang, K. , & Liu, M. (2025). A cross-lingual text similarity model based on alternate language data reconstruction method. *Journal of Jilin University (Science Edition), 63*(02), 551-558. DOI: 10. 13413/j. cnki. jdxblxb. 2024078.

[20] Wang, Z. , & Chen, H. (2019). Practical machine learning on mobile devices with TensorFlow. *Beijing: Publishing House of Electronics Industry.*

[21] Ma, Y. , & Cao, J. (2024). Research on knowledge graph of data element market based on NLP and LLM. *Information & Computer, 36*(24), 129-133.