Dean&Francis
ISSN 2959-6157

Evolution and Challenges of Natural Language Processing Technologies Based on Text Understanding and Generation

Junjie Li

School of Computer Science, Xi'an University of Posts & Telecommunications, Xi'an, 710500, China.
1686799661@qq.com

Abstract:

Natural Language Processing (NLP) began in 1950, initially relying on rules established by linguists to parse text, but this approach proved insufficient for handling complex linguistic phenomena. With the development of technology, machine learning methods have been introduced to improve processing efficiency and accuracy through data-driven approaches. Today, neural networks and pre-trained models, such as BERT and GPT, have become mainstream. With their powerful data learning capabilities and deep semantic understanding, they have greatly expanded the application boundaries of natural language processing, showing unprecedented performance in everything from machine translation to text generation. This paper systematically reviews the development of natural language processing technology, analyzes the evolution from early rule-based methods to modern neural networks and pre-trained models (such as RNN, Transformer, BERT, GPT, etc.), and discusses the current status and problems of these technologies in text understanding, generation and cross-modal applications. The findings indicate that although natural language processing has shifted from "understanding" to "generation" and has gradually achieved "cross-modal intelligence," problems such as model illusion, bias propagation, high energy consumption, poor interpretability, and data quality constraints remain current bottlenecks. Future research should focus on improving the reliability and security of models, optimizing resource utilization, designing interpretable artificial intelligence systems, and exploring knowledge enhancement and dynamic update mechanisms.

Keywords: Natural Language Processing, Interpretability, Pre-trained Models, Text Understanding, Text Generation

1. Introduction

Natural Language Processing (NLP) has been an important research direction in the field of artificial intelligence since its birth in the 1950s, with the goal of enabling machines to understand, generate, and interact with human language [1,2]. Early linguistic rule-based methods were highly interpretable but complex to maintain; subsequent statistical models improved the robustness and automation of the system; deep learning has driven a revolution in language representation learning, enabling NLP to enter a new stage and achieve significant breakthroughs in semantic understanding and generation. While the rapid development of NLP has yielded significant results, it has also exposed many new challenges. While the performance of deep models has improved, the huge computational and data consumption limits their universality and sustainability [3]. Problems such as insufficient interpretability of the model, hallucination generation, data bias and ethical risks have not yet been systematically resolved [4]. Besides, the lack of cross-modal, multilingual, and domain-adaptive capabilities also limits the widespread application of NLP in complex real-world scenarios. These issues suggest that, even with advancements in model performance and broader applications, substantial research gaps persist in aspects like reliability, fairness, and controllability. This paper reviews the development history of natural language processing technology, focusing on the technological evolution logic from rule systems and statistical models to deep learning and pre-training paradigms. It also summarizes the representative achievements and shortcomings in text understanding, text generation and cross-modal applications, and discusses the key challenges and future development directions of current research. Through literature review and comparative analysis, this paper reviews representative models such as RNN, Transformer, BERT, and GPT, along with their applications. It summarizes the characteristics, evolution, and limitations of each stage of technology development, revealing the overall trends and underlying logic of NLP research. It is hoped that these efforts will provide valuable insights and suggestions for the development of natural language processing technology, promoting its continuous optimization and progress.

2. Development of Natural Language Processing Methods and Models

2.1 Rule-Based Methods and Statistical Models

In the early stages of the development of natural language processing, it mainly relied on rule systems, and later gradually shifted to statistical models [3]. This process actually reflects that our way of acquiring language knowledge has changed from manual encoding to data-driven learning, and has also made great progress in the stability and scalability of models.

In the rule-based approach era (1950s-1980s), formal grammars and expert rules were primarily used to construct the system. By specifying grammar rules and lexical patterns, parsing systems can be developed. The advantage of doing so is that the model is easy to understand and the theoretical basis is solid [2]. However, while this method works well for some structured linguistic phenomena, it falls short when dealing with ambiguity, polysemy, and variations in language. As language phenomena become increasingly complex and rules become more numerous, the maintenance costs of the system also rise sharply, which greatly limits its application in a wider range of fields.This approach, which relies on explicit knowledge representation, has always presented a contradiction between controllability and universality. In the 1990s-2000s, the statistical modeling phase marked a shift in natural language processing research from rule-based approaches to data-oriented learning methods. Probabilistic models, conditional random fields, and various machine learning algorithms have become the main tools. Models can learn the statistical features of language through a large amount of corpus, improving generalization ability and application performance [3]. This approach notably reduces reliance on manual rules and makes the model more adaptable to noise and changes. However, statistical models still require meticulous feature engineering, and their expressive power is limited by manually designed feature spaces, making it difficult to fully capture deep semantic and contextual information. It can be observed that rule-based methods place more emphasis on theory and interpretability, relying on explicit linguistic knowledge; while statistical models are centered on data and probability, mastering linguistic rules through experiential learning. This shift not only laid the foundation for deep learning, but also showed us the ongoing balance that natural language processing needs to strike between controllability, automation, and performance.

2.2 Deep Learning and Model Development

The introduction of deep learning has transformed natural language processing from statistical learning that relies on manually designed features to end-to-end representation learning. This marks both a technological progress and an enhancement in model capability, evolving from local feature extraction to semantic-level understanding. Specifically, deep learning enables models to better understand and generate natural language by automatically learning

ISSN 2959-6157

complex patterns in data.

In the 2010s, deep neural networks emerged as a dominant force in NLP, leading to substantial performance improvements. By capturing sequential dependencies, RNNs and their LSTM and GRU variants have driven notable advancements in language modeling and machine translation [7]. Word embedding technology uses continuous vectors to represent semantics, laying a solid foundation for deep semantic modeling. The model at this stage employs an end-to-end training mechanism, enabling it to automatically learn feature representations and significantly reducing the workload of manually designing features. However, RNN structures still suffer from gradient vanishing and computational inefficiency when dealing with long-range dependencies, which limits their application in large-scale tasks to some extent. Since 2018, the emergence of the Transformer architecture has completely changed the structure and training method of NLP models [4]. By processing sequences in parallel and effectively modeling global dependencies, self-attention-based models improve their representational power and training efficiency. The "pre-training-fine-tuning" model, represented by BERT [5] and GPT [6], has gradually become the mainstream, enabling the model to learn language representations on general corpora and quickly adapt to specific tasks with a small amount of task data. However, as the capabilities of models expand, problems such as high computational costs, energy consumption, and insufficient interpretability have gradually become prominent. This reflects that while pursuing high performance, deep learning methods must also face a trade-off between resource consumption and performance. Deep learning has propelled NLP models from shallow statistical modeling to a new stage of semantic understanding and generation, significantly enhancing their generalization and transfer capabilities in open domains. However, model complexity, data dependency, and controllability remain core challenges.

3. Task Classification and Cross-Modal Applications of Natural Language Processing

3.1 Text Comprehension Tasks

Text understanding in NLP enables machines to extract meaning from unstructured text to support tasks like decision-making, content generation, and analysis. These tasks do not generate new text directly; instead, they extract and summarize information from existing text, forming the foundation of machine language comprehension, and moving from "perceiving" to "understanding."

The basic tasks involve preliminary text processing steps like word segmentation, part-of-speech tagging, and syntactic analysis. These steps primarily convert the raw text into structured linguistic units, thus establishing a framework for further text analysis [8]. The core semantic understanding task is to delve into the meaning of the text, including named entity recognition, relation extraction, temporal extraction and semantic role labeling. These tasks grasp the key to text understanding [9]. Document-level and sentiment understanding tasks have been further expanded to include text classification, sentiment analysis and text summarization, which not only focus on the content of the text, but also involve the interpretation of subjective information, moving from the sentence level to the discourse level [10]. Advanced understanding and reasoning tasks are cutting-edge fields of text understanding, requiring models to have knowledge, logic and reasoning abilities, and are mainly divided into question answering systems and natural language reasoning [5]. Text understanding tasks progress from simple to complex, providing essential foundational capabilities for almost all advanced NLP applications. The emergence of pre-trained language models has brought significant improvements to these tasks, as they provide models with deeper semantic representations and broader knowledge bases.

3.2 Text Generation Tasks

Text generation tasks differ from text understanding tasks in that their main goal is to enable machines to automatically generate new text based on specific requirements or instructions. This is a crucial step in natural language processing from "understanding" to "creation," and it is also the core of current AIGC. Text generation is essentially conditional creation, with input conditions and core technologies forming a rich technological ecosystem. Its core technologies have evolved from early template-based and statistical generation methods to modern neural text generation, relying on the Seq2Seq framework and Transformer architecture to achieve efficient contextual modeling through autoregressive generation and attention mechanisms. Today, the mainstream approach integrates pre-trained models with few-shot learning [4-6].

Text-to-text generation tasks convert text from one language to another, as seen in online translation services. Data-to-text generation tasks transform structured data into natural language descriptions, such as weather forecasts or sports reports. Free-creation and AIGC tasks generate stories, poems, or articles based on user-provided creative instructions, while dialogue generation tasks produce natural and coherent conversational content for applications such as chatbots and AI assistants. These

task types are widely used in fields such as online translation, content creation, and AI assistants. However, the illusion problem may cause the generated text to contain information inconsistent with the input conditions [15]; insufficient controllability makes it difficult to adjust the generated results precisely; lack of long-term planning may cause the generated long text to lack coherence; lagging evaluation indicators make it difficult to accurately assess the quality of the generated content; bias and security issues may lead to the generated content being discriminatory or unsafe. These factors collectively limit the reliability of text generation systems. By leveraging retrieval enhancement generation technology, the accuracy and richness of generated content can be improved by combining external knowledge bases; multimodal fusion technology can combine multimodal information such as images and audio to improve the diversity and relevance of generated content; value alignment technology can ensure that generated content conforms to human values and ethical standards [12].

3.3 Cross-Modal Applications

The primary goal of cross-modal learning and research is to equip machines with the ability to interpret and produce information across multiple modalities, similar to human capability, and to facilitate the exchange and mutual enhancement of information between these modalities. This breaks down the barriers between different modes in traditional AI tasks and is a key pathway to more general AI. The key lies in solving the modality gap and alignment problem, relying on technologies such as pre-trained models, contrastive learning and Transformer architecture [4]. For example, CLIP achieves image-text alignment via contrastive learning, effectively matching text descriptions with image content, thereby supporting efficient image-text retrieval [11]; while DALL·E generates high-quality images based on text and is widely used in creative design and content generation. These technologies demonstrate the great potential of cross-modal applications. At present, the main directions of cross-modal applications include image and text understanding and retrieval, text-to-image generation, audio-related applications, and video-related applications. Among them, image and text understanding and retrieval aim to achieve efficient matching and semantic understanding between text and images; text-to-image generation generates realistic images based on text descriptions and is widely used in creative design and content production; audio-related applications integrate multimodal information such as audio and text, supporting speech recognition, audio classification, and multimodal interaction; and video-related applications combine video frames with text descriptions to achieve the understanding, generation, and retrieval of video content, providing support for intelligent video editing and content recommendation. However, the current model still faces several challenges. It lacks fine-grained understanding in complex scenarios, has limited common-sense reasoning ability, and struggles to provide necessary background knowledge support; furthermore, illusion problems and security challenges also affect its reliability in practical applications. Future research directions include: developing multimodal large models, optimizing cross-modal alignment techniques, enhancing common-sense reasoning ability, and improving the reliability and fairness of generated content through security audits and dynamic knowledge updates.

4. Challenges and Future Directions

4.1 Reliability and Security of Generation

Large Language Models (LLMs) are still susceptible to illusion problems when generating text, which may produce content that does not conform to the facts, like fictitious information, misquotes, and logically inconsistent statements. The social biases learned by the model during training may be unintentionally amplified, resulting in unfairness in outputs in terms of gender, race, or social group [13]. Malicious input may also lead to the model generating harmful, illegal or violent content. These deficiencies in reliability and security limit the application of the model in high-risk fields such as finance, healthcare and law, and also increase social and ethical risks [15]. To address these issues, retrieval-enhanced generation techniques are combined to verify the consistency between the model output and the external knowledge base, while also improving pre-training strategies and alignment techniques to reduce bias [12]. Besides, red team testing, content filtering, and security audits can further mitigate potential risks. Future research also needs to explore dynamic knowledge updates and real-time verification mechanisms to continuously improve the reliability and security of generated content.

4.2 Resource Consumption and Interpretability Constraints

The training cost of modern large-scale models is very high, often requiring tens of millions to even hundreds of millions of dollars, and consumes substantial amounts of energy. This not only limits access for small and medium-sized organizations and academic groups but also exacerbates carbon emissions and environmental pressures

ISSN 2959-6157

[14]. Furthermore, the complexity of the model makes the decision-making process difficult to interpret, the source of errors difficult to trace, and the division of responsibility unclear. And this black-box characteristic limits the application of the model in scenarios requiring a high degree of controllability, like law, healthcare, and finance. T he demand for high computing resources also affects the iteration speed and innovation capability of models, and the testing and application of new technologies face high barriers. To address these challenges, researchers employ model compression, pruning, quantization, and distillation techniques to reduce training and inference costs. They also develop explainable AI research by exploring the internal decision-making mechanisms of models through feature visualization, probe analysis, and circuit analysis, thus improving traceability and reliability. Lightweight solutions such as sparse activation networks to reduce computational load and using distributed and mixed precision training are being developed to create more efficient model architectures that maintain performance while reducing energy consumption and environmental impact.

4.3 Data Knowledge Limitations and Bottlenecks

The limitations and bottlenecks of data and knowledge are major challenges facing current model development. And model performance is highly dependent on high-quality, comprehensive data and knowledge, but real-world data often contains noise, is scarce, or is biased. This not only affects the stability of models in complex tasks but also limits their generalization ability in cross-domain applications and long-tail tasks. Incomplete knowledge coverage can make it difficult for models to understand domain-specific terminology, professional concepts, or cultural context, increasing the risk of generating inaccurate, ambiguous, or irrelevant content. Meanwhile, inconsistencies between different data sources, differences in annotation, and lag in corpus updates may further weaken the model training effect and output reliability. To address these issues, relevant studies have adopted synthetic data generation, data cleaning, denoising and management techniques to boost data quality, and combined knowledge graphs or domain knowledge bases to expand the scope of knowledge and strengthen cross-domain understanding [12]. By utilizing multi-source data fusion, transfer learning, and minimally supervised learning, the model's adaptability in low-resource scenarios is enhanced, supporting long-tail and low-resource domains and boosting generalization performance.

5. Conclusion

This paper finds that natural language processing has gone through four stages of development since its birth in the 1950s, from rule-based methods to pre-trained models. This process shifts from explicit encoding relying on manual rules to data-driven end-to-end semantic models, achieving a transformation from "understanding" to "generation." The Transformer architecture, along with the BERT/GPT pre-trained model paradigm, has greatly enhanced performance in text understanding, generation, and cross-modal applications, enabling the widespread adoption of NLP in question answering, dialogue generation, image-text retrieval, and AIGC. However, NLP still faces many challenges. The reliability of generated content is insufficient, resource consumption is too high, the model decision-making process lacks interpretability, data quality and quantity are bottlenecks, security risks are high, and ethical issues are also prominent. These issues limit the application of NLP in some key areas. Future research needs to focus on technology optimization, improving interpretability, data innovation, building secure frameworks, and sustainable computing to enhance the performance and reliability of NLP and expand its application scope.

References

[1] Turing, A. M. (2021). Computing machinery and intelligence (1950). Mind, 59(236), 33-60.

[2] Chomsky, N. (2002). Syntactic structures. Walter de Gruyter.

[3] Lafferty, J., McCallum, A., & Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML, 282-289.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[5] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[6] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[7] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[8] Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. Using Large Corpora, 273, 31.

[9] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.
[10] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.
[11] Radford, A., Kim, J. W., Hallacy, C., et al. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.
[12] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
[13] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
[14] Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 09, pp. 13693-13696).
[15] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.