Dean&Francis

ISSN 2959-6157

Natural Language Processing Models and Applications

Yawen Zhang

Department of Computer Science and Technology, Fuzhou University, Fujian, China
*Corresponding author: 102401503@ fzu.edu.cn

Abstract:

This paper reviews the technological evolution of Natural Language Processing (NLP), tracing its development from traditional rule-based and statistical methods to modern deep learning paradigms. It particularly emphasizes the profound impact of neural models. By systematically examining NLP's significantly enhanced capabilities in understanding, generating, and integrating natural language, this study aims to comprehensively analyze the practical value of this technology across diverse application domains. The paper first elaborates on the limitations of rule-based and statistical models, then demonstrates the emerging capabilities of large-scale language models in cross-modal data fusion through applications such as medical text mining and data generation. Finally, it highlights current challenges faced by NLP, including robustness, computational resource consumption, and ethical biases, while proposing that lightweight models and multimodal unified frameworks represent critical future research directions. This review not only underscores the transformative potential of NLP across various fields but also provides insights into future research trajectories, encouraging the development of more efficient, interpretable, and ethically responsible language technologies.

Keywords: NLP; Artificial Intelligence; Applications.

1. Introduction

Natural Language Processing, as a core discipline of artificial intelligence, has long been dedicated to achieving machines' profound understanding and generation of human language. Early research primarily relied on rule-based systems and statistical models. While progress was made in specific tasks, limitations emerged due to insufficient generalization capability and high manual annotation costs. In 1950, Alan Turing proposed the „Turing Test"; early researchers developed simple dialogue programs using hand-crafted grammatical rules and dictionaries. However, due to challenges such as linguistic ambiguity and limited computational power, NLP entered its first winter in the 1970s. With the rise of the internet bringing massive textual data, probability-based statistical methods became mainstream. During this period, NLP paradigms began shifting from rule-based approaches to data-driven methodologies. In 2013, Word2Vec word embedding technology demonstrated that word semantics could be computationally represented. Recurrent Neural Networks and their variants—LSTM and GRU—demonstrated outstanding performance in machine translation and text generation tasks. Currently, large language models represented by ChatGPT exhibit human-like conversational capabilities, marking NLP's entry into a new phase of exploring Artificial General Intelligence (AGI).

The significance of this evolution lies not only in technological progress but also in its profound impact on human-computer interaction, knowledge acquisition, and social development. Natural language processing has evolved from a tool limited to specific domains into a general intelligent technology capable of supporting diverse applications—such as real-time translation, intelligent customer service, and automated content generation—significantly improving efficiency across various industries and breaking down language barriers. More importantly, the emergence of large language models has opened up new possibilities for achieving Artificial General Intelligence (AGI), prompting people to rethink in-depth issues such as machine cognition, language understanding, and even the nature of intelligence. At the same time, natural language processing technology has also raised important ethical and social issues such as data bias, privacy protection, and responsibility attribution, highlighting the necessity of integrating technology and humanities in future development.

2. The Development of NLP

Traditional natural language processing models can be formally divided into two subtypes: rule-based and statistical. Each of these subtypes further consists of several specific models, each with its own characteristics and applicable scope.

2.1 Rule-Based Stage

In the early stages of natural language processing, researchers formulated a series of standard templates based on the tasks that needed processing [1]. The processed corpus was compared against these standard templates, and the template with the highest similarity was selected to represent the meaning of the corpus. This model was called the template method. In 1956, drawing on Shannon's work, people used finite-state automata as a tool to describe grammar, establishing a mathematical model of language. Regardless of the model, significant defects existed. Consequently, after a period of popularity, rule-based natural language processing gradually fell out of favor among scholars, who turned their attention to statistical-based natural language processing, proposing various models and gradually achieving substantial results.

2.2 Statistical-Based Stage

The Support Vector Machine (SVM) model is used to separate samples with a maximum margin [2]. It possesses a solid mathematical theory and strong generalization ability [3], making it suitable for classification problems with small sample sizes. However, when facing large-scale, high-dimensional corpora, this method places high demands on machine performance and has low resistance to noise, leading to poor model robustness. The Maximum Entropy model was used for estimating the power spectral density of signals, and later introduced into natural language processing. The Maximum Entropy principle holds that when learning a probability model, among all possible probability model distributions, the model with the highest entropy is the best. Please refer to the following Table 1.

Table 1. Traditional Natural Language Processing Models [4]

Category	Specific Model	Launch Date	Advantage	Disadvantage
Rule-Based	Template Method	—	Simple Theory	Poor robustness, insufficient generalization ability, and a lack of unified template standards
	Automata Theory	1956	Has a perfect theory and performs well on closed corpora	Poor robustness, insufficient generalization ability, and low practicality
Statistics-Based	Support Vector Machine	1964	Sound theory, capable of solving high-dimensional sparse data problems	Weak noise resistance, low stability, and strict hardware requirements in traditional models

ISSN 2959-6157

	Maximum Entropy Model	1967	Flexibly set constraints to adjust the model's fitting degree to the corpus.	Huge computational cost, difficult to apply in practice.
	Decision Tree	1975	Simple and practical, easy to understand and implement, facilitating processing and visualization	Sensitive to data, neglects relationships between attributes, and prone to overfitting

3. Application Areas of NLP

3.1 Data Generation Field

Large language models possess in-context learning capabilities, thus, they can be used to generate training datasets for smaller models in specific domains. DAI et al. attempted to use ChatGPT to solve model training problems in few-shot scenarios [5]. They designed appropriate prompts to make ChatGPT generate augmented data semantically similar to the original data. Subsequently, they used the original data combined with the augmented data to train a downstream BERT model [6]. Experimental results showed that training the model with augmented data generated by ChatGPT yielded better results compared to using data generated by traditional augmentation methods like back-translation or word replacement.

3.2 Artificial Intelligence Field

Large models have wide applications in the field of natural language processing, which can be further divided into four specific applications: text classification, named entity recognition, sentiment analysis, and machine translation. Among these, text classification is the foundation of text information processing and the basis for large models to generate natural language. The algorithmic logic of large AI models in processing natural language is similar to human learning: by continuously absorbing knowledge to understand the probability of various events, they master probabilistic judgment skills, thereby serving the user.

The application of large models in the field of natural language processing demonstrates significant advantages[7]. Their powerful representational ability allows models to better understand and generate natural language, thereby achieving higher accuracy. Simultaneously, the application scope of large models is very broad, covering text classification, sentiment analysis, question-answering systems, and more. However, issues such as high cost, insufficient privacy, and security still exist during use. Therefore, corresponding measures should be taken to reduce costs, improve computational efficiency, and strengthen data privacy and security protection to ensure that large models can play a greater role in the field of natural language processing.

3.3 Healthcare Field

The core applications of NLP in the healthcare sector focus on the in-depth mining and analysis of medical textual data to enable intelligent assistance [8]. By extracting key information from electronic health records and converting it into standardized codes, NLP enhances diagnostic and treatment efficiency. It supports physicians in symptom analysis, generates diagnostic suggestions, and provides evidence-based decision support drawing on medical literature. NLP also accelerates drug development processes, including detecting adverse drug reactions, matching patients to clinical trials, and exploring new uses for existing drugs.[9] Furthermore, intelligent consultation robots and patient health advisory systems improve doctor-patient communication and health management. Overall, NLP technology is driving a transformation toward data-driven, efficient, and precise healthcare, significantly improving the efficiency, accuracy, and safety of medical services.

3.4 Geological Domain

Geological research focuses on the lithosphere and its adjacent spheres, which are not directly visible [10]. It involves the study of stratigraphic structures, sedimentary evolution, minerals and rocks, geophysics, geochemistry, laboratory testing, analysis, and computer-based simulations. At the same time, geological structures are often shaped by multiple phases of tectonic activity, resulting in diverse regional geological characteristics with significant variations. Consequently, geological data often exhibit multi-source and multi-interpretation features, while outcomes are typically presented in a mix of unstructured, semi-structured, and structured forms. These complexities pose challenges for further research and information extraction.

With the support of NLP technology, quantitative analysis, development, and utilization of geological texts can be conducted in a scientific and feasible manner. Geologists will have access to domain-specific NLP models and techniques, enabling the construction of more scientific, systematic, and reliable geological knowledge graphs. This will facilitate deeper integration of geological data—including texts, images, audio, and video—with artificial intelligence. Ultimately, it will allow for the fully automated, scientific, and standardized generation of geological reports, one-click 3D visualization of geological texts and knowledge, streamlined interaction between computers and geological experts, and improved support for industry and government decision-making.

4. Conclusion

Natural language processing has made significant progress, has been deployed in industrial applications in many fields, and has demonstrated certain market value and potential. However, natural language processing still faces many bottlenecks. Therefore, this paper summarizes the achievements of natural language processing models and looks into future trends of natural language processing. Firstly, the robustness issue remains to be solved. In practical applications, spelling errors, grammatical errors, and punctuation issues contained in text data may lead to performance degradation of NLP models. Developing NLP models that can adapt to various extreme conditions in real application scenarios is a challenge. Solving these problems will help improve the practicality and reliability of natural language processing technology. Meanwhile, NLP also faces the problem of large model parameter counts. Existing large models require substantial computational resources for training and inference. Therefore, running these models may require expensive hardware infrastructure. Models require significant time, increasing the time cost for researchers and engineers in model development and optimization. Thus, developing more efficient and lightweight NLP models is necessary to make NLP technology more sustainable and widely applicable. Furthermore, techniques such as transfer learning, model compression, and quantization should be employed to reduce the computational and storage demands of large NLP models, making them easier to use in different environments.

References

[1] Jiang L, Tang H L, Chen Y J. A survey of transformer-based natural language processing research. Modern Computer, 2024, 30(14): 31–35.

[2] He X F, Zhou J, Chen D G, Liao H. A review of deep learning models in natural language processing. Computer Applications and Software, 2025, 42(02): 1–19+101.

[3] Xu D L, Lin M, Wang Y R, Zhang S J. A review of NLP data augmentation methods based on large language models. Journal of Computer Science and Exploration, 2025, 19(06): 1395–1413.

[4] Wei X L, Sun X. Research progress and development trends in natural language processing techniques. China-Arab Science and Technology Forum (Chinese and English), 2025, (05): 84–88.

[5] Zhu H H, Liang F, Jiang J P. A review and outlook of natural language processing research in international library and information science: A visual analysis based on CiteSpace. Journal of Guilin University of Aerospace Technology, 2025, 30(02): 322–333.

[6] Shen L, Yin Y N. Research on translation strategies of academic works based on NLP technology. Shanghai Journal of Translators, 2025, (03): 56–62.

[7] Wu D. Methods and challenges of large models in NLP benchmarking. Journal of Liming Vocational University, 2025, (02): 85–92.

[8] Zhang Y, Nie Y M. A survey of large language models and their application prospects in security. Intelligent Security, 2023, 2(04): 100–112.

[9] Zhang J Y, Li X Q. A review of the application of deep learning in natural language processing. Computer Knowledge and Technology, 2025, 21(19): 23–25.

[10] Xu R, Li Z X, Zhang P Q, Yan J B. Applications and prospects of natural language processing technology in the field of geology. Energy Technology and Management, 2025, 50(04): 151–154.