Challenges and Application Models of Natural Language Processing

Weiqi Huang[a]
Business School, GuangZhou Nanfang Collage, 882 Hot Spring Avenue, Conghua District, Guangzhou, China

Keywords: Natural Language Processing, Natural Language Understanding, Natural Language Generation, Natural Language Processing Model.

Abstract: With the development of artificial intelligence, natural language processing has become an important research field of human-computer interaction, and its importance has become increasingly prominent. This paper outlines three major challenges facing natural language processing today: First, there are a large number of ambiguous words in natural language; Second, natural language processing is highly dependent on contextual information. Third, differences between different languages introduce additional complexity to processing. By sorting out the challenges faced, it provides new ideas for the future research direction. Next, it introduces the classification of natural language applications (natural language understanding and text generation) and the actual realistic scenarios applied to it, reflecting the application of natural language processing in silence to help and affect people's lives. Finally, the paper discusses three major models (Transformer model, Bert model, GPT model) which play an indispensable role in promoting the progress of natural language processing technology. These models show excellent processing power in a multitude of natural language tasks.

1 INTRODUCTION

In the contemporary digital era, the interaction between human and machine is becoming more and more frequent, and natural language processing (NLP), as a key technology connecting human language and computer system, is gradually becoming one of the most dynamic and influential research directions within the domain of artificial intelligence. With the popularization of the Internet and the development of big data technology, massive text data continues to emerge, which contains rich information and knowledge, but also brings huge challenges. How to process, analyze and understand these text data effectively and transform them into valuable information has become the focus of scientific and technological circles and academic circles. NLP technology is the key to addressing this challenge. The objective is to enhance the capacity of computers to understand, interpret, and generate human language, thereby promoting more natural and seamless interactions with humans.

The development of natural language processing is full of challenges and opportunities. From early rule-following methods to modern deep learning-based models, the range of applications is increasingly wide. Nowadays, NLP technology has penetrated into all aspects of our lives, from intelligent assistants, machine translation, sentiment analysis, to text mining, automatic summary, question and answer system, etc. The application of NLP not only improves the efficiency of information processing, but also brings great convenience to people's life and work. However, despite the remarkable progress, natural language processing still faces many challenges, such as language ambiguity, context understanding, and multilingual processing, which limit the further innovation and utilization of NLP technology.

As technology evolves and becomes more advanced, people are generating more and more data, which provides a huge opportunity to enhance the training of natural language models. From the early classic Transformer model (Vaswani, Shazeer, Parmar et al, 2017), which uses about 100 million parameters, to the current GPT model (Radford, Narasimhan, Salimans et al, 2018) with the escalation in parameters numbers, the model's capabilities are constantly improved to achieve more natural human-computer interaction.

[a] https://orcid.org/0009-0006-9968-3658

Huang, W.
Challenges and Application Models of Natural Language Processing.
DOI: 10.5220/0013704100004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 673-678
ISBN: 978-989-758-765-8

Through this review, the paper hope to provide readers with a comprehensive and systematic overview of natural language processing, and help readers better understand the current situation and future development direction of this field.

2 OVERVIEW OF NATURAL LANGUAGE PROCESSING

NLP serves as a bridge between multiple disciplines, focusing on how humans and machines can communicate effectively using natural language. It aims to equip machines with the ability to comprehend, decipher, and generate human-like written or spoken expressions. By leveraging advanced computational techniques and linguistic theories, NLP enables machines to analyze, understand, and interact with human language in a way that conveys meaningful response and contextually appropriate (Das, 2024). NLP seeks to equip with the ability to understand and manipulate human language through various computational models that automatically analyze linguistic structures. In NLP, the foundational elements of language are typically referred to as atomic terms, like "bad," "old," or "fantastic." When these atomic terms are combined, they create compound terms, such as "very good movie" or "young man." At its core, an atomic term refers to a single word, whereas a compound term refers to a multi-word phrase. Words serve as the basic units of human language, and their comprehension is essential for any NLP task (Satpute, 2023). At present, NLP has been applied to many fields, such as: speech recognition, question answering system, online translation, text classification and so on NLP has played an important role.

2.1 Challenges in natural language processing

Trying to make machines understand human language is very challenging. The machine not only understands the surface meaning of the sentence, but also understands the underlying information contained in the sentence. The following are the difficulties that machines face in NLP.

First, there are a large number of words with multiple meanings in natural language processing. For example, in the sentence "I want an apple", the "apple" that a person wants to express refers to the "apple phone", but the machine may understand it as a fruit, that is, the meaning of a word will have varying interpretations in different contexts. This illustrates the ambiguity inherent in language expression. To address this issue, artificial intelligence techniques, such as machine learning and deep learning models are employed to identify and resolve textual ambiguities. Moreover, parameters in the model can be adjusted to better adapt to different language environments and contexts, thus improving the accuracy and effectiveness of ambiguity detection (Satpute, 2023).

Second, natural language processing relies on understanding context, but the processing power of computational models is limited. One reason for this is the amount and quality of data. Although the data sets used by modern NLP models are already very large, they may not perform well when dealing with rare or domain-specific processes. For example, a sentiment analysis model may perform well when dealing with common movie reviews, but poorly when faced with rare, domain-specific reviews, such as professional cinematography reviews. The reason for this long-tail phenomenon is that many words and expressions appear infrequently in natural language, and these long-tail data may be ignored in the training set, resulting in poor performance of NLP models in dealing with these rare cases. Another reason is limited computing resources. The NLP model requires significant memory and computing power resources to process millions of pieces of text. Therefore, when we use a computer with insufficient computing power, it takes a very long time to process the context (Ardehkhani, 2023). To solve this problem, the number of Gpus can be increased to reduce processing time, but this approach will lead to higher training costs (Strubell, 2019).

Third, there are linguistic differences in natural languages. In today's Internet society, A robust NLP model that can handle language diversity is important (Mitra, 2020). For some small languages, the data set that can be used for training is very small, which brings a lot of challenges to model training. Owing to the scarcity of large-scale training data, models generally exhibit poorer performance in smaller languages compared to larger ones. Moreover, the primary distinctions between languages are manifested in their grammatical differences. For example, an adjective in French usually comes after a noun, while an adjective comes before a noun in English. In text generation, if the machine does not understand these grammar rules correctly, the generated text will be meaningless or difficult to understand (Nadkarni, 2011).

2.2 The applied branch of natural language processing

NLP primarily consists of two key branches. One of these is Natural Language Understanding (NLU), which refers to the process of understanding and inter preting human natural language by computer. This represents a crucial research avenue within the realm of artificial intelligence, with the goal of endowing computer with the capability to understand and process human language, thereby facilitating natural and seamless interactions with humans (Guo, 2024). Another branch is Natural Language Generation (NLG), whose primary task is to generate human-readable text from structured and unstructured data to provide feedback in a way that is easy for humans to understand (Azhar, 2024). As shown in Figure 1, this is the main branch module for NLP applications.

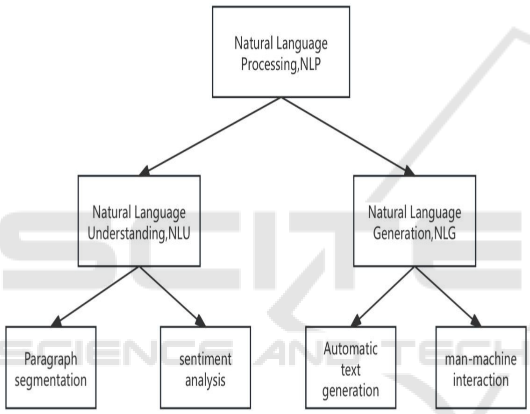

Figure 1: The main branch module of NLP applications (Picture credit : Original)

NLU is mainly used in text segmentation and sentiment analysis. Text segmentation refers to the prediction of reliable paragraph boundaries based on the same topic belonging of the sentences in the same paragraph for an article containing multiple sentences (Zhao, 2024). This segmentation process helps organize data into logically coherent units and is essential for improving the readability of text. Emotion analysis is to analyze the words and expressions in the text to judge the emotional tendency conveyed by the text, such as positive, negative or neutral. Businesses leverage sentiment analysis to discern what user reviews indicate about their goods or services.

NLG is mainly used for automated text generation and human-computer interaction. In terms of automated text generation, NLG helps individuals efficiently generate various types of text content by converting structured data into natural language text. For example, in the field of news reporting, NLG is able to quickly generate accurate and timely news content, reducing the workload of human editors. In terms of human-machine interaction, NLG technology enables the machine to generate natural and accurate replies, enhancing the user interaction experience. For example, artificial intelligence such as Siri and Google Assistant use NLG technology to provide users with accurate interactive questions and answers.

3 APPLICATION MODEL OF NATURAL LANGUAGE

The development of NLP cannot be separated from several important models. In 2017, Google introduced the Transformer model for the first time, which relies on the self-attention mechanism (Vaswani, Shazeer, Parmar et al, 2017). has laid an important foundation for the development of NLP and exerted a profound influence. Next, in 2018, Google introduced the Bert model (Devlin, Chang, Lee et al, 2018) and OpenAI also developed the GPT-1 model (Radford, Narasimhan, Salimans et al, 2018). Both models incorporate pre-training modules on top of the Transformer model.

3.1 Transformer model

The Transformer model was initially used in natural language translation. Compared with previous mainstream models LSTM and GRU, Transformer has two obvious advantages: First, Transformer uses multi-head attention mechanism model and can use distributed CPU for parallel training to improve model training efficiency and accuracy (Lei, 2024); Second, traditional RNNS and LSTMS tend to perform poorly at capturing long-term dependencies because these models need to process sequence data progressively, leading to gradient disappearance or gradient explosion problems. With its self-attention mechanism, the Transformer model can directly calculate the dependencies between any two locations in the sequence, making it more efficient to capture long-term dependencies. The Transformer model is mainly composed of input part, N-layer encoder part, N-layer decoder part and output part. The complete flow of Transformer is shown in Figure 2.

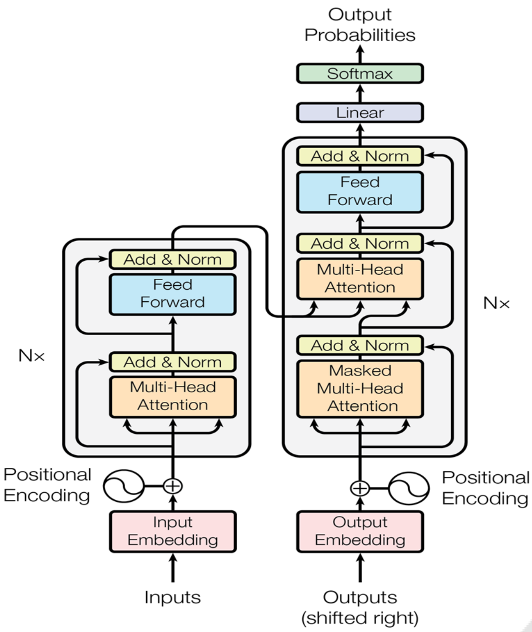

Figure 2: Complete process of Transformer model (Vaswani, Shazeer, Parmar et al, 2017).

Among them, the multi-head attention mechanism in the decoder is the cornerstone of the Transformer model (Vaswani, Shazeer, Parmar et al, 2017). The attention mechanism multiplies the attention weight by the QK matrix and multiplies the weight value with V to get the final attention result. Figure 3 shows the calculation process of attention results. Multi-head refers to the parallel feature extraction of multiple single-head self-attention mechanisms. By searching for parameters in multiple parameter Spaces, the accuracy of the model will naturally increase as the parameters extracted increase. However, the shortcoming of Transformer model is that it can only be applied to the data analysis of short sentences, because it cannot capture the position relationship between words for analysis (Li, 2021).

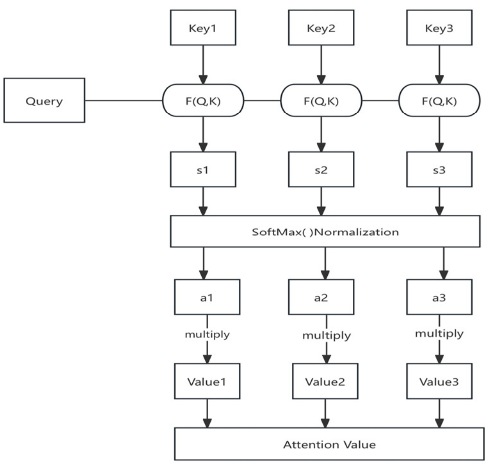

Figure 3: The process of calculating the outcome of attention (Picture credit : Original).

The Transformer model uses A training set : The standard WMT 2014 English-German dataset contains 45,000 sentence pairs. The sentence preprocessing adopts Byte-pair encoding to make the dictionary of the training set smaller and get a dictionary composed of 37000 tokens. The devices used include 8 NVIDIA P100 GPUs. The Evaluation criteria used are: BLEU (Bilingual Evaluation Understudy) is a method to evaluate the quality of machine translation, in particular to measure the similarity between machine translation output and human translation. In particular, it evaluates the overlap of n-grams, which are contiguous sequences of n items from a given text sample, between the translated output and the reference translation. Higher BLEU scores denote a higher level of similarity, suggesting that the machine-generated translation is more comparable to human translation quality. This metric is particularly valuable in evaluating the fluency and accuracy of translation models, making it a cornerstone in the field of machine translation research and development. Specifically, it measures the overlap of n-grams (contiguous sequences of n items from a given sample of text) between the translated output and the reference translation. And the other training set is: WMT 2014 English-French dataset, which is a large-scale corpus, comprises 36 million sentences and segments the tokens into 32,000 distinct phrases. The hardware utilized includes eight NVIDIA P100 GPUs.

3.2 Bert model based on Transformer model

The Bert model uses the Encoder part of Transformer (Devlin, Chang, Lee et al, 2018). The biggest difference from the Transformer Model is the introduction of two pre-training tasks: Mask Language Model (MLM) and Next Sentence Prediction(NSP) are two fundamental tasks in pre-training language models. MLM involves randomly concealing certain words in the input text and then training the model to infer these hidden words based on the surrounding context. Meanwhile, the NSP task focuses on training the model to determine if two sentences are adjacent in a coherent text sequence. Another difference is the introduction of bidirectional processing, that is in the pre-training phase. For each word, the Bert model can evaluate both the preceding and subsequent context information. Bert model is widely used in NLU work because of its strong context understanding ability (Kurt, 2023).

3.3 GPT model based on Transformer

GPT uses Transformer's encoder architecture (Radford, 2018) because GPT is primarily used to generate text, and the decoder is designed for generation tasks. GPT is capable of predicting the next word using the previous word, which means the input sentence is directional. The one-way self-attention mechanism is used, which only focuses on the previous words in the sequence, so it has strong ability to generate and interactive question answering. In addition, the biggest feature of GPT is the large amount of parameters, so as to ensure the quality and accuracy of interactive dialogue. The pre-trained GPT model is composed of 12 Transformer layers, and the model dimensions are

$$d_{model}=768 \quad (1)$$

The total number of parameters reached 110 million (Zheng, 2021).

With the continuous increase of model parameters, Google has continuously iterated the GPT model. Launched in 2020, GPT-3 (commonly known as chatGPT) is the most powerful and extensive language model to date (Gupta, 2023), and its output is highly consistent and contextually relevant. GPT-3 ability to learn from a small number of samples is a major advance, quickly grasping new information from small samples. This improvement is due to the increasing number of parameters used in GPT-3 compared to GPT-2 and GPT-1. GPT encompasses 175 billion parameters, and its predecessors GPT-1 and GPT-2 have 117 million and 1.5 billion parameters, respectively.

4 CONCLUSIONS

As an important branch of artificial intelligence, NLP has made remarkable progress in theory and application. NLP has undergone significant evolution, transitioning from early rule-based methods to contemporary models powered by deep learning. This series of advancements has substantially improved the ability of computers to comprehend and generate human language, enabling more sophisticated and natural interactions. Despite these advancements, the field of NLP continues to confront numerous challenges that impede its further development and broader application. In an effort to surmount these obstacles, researchers are persistently exploring novel approaches and techniques, including pre-training language models, multimodal learning, and reinforcement learning, to boost the performance and adaptability of the models. In the future, natural language processing technology will continue to develop rapidly and deeply integrate with other technologies such as computer vision, speech recognition, machine learning, etc., to form a more intelligent and efficient artificial intelligence system. These systems will be able to better understand human language and enable smoother human-computer interaction.

REFERENCES

Ardkhani, P., Vahedi, A., & Aghababa, H. 2023. Challenges in natural language processing and natural language understanding by considering both technical and natural domains. 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA), Qom, Iran, Islamic Republic of, pp. 1-5.

Azhar, U., & Nazir, A. 2024. Exploring the natural language generation: Current trends and research challenges. 2024 International Conference on Engineering & Computing Technologies (ICECT), Islamabad, Pakistan, pp. 1-6.

Chang, M. W., Devlin, J., Lee, K., & Toutanova, K. 2018. Bert: Pretraining of deep bidirectional transformers for language understanding. OpenAI Blog.

Das, S., & Das, D. 2024. Natural language processing (NLP) techniques: Usability in human-computer interactions. 2024 6th International Conference on Natural Language Processing (ICNLP), Xi'an, China, pp. 783-787.

Ganesh, A., Strubell, E., & McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Gupta, N. K., Chaudhary, A., Singh, R., & Singh, R. 2023. ChatGPT: Exploring the capabilities and limitations of a large language model for conversational AI. 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, pp. 139-142.

Guo, D. C. 2024. Research on natural language understanding problems in open-world scenarios (Master’s thesis, Beijing University of Posts and Telecommunications).

Jiang, L., Tang, H. L., & Chen, Y. J. 2024. A review of natural language processing based on the Transformer model. Modern Computer, (14), 31-35.

Kurt, U., & Çayir, A. 2023. A modern Turkish poet: Fine-tuned GPT-2. 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, pp. 01-05.

Li, X., Wang, S., Wang, Z., & Zhu, J. 2021. A review of natural language generation. Computer Applications, 41(05), 1227-1235.

Mitra, A. 2020. Sentiment analysis using machine learning approaches (lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies (UCCT), 2(3), 145-152.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. 2011. Natural language processing: An introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. 2018. Improving language understanding by generative pre-training. OpenAI Blog.

Satpute, R. S., & Agrawal, A. 2023. Machine learning approach for ambiguity detection in social media context. 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, pp. 516-522.

Shazeer, N., Vaswani, A., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Xia, Z., Zhang, C., & Woodland, P. C. 2021. Adapting GPT, GPT-2 and BERT language models for speech recognition. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, pp. 162-168.

Zhao, Y. B., Jiang, F., & Li, P. F. 2024. A multi-level coherent text segmentation method based on BERT. Computer Applications and Software, (10), 262-268+324.

Zheng, X., Zhang, C., & Woodland, P. C. 2021. Adapting GPT, GPT-2 and BERT language models for speech recognition. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, pp. 162-168.