Big Data-based Natural Language Processing and Speech Recognition System

Xuye Wang, Jiahao Zheng, Qihao Dong

Shenyang Aerospace University, Liaoning, China

ABSTRACT

The aim of this paper is to build a big data based natural language processing (NLP) and speech recognition system to meet the increasing demand for text and speech data processing. This system integrates big data technologies and advanced deep learning algorithms to improve the accuracy of text understanding, speech recognition, and has multilingual and multi-accent adaptability. Key features include efficient processing of large-scale data, application of deep learning in NLP and speech recognition, multi-language and multi-accent processing capabilities, and user privacy and data security considerations. With these innovative features, the system aims to provide a high-performance, high-accuracy solution for processing large amounts of text and speech data in the real world, advancing the field of natural language processing and speech recognition.

Keywords: Big Data；Natural Language Processing (NLP)；Speech Recognition；Deep Learning；Adaptability

1 INTRODUCTION

In today's digital age, the rapid growth and diversity of information makes big data processing a central challenge in science and engineering. In this context, the development of Natural Language Processing (NLP) and Speech Recognition technologies has become particularly critical as they are core tools for understanding and processing human language. This research aims to explore and innovate Big Data-driven Natural Language Processing and Speech Recognition systems to address the challenges faced by existing technologies in processing large-scale text and speech data [1].

With the widespread use of social media, online documents, voice assistants, and other communication tools, we are increasingly generating text and speech data of unprecedented size and diversity. Traditional NLP and speech recognition systems often face the problem of decreasing accuracy and efficiency when facing such large and complex datasets [2]. Therefore, this research aims to improve the adaptability, accuracy and processing efficiency of the system by integrating big data technologies to meet the growing demand for information processing.

One of the key objectives of this research is to overcome the effects of linguistic diversity and accent differences on the performance of NLP and speech recognition systems [3]. We introduce deep learning and neural network techniques into the system to better understand the differences between cultures and languages and achieve more accurate processing of multiple languages and accents [4].

Through the use of large-scale datasets, this system is able to achieve high-performance computation when confronted with large amounts of information, thus ensuring that it can still operate efficiently in real-world practical applications [5]. The results of this study are expected to provide an important reference and guidance for the future development of NLP and speech recognition technologies, to promote innovation in this field and to provide society with smarter and more convenient information processing tools.

2 RELEATED WORK

Prior to this study, many scholars and research teams have conducted extensive work in the areas of big data, natural language processing and speech recognition. Prior studies have covered the application of big data techniques to natural language processing. These works have explored how to efficiently process large-scale text data and how to utilise distributed computing and storage resources to improve processing efficiency [6]. Deep learning techniques have made significant progress in the field of natural language processing. Previous research has focussed on the use of deep neural networks to improve the performance of tasks such as semantic understanding, text classification and entity recognition. In the field of speech recognition, researchers have worked on processing large-scale speech datasets to improve the accuracy of speech recognition systems [7]. These works have focused on how to train deep learning models to adapt to different speech features and contexts. In terms of dealing with multilingualism and multiple accents, previous research has emphasised the challenge of building highly adaptable systems that can handle different languages and accents on a global scale [8]. Research on big data processing also includes how to effectively utilise distributed computing and storage systems to meet the challenges of large-scale data, ensuring scalability and high performance of systems.

3 PROBLEM ANALYSIS

When building natural language processing and speech recognition systems based on big data, we face a series of key problems that require careful analysis and solutions. The problem of large-scale data processing, how to effectively process large amounts of text and speech data to ensure that the system can maintain high performance in the face of massive amounts of information [9]. Deep Learning in NLP and Speech Recognition problem, how to use deep learning techniques to improve the accuracy of natural language processing and speech recognition, especially when dealing with complex contexts and multilingual scenarios. Multi-language and multi-accent processing issues, how the system can adapt to different languages and accents to ensure that language processing and speech recognition tasks can be performed efficiently on a global scale [10]. User privacy and data security issues, how to ensure that user privacy is adequately protected when processing large amounts of user-generated text and speech data. Whether the system is sufficiently secure against potential risks of data leakage or misuse [11]. System scalability and performance optimisation issues, how to ensure that the system is scalable in the face of increasing data loads and can scale without degrading performance.

By deeply analysing and solving these problems, we can provide strong support for

building out an efficient, accurate and adaptable big data-driven natural language processing and speech recognition system.

4 COUNTERMEASURE RESEARCH

Countermeasures for large-scale data processing, the use of distributed computing and storage technologies, and the use of cloud computing platforms to achieve efficient data processing and storage. Use techniques such as data compression and indexing to reduce the time and cost of data processing. Deep Learning in NLP and Speech Recognition Countermeasures to select appropriate deep learning model structures, such as Transformer, to improve the ability to learn semantic and speech features [12]. Perform data enhancement and introduce techniques such as adversarial training to improve the generalisation and robustness of the model. Multi-language and multi-accent processing countermeasures to construct a multi-language corpus for model training to ensure that the system can adapt to different languages. For different accents, training for specific accents or introducing transfer learning techniques to improve the recognition of accents. User privacy and data security countermeasures, using encryption technology to ensure the security of user data during transmission and storage. Implement data anonymisation strategies to minimise the handling of sensitive information and follow relevant privacy regulations and standards [13]. Countermeasures for system scalability and performance optimisation, designing a system architecture with good scalability and making full use of distributed computing and storage resources. Use parallel computing and caching techniques to improve system response speed and performance. Real-time processing and streaming data processing countermeasures, introducing real-time processing frameworks, such as Apache Kafka, Apache Flink, etc., to deal with text and speech data generated in real time [14]. Optimisation of algorithms and models for streaming data processing to ensure that the system can maintain efficient performance in a streaming data environment [15]. Quality assessment and model optimisation countermeasures to implement system performance monitoring and quality assessment mechanisms to identify and resolve performance degradation or accuracy issues in a timely manner. Regular optimisation and updating of deep learning models to adapt to changing language and speech features.

The combined use of the above countermeasures can better address the various challenges faced by big data-driven natural language processing and speech recognition systems, and improve the robustness, performance and user experience of the system.

5 CONCLUSION

This research is dedicated to building a big data-based natural language processing and speech recognition system to address the challenges faced when confronted with increasing amounts of text and speech data. Through comprehensive analyses of related work and in-depth study of the problem, we propose a series of countermeasures to address the key issues in system design and implementation.

In terms of large-scale data processing, we emphasise the use of distributed computing and storage techniques, as well as data compression and indexing to ensure that the system can process efficiently in the face of massive amounts of information. In terms of deep learning applications, we suggest choosing appropriate model structures and adopting techniques such as data augmentation and adversarial training to improve the system's ability to learn semantic and phonetic features.

For multilingual and multi-accent processing, we emphasise the construction of multilingual corpora and training for specific accents to ensure that the system can efficiently

perform language processing and speech recognition tasks globally. At the same time, we paid attention to user privacy and data security, and proposed countermeasures such as adopting encryption technology and data anonymisation to ensure that user data is fully protected.

In terms of system scalability and performance optimisation, we propose the adoption of distributed computing and storage systems, and the use of parallel computing and caching techniques to ensure that the system is scalable in the face of increasing data loads, and that it scales without degrading performance.

Overall, by adopting a combination of these countermeasures, we can build a more robust, efficient, and adaptable big data-driven natural language processing and speech recognition system. This will provide effective tools for processing large-scale text and speech data, promote the further development of NLP and speech recognition technologies, and provide smarter and more convenient information processing services for the society.

6 DISCUSSION

In the process of building a natural language processing and speech recognition system based on big data, we need to focus on a series of key issues and discuss the corresponding solutions. The discussion of large-scale data processing, by adopting distributed computing and storage technology, the system can make full use of resources such as cloud computing platforms to achieve efficient processing of large-scale data. This helps to solve the problem of inefficiency of traditional processing methods in a big data environment. However, issues such as data consistency and communication overheads also need to be considered when implementing distributed systems. Deep Learning in NLP and Speech Recognition Discussion, Deep learning plays a key role in improving the accuracy of natural language processing and speech recognition. Choosing the right deep learning model structure and optimising the training process are crucial for system performance. However, the complexity of deep learning models also brings problems such as increased demand for computational resources and reduced model interpretability. Multilingual and Multi-Accent Processing Discussion, in the context of multilingual processing, building a multilingual corpus is crucial to ensure that the system has a wide range of language support. However, contextual differences between certain languages may lead to a degradation of model performance in some cases. For accent processing, model training needs to include data with representative accents to ensure the system's adaptability to different accents. User Privacy and Data Security Discussion, User privacy and data security are important concerns in big data processing. The use of encryption and anonymisation techniques in system design is part of ensuring user data security. However, as regulations and standards for data privacy continue to evolve, systems also need to be updated to comply with the latest legal requirements. System scalability and performance optimisation are discussed, and system scalability is key to ensuring high performance under growing loads. Optimisation of distributed computing and storage requires balancing the various components of the system to achieve a high degree of scalability. However, an overly distributed design can lead to increased complexity. Real-Time Processing and Streaming Data Processing are discussed, and the introduction of a real-time processing framework is an effective way to ensure that the system is able to process text and speech data generated in real time. However, real-time processing may also increase the complexity of the system, requiring a trade-off between the need to handle real-time and performance in the system design. Quality assessment and model optimisation are discussed as an ongoing process after the system has gone live. The system needs to establish a robust monitoring mechanism to detect performance degradation or accuracy issues in a timely manner and take corresponding optimisation measures. However, interpretability for deep learning models is still a challenge because the complex model structure is difficult to intuitively explain its decision-making process.

REFERENCES

[1] Qian, R., Sengan, S., & Juneja, S. (2022). English language teaching based on big data analytics in augmentative and alternative communication system. *International Journal of Speech Technology*, *25*(2), 409-420.

[2] Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, *31*(1), 113-126.

[3] Arora, M., & Sharma, R. L. (2023). Artificial intelligence and big data: ontological and communicative perspectives in multi-sectoral scenarios of modern businesses. *foresight*, *25*(1), 126-143.

[4] Iskamto, D. (2023). Data Science: Trends and Its Role in Various Fields. *Adpebi International Journal of Multidisciplinary Sciences*, *2*(2), 165-172.

[5] Eom, G., Yun, S., & Byeon, H. (2022). Predicting the sentiment of South Korean Twitter users toward vaccination after the emergence of COVID-19 Omicron variant using deep learning-based natural language processing. *Frontiers in Medicine*, *9*, 948917.

[6] Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. *Artificial Intelligence Review*, 1-37.

[7] Wang, H., Liu, W., & Wang, Y. (2022, June). Research on Transaction Mode and Dynamic Pricing Model of Big Data Service Based on Sequential Recommendation Algorithm. In *International Conference on Applications and Techniques in Cyber Intelligence* (pp. 993-1001). Cham: Springer International Publishing.

[8] Zhao, Y., & Miao, R. (2022). Network media public opinion and social governance supported by the internet-of-things big data. *Security and Communication Networks*, *2022*.

[9] Anjum, M., & Shahab, S. (2023). Improving Autonomous Vehicle Controls and Quality Using Natural Language Processing-Based Input Recognition Model. *Sustainability*, *15*(7), 5749.

[10] Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., & Cho, J. (2023). Development of Language Models for Continuous Uzbek Speech Recognition System. *Sensors*, *23*(3), 1145.

[11] Tyagi, N., & Bhushan, B. (2023). Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions. *Wireless Personal Communications*, *130*(2), 857-908.

[12] Varaprasad, R., & Mahalaxmi, G. (2022). Applications and Techniques of Natural Language Processing: An Overview. *IUP Journal of Computer Sciences*, *16*(3), 7-21.

[13] Lareyre, F., Nasr, B., Chaudhuri, A., Di Lorenzo, G., Carlier, M., & Raffort, J. (2023, September). Comprehensive review of Natural Language Processing (NLP) in vascular surgery. In *EJVES Vascular Forum*. Elsevier.

[14] Baskar, S., Dhote, S., Dhote, T., Jayanandini, G., Akila, D., & Doss, S. (2022). A predictive typological content retrieval method for real-time applications using multilingual natural language processing. *Expert Systems*, e13172.

[15] Balli, C., Guzel, M. S., Bostanci, E., & Mishra, A. (2022). Sentimental analysis of Twitter users from Turkish content with natural language processing. *Computational Intelligence and Neuroscience*, *2022*.