ISSN 2618-1584 Vol. 6, Issue 11: 25-29, DOI: 10.25236/FMSR.2024.061105

Research on the Application of Medical Text Matching Technology Combined with Twin Network and Knowledge Distillation in Online Consultation

Jinzhu Yang*

AI Research, Dyania Health Inc, Jersey City, New Jersey, 07310, United States
jinzhu.yang0625@yahoo.com
**Corresponding author*

Abstract: *Online consultation has become an important way for people to seek advice from doctors during the epidemic, but the lack of face-to-face communication and differences in patients' self-reported symptoms make it difficult for doctors to accurately assess the condition. In this paper, a Chinese medical text matching model combining twin network and knowledge distillation technology is proposed, and a medical question answering system is established. The system uses the Siamese-UniLM model to capture deep text information, greatly improving the accuracy and F1 value, and reducing model parameters and computing resources through knowledge distillation technology to achieve efficient online reasoning. The experimental results show that the model significantly improves the system response speed and resource utilization efficiency, and provides a reliable solution for providing scientific and accurate medical advice.*

Keywords: *Medical Text Comparison, Twin Networks, Knowledge Distillation, Online Consultation*

1. Introduction

With the rapid development of the economy, accompanied by an increasing demand for medical services that current resources are inadequately equipped to meet—highlighted by statistics revealing only 2.44 practicing physicians per 1,000 people in China—a range of challenges has arisen, including the uneven distribution of medical resources, overwhelming workloads for healthcare professionals, and extended patient wait times; amidst this context, the rapid evolution of the internet has notably improved access to medical services, positioning online question-and-answer systems as a robust complement to traditional healthcare delivery methods, effectively addressing the scarcity of medical resources.

Nevertheless, traditional question-answering systems, particularly in highly specialized fields like medicine, have yet to achieve optimal maturity, underscoring the urgent need for developing an intelligent medical question-answering system grounded in advancements in natural language processing tasks, pivotal for enhancing the medical domain by facilitating the dissemination of specialized medical information and enhancing the efficiency and quality of healthcare services.

The on-going advancement of these systems not only relies on technological progress but also demands continual updates to medical knowledge and a profound understanding of user requirements; future iterations must aim for heightened accuracy and response speeds, while simultaneously prioritizing data privacy and security to protect users' sensitive medical information.

Through the integration of advanced technology with empathetic design principles, the intelligent medical question-answering system is poised to significantly alleviate shortages in medical resources and elevate standards of healthcare provision, representing a transformative approach that promises to bridge critical gaps in healthcare delivery, ensuring equitable access to enhanced medical expertise and fostering improved patient outcomes.

ISSN 2618-1584 Vol. 6, Issue 11: 25-29, DOI: 10.25236/FMSR.2024.061105

2. Related Research

2.1. Present State and Future Outlook of Text Matching Research

This study applies advanced medical text matching technology to online consultation scenarios, and explores the key challenges and innovative solutions of semantic matching in the medical field in depth. It highlights how traditional semantic matching models face significant obstacles due to the lexical specialization and semantic complexity inherent in medical language. This often makes it difficult to accurately capture the nuances of meaning and technical terms that are crucial in the medical field. In this study, we propose a Siamese-UniLM model, which combines the advantages of a pre-trained model and dual network architecture to effectively capture deep semantic information from text. This study greatly improves the performance of medical question answering systems by utilizing the ability of pre-trained language models to make fine adjustments on a wide range of medical corpora to accurately understand and interpret complex medical queries and texts. Furthermore, we introduce the Tiny Siamese-UniLM model, which has been compressed and optimized through the application of knowledge distillation technology, where a smaller, more efficient model (the student model) is trained to replicate the performance of a larger, more complex model (the teacher model), significantly enhancing model parameters and inference speed without substantially sacrificing accuracy, making it well-suited for deployment in real-time online question-answering systems. These innovations, which not only enhance the efficiency and user experience of medical question-answering systems but also lay a robust technical foundation for future applications in medical information processing, achieve their impact by reducing computational resources required and improving response times, thereby making real-time online consultations more accessible and practical for widespread use. The adoption of such advanced models holds significant promise for the future of healthcare by enabling more accurate and timely responses to patient inquiries, thus improving patient satisfaction and outcomes, and the ability to deploy these models in real-time online environments ensures that patients can receive immediate support, which is crucial in urgent or time-sensitive medical situations. Moreover, the success of these models in capturing deep semantic information and handling complex medical terminology paves the way for their application in various other areas of medical information processing, including automated medical record analysis, intelligent clinical decision support systems, and advanced medical research tools, with the advancements presented in this paper representing a significant step forward in the field of medical text matching and semantic understanding, by harnessing the power of twin networks, pre-trained models, and knowledge distillation to develop a solution that addresses the unique challenges of medical language processing, offering a powerful tool for improving healthcare delivery and patient care.S Ishii et al. implicitly learned the correspondence between words in text and regions in image from text-image pairs through the attention mechanism [1]. Chen aims to solve this problem. The key idea is to test whether existing deep text matching methods meet some basic heuristic requirements in information retrieval [2].

2.2. The Development of Medical Text Matching Technology in Intelligent Question-answering Systems Enhances Patient-provider Interactions and Enables Future Innovations in Automated Medical Support

The medical question-answering system, representing an emerging research field with significant potential, has its critical need underscored by the novel coronavirus epidemic and the consequent surge in demand for medical resources, as establishing a robust system can substantially enhance the efficiency of medical services, alleviate the burden on healthcare professionals, and offer wide-ranging application prospects; however, while advancements in technologies such as automatic diagnosis and medical image recognition have gained popularity, these typically require substantial guidance and manual intervention from doctors, making them less efficient and more resource-intensive, in contrast to constructing a comprehensive corpus for medical Q&A, which is an extensive and labor-intensive endeavor, as traditional systems often based on search engines and fixed dialogue templates frequently fail to accurately capture the true intent of patients, leading to a suboptimal consultation experience; to address these limitations, an intelligent medical question-answering system must go beyond mere template-based interactions by dynamically interacting based on the patient's historical questions and answers to ensure that the outputted treatment information is both relevant and precise, requiring the system to identify the patient's intent through analysis of conversation history data and leverage advanced entity relationship extraction models to retrieve accurate information from various datasets such as CoQA (Conversational Question Answering), ATIS (Airline Travel Information System), and SNIPS (a dataset for spoken language understanding); furthermore, the system should incorporate

ISSN 2618-1584 Vol. 6, Issue 11: 25-29, DOI: 10.25236/FMSR.2024.061105

natural language processing (NLP) and machine learning algorithms to improve its understanding and responsiveness, enabling it to learn from interactions and continuously enhance its accuracy and relevance by integrating deep learning techniques to perform sophisticated semantic analysis, thereby ensuring accurate interpretation and response to patient inquiries; additionally, technical accuracy must be complemented by prioritizing user experience through a user-friendly interface, intuitive navigation, and multi-device accessibility, ensuring that patients can easily access and utilize the system, while data security remains a critical aspect, necessitating robust measures to protect sensitive patient information and comply with healthcare regulations; ultimately, an advanced medical question-answering system holds the promise of transforming healthcare delivery by providing accurate, timely, and efficient responses to patient inquiries, significantly improving patient outcomes, reducing the workload on healthcare providers, and streamlining medical services, and as research and development in this field progress, the potential applications and benefits of intelligent medical question-answering systems will continue to expand, offering innovative solutions to some of the most pressing challenges in healthcare.X Zhao et al. applied the pre-trained language model in the medical field to the text matching stage of the medical question answering system [3]. JS Jang et al discuss the QA pair matching method in the QA model, which finds the most relevant question and its recommended answer for a given problem. Existing studies of this method have been performed on entire datasets or on datasets within categories manually specified by the question writers [4].

3. Application of Semantic Matching Model Based on Knowledge Distillation in Modern Medical Question Answering System

As BERT has successfully combined Transformer architecture with pre-training fine-tuning of downstream tasks and made significant achievements in several areas of natural language processing, researchers have begun to use this idea to propose a variety of pre-training models. These models use large data sets to learn the universal representation of text and adjust parameters to achieve better performance in multiple text matching tasks. Effectively matching medical texts in medical question answering systems often requires fine-tuning all parameters of the pre-trained model, which leads to high storage costs at deployment and forgetting problems in continuous learning. The number of parameters in large models has increased rapidly in recent years, from one billion to one billion. Compared to the exponential growth in model size, performance gains for high-performance devices are limited. Addressing these scalability issues requires innovative approaches, such as combining twin networks and knowledge distillation techniques, to simplify model deployment and reduce the computational burden.

In supervised or unsupervised training, a large-scale deep neural network model can simulate the behavior of the teacher model and transmit its knowledge to the student model. The key lies in the effective transmission of information to ensure that students better understand and apply this knowledge. With the increasing number of parameters and network depth of large-scale models, the research shows that there is a problem of parameter redundancy. Knowledge transfer from large-scale models to small-scale models can be achieved with a modest reduction in performance. Two different knowledge distillation strategies of isomerization distillation and isomorphism distillation are discussed in this paper. Heterogeneous distillation involves the teacher model and the student model using different structures, while isomorphic distillation refers to the two using similar model structures. Because of its strong generalization ability, the teacher model can transfer deep knowledge to the student model through soft goals. Siamese-UniLM was introduced as a teacher model, which fully learned the medical knowledge in the medical question answering dataset. The student model, Tiny-Siamese-UniLM, consists of three layers of Transformer-Encoder and one full connection layer. The initialization parameters are partially weighted by Siamese-UniLM. The size of the teacher model is about 340M. Tiny-Siamese-UniLM's Transformer architecture captures the different layer features of medical text layer by layer, from surface features to semantic features, to deal with the long-distance dependencies of text.

The application of the medical text matching technology combining twin networks and knowledge distillation in online consultation shows that we have significantly improved the performance of the medical question answering system by using the Tiny-Siamese-UniLM model optimized by knowledge distillation. Compared with the original teacher model, the student model not only significantly improves the reasoning speed, but also significantly reduces the hardware resource consumption required by the system, which is crucial for the real-time response of the online medical Q&A system, and can better meet the needs of users.

ISSN 2618-1584 Vol. 6, Issue 11: 25-29, DOI: 10.25236/FMSR.2024.061105

Further analysis shows that the success of the student model over the teacher model is partly due to its more concise parameter Settings and more efficient learning ability. This optimization not only helps to maintain the complexity of the model, but also effectively avoids the risk of overfitting and performance degradation. Through knowledge distillation technology, the student model can effectively extract key features from the teacher model, and improve the generalization ability and actual operation efficiency of the model by simplifying the number of parameters.

This research not only promotes the progress of medical question answering system technically, but also provides a feasible technical path and practical reference for future applications in the field of medical information processing. By optimizing model performance and efficiency, we can better support the digital transformation of healthcare delivery and improve patient access to health information with greater convenience and accuracy.

4. Design and Implementation of Medical Question Answering System Based on Distillation Model

In the medical field, the design of a medical question-answering system must encompass not only the technical implementation but also prioritize user experience and the security of medical knowledge, with these systems expected to offer a straightforward browser access interface enabling users to effortlessly input medical and disease-related queries and receive instant responses to standardized questions, thereby ensuring timely and effective patient support; to uphold the accuracy and reliability of medical knowledge, the construction of the knowledge base should undergo a rigorous manual audit process, ensuring concise and clear content that prevents the dissemination of invalid or incorrect information, while the system's interface design should be intuitive and user-friendly, minimizing the chance of users receiving irrelevant or erroneous answers and allowing for straightforward navigation to obtain accurate information with minimal effort, and additionally, the system should support access from multiple terminal devices, ensuring convenience and a seamless user experience across various platforms; by leveraging deep learning technology, the system can conduct sophisticated sentence semantic analysis, which, when combined with an established medical knowledge base, enables automatic inquiry functionality, allowing patients to easily access and receive precise medical information through the system, and furthermore, robust data security measures must be implemented to protect sensitive user information and comply with healthcare regulations, with continuous updates and improvements to the knowledge base being essential to keep the information current and relevant, thereby integrating these considerations to enhance the medical question-answering system's utility as a valuable tool for both patients and healthcare providers, facilitating improved health outcomes by ensuring better access to reliable medical information.

5. Conclusion

This paper delves into the pivotal challenges and innovative solutions associated with semantic matching in the medical field by applying advanced medical text matching technology in online consultation scenarios, highlighting how traditional semantic matching models encounter significant obstacles when dealing with medical texts due to the lexical specialization and semantic complexity inherent in medical language, often struggling to accurately capture the nuanced meanings and specialized terminology critical in the medical domain [5]; to address these challenges, we propose the Siamese-UniLM model, which combines the strengths of pre-trained models and twin network architectures to effectively capture deep semantic information from texts, resulting in substantial performance improvements in medical question-answering systems by leveraging the power of pre-trained language models fine-tuned on extensive medical corpora to understand and interpret complex medical queries and texts accurately; furthermore, we introduce the Tiny Siamese-UniLM model, which has been compressed and optimized through the application of knowledge distillation technology, involving training a smaller, more efficient model (the student model) to replicate the performance of a larger, more complex model (the teacher model), significantly enhancing the model parameters and inference speed without substantially sacrificing accuracy, thus resulting in a highly efficient model well-suited for deployment in real-time online question-answering systems [6]; these innovations not only enhance the efficiency and user experience of medical question-answering systems but also lay a robust technical foundation for future applications in medical information processing by reducing the computational resources required and improving response times, making real-time online consultations more accessible and practical for widespread use; the adoption of such advanced models holds significant promise for the future of healthcare by enabling more accurate and

ISSN 2618-1584 Vol. 6, Issue 11: 25-29, DOI: 10.25236/FMSR.2024.061105

timely responses to patient inquiries, thus improving patient satisfaction and outcomes, and the ability to deploy these models in real-time online environments ensures that patients can receive immediate support, which is crucial in urgent or time-sensitive medical situations; the success of these models in capturing deep semantic information and handling complex medical terminology paves the way for their application in various other areas of medical information processing, including automated medical record analysis, intelligent clinical decision support systems, and advanced medical research tools; the advancements presented in this paper represent a significant step forward in the field of medical text matching and semantic understanding, and by harnessing the power of twin networks, pre-trained models, and knowledge distillation, we have developed a solution that addresses the unique challenges of medical language processing, offering a powerful tool for improving healthcare delivery and patient care.

References

[1] Ishii S, Yamazaki T, Ito S, et al. Improving Object Coverage of Text-to-Image Generation by Object Matching. Proceedings of the Annual Conference of JSAI, 2022: 705. DOI: 10. 11517/pjsai. JSAI2022. 0_2O1GS705.

[2] Chen L, Lan Y, Pang L, et al. Toward the Understanding of Deep Text Matching Models for Information Retrieval. J Intell Inf Syst, 2021. DOI: 10. 48550/arXiv. 2108. 07081.

[3] Zhao X, Li Z, Wu S, et al. Deep Text Matching in Medical Question Answering System. IEEE Access, 2021, 5, 59-61.

[4] Jang J S, Kong H Y. Question-Answering Pair Matching Based on Question Classification and Ensemble Sentence Embedding. Computer Systems Science and Engineering, 2023, 46 (9): 3471-3489. DOI: 10. 32604/csse. 2023. 035570.

[5] Cao Y, Cao P, Chen H, Kochendorfer K. M, Trotter A. B, Galanter W. L, & Iyer R. K. Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model. Cham: Springer International Publishing. 2022, 6, 245-256

[6] Varatharajah Y, Chen H, Trotter A, & Iyer R. K. A Dynamic Human-in-the-loop Recommender System for Evidence-based Clinical Staging of COVID-19. In HealthRecSys@ RecSys, 2020, 8, 21-22