Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: September 21, 2024 at 1:30 PM

Leveraging the Domain Adaptation of Retrieval Augmented Generation (RAG) Models in Conversational AI for Enhanced Customer Service

Recent advancements in Large Language Models (LLMs) have significantly transformed the field of conversational AI. Retrieval Augmented Generation (RAG) model stands out to be highly effective in knowledge-intensive NLP tasks. Rcecently, the RAG-end2endmodelfurtheroptimized the architecture and demonstrated notable performance improvements by jointly updating the query encoder and the passage encoder during training. However, the effectiveness of the RAG-based architectures remains relatively unexplored when fine-tuned on downstream NLP tasks in specialized domains such as customer service. Furthermore, a critical challenge persists in reducing the occurrence of hallucinations while maintaining high domain specific accuracy for building reliable conversational AI systems. In this research, weinvestigated the effectiveness of diverse RAG and RAG-like architectures through domain adaptation. To facilitate the evaluation of the models, we constructed a comprehensive dataset featuring domain-specific knowledge base and question answer pairs sourced from wide range of hotel-related conversations. We finetuned all the models to explore the potential of these models to achieve effective domain adaptation on our dataset and evaluated their ability to generate accurate and relevant responses. While achieving domain adaptation we also addressed a critical research gap on determining the impact of reducing hallucinations with domain adaptation across different RAG architectures, an aspect that was not properly measured in prior work. Our evaluation shows positive results in all metrics by employing domain adaptation, demonstrating strong performance on QA tasks and by providing insights into their efficacy in reducing hallucinations. Our findings clearly indicate that domain adaptation not only enhances the models' performance on QA tasks but also significantly reduces hallucination across all evaluated RAG architectures. Our work highlights the impact of domain adaptation in enhancing the reliability and accuracy of conversational AI models and contributes to the field by providing open-source domain-specific datasets and detailed performance analyses, offering valuable insights into the practical application of RAG models. By leveraging domain adaptation techniques, RAG models can be refined for superior performance with significant implications for improving the trustworthiness and user experience in automated customer service systems.