Hypothesis

In the pursuit of advancing the realm of question answering (QA) systems, this paper delves into the unexplored terrain of post-deployment optimization through the astute utilization of user interactions and feedback. The primary objective of the authors is to intricately address dual facets pivotal to system refinement: augmentation of the QA system's performance and endowing the model with the acumen to expound upon the validity or fallacy of its responses.

Contribution

The paper addresses the question of whether question answering (QA) systems can be improved in post-deployment based on user interactions and feedback.The authors collect a dataset called FEEDBACKQA, which contains interactive feedback from users of a deployed QA system.The paper investigates the usefulness of this feedback for improving the accuracy and explainability of retrieval-based QA systems. The authors train a neural model using the feedback data, which can generate explanations and re-score answer candidates. The experiments conducted in the paper demonstrate that the feedback data not only improves the accuracy of the deployed QA system but also benefits other stronger non-deployed systems

Methodology

There are two stages of data collection and they are pre-deployment and post deployment data collection. In the first stage, the researchers collected Covid-19-related content from various official websites such as WHO, US Government, UK Government, Canadian government, and Australian government.They trained a base RQA model with this dataset for each source separately.They use a type of neural network model which is a BERT-based dense retriever for the base model. In the second stage, questions are collected for individual passages beforehand and used as interactive questions to cover a range of topics in each source. Feedback is collected from crowdworkers for each question-answer pair, including a rating and a natural language explanation. The feedback is obtained from three different workers to ensure quality. They used dense passage retrievers to build RQA models, with BERT and BART pre-trained models used for obtaining embeddings. A poly-encoder is used to build question-sensitive document representations, which are computed using attention between question and passage embeddings.The RQA model is trained to maximize the log-likelihood of the correct answer using in-batch negative sampling technique. They compared two variants of the FEEDBACKRERANKER model and found that the one directly predicting the rating performed slightly better. To eliminate confounding factors, they trained a reranker model on pre-deployment data and compared it to the reranker trained on feedback data. They also combined the training data of FEEDBACKRERANKER and VANILLARERANKER to train a third reranker called COMBINEDRERANKER. The feedback-enhanced RQA model outperforms the base RQA model, with the COMBINEDRERANKER model being the strongest, indicating that feedback data and pre-deployment data complement each other, leading to improved accuracy even in unseen domains.

Limitation

The feedback collection setup in this paper is a simulation of a deployed model, and real-world feedback may contain sensitive information that needs to be handled with care, ensuring privacy and data protection.Real-world feedback can be noisy, meaning it may contain irrelevant or misleading information that can impact the accuracy of the feedback analysis.The training and inference of a reranker with feedback data increase the usage of computational resources

Future Plan

Dataset Augmentation:
Acquiring and curating a novel dataset named FEEDBACKQA, characterized by interactive feedback in the form of ratings and natural language explanations.

Model Enhancement:
Iteratively refining the QA model by leveraging the invaluable feedback data. This involves the implementation of a reranker to discern optimal answer candidates and generate coherent explanations.

Utility Assessment:
Scrutinizing the efficacy of feedback-driven explanations in augmenting the performance of our deployed system. This entails conducting human evaluations to gauge the extent to which explanations facilitate users in distinguishing between correct and incorrect answers.

Explanatory Content Analysis:
Delving into the diverse types of explanations present in the feedback data, including review-style narratives, succinct summaries of pertinent and extraneous content, and identification of missing information.

User-Centric Evaluations:
Undertaking meticulous human evaluations to appraise the practical utility of the explanations, focusing on their effectiveness in aiding users in making informed decisions within the QA system.

Comparative Performance Analysis:
Rigorously comparing the performance of our deployed QA system against other non-deployed systems, utilizing the feedback data as a benchmark for comprehensive assessment.

This multifaceted approach ensures a holistic exploration of the feedback-driven enhancements to our QA system, paving the way for a more robust and user-friendly solution in the realm of computer science research.