NLP Report

Hala Medhat 49-4186 Rana Emad 49-5267

March 14, 2024

1 Introduction

In an era marked by rapid advancements in artificial intelligence and natural language processing, the demand for efficient and intuitive question-answering systems has escalated exponentially. These systems play a pivotal role in numerous domains, ranging from customer service and education to healthcare and information retrieval. With the ever-increasing volume of digital information, the necessity for intelligent systems capable of swiftly and accurately responding to user queries becomes increasingly critical.

Traditional question-answering systems have primarily focused on retrieving relevant information from vast repositories of structured or unstructured data. While these systems have achieved notable success, they often struggle to maintain coherent and contextually relevant conversations over extended interactions. This limitation highlights the importance of integrating chat history into question-answering systems, thereby enhancing user experience and enabling more engaging interactions.

2 Motivation

The motivation behind the development of a question-answering system with integrated chat history stems from the recognition of the evolving nature of human-machine interactions. Users increasingly expect conversational agents to exhibit a deeper understanding of context and maintain continuity across interactions. By incorporating chat history, a question-answering system gains the capability to leverage past exchanges to enrich subsequent conversations, thereby fostering a more personalized and seamless user experience. Additionally, the ability to analyze aggregated chat history data provides invaluable insights into user behavior, preferences, and emerging trends, empowering organizations to make data-driven decisions and optimize system performance. In essence, the integration of chat history into question-answering systems represents a strategic evolution towards more adaptive, context-aware, and user-centering conversational AI solutions, poised to revolutionize human-computer interaction across a myriad of domains and applications. Furthermore, Analyzing past interactions enables question-answering systems to identify and rectify errors or misunderstandings that may have occurred during previous exchanges. By learning from past interactions, these systems can refine their understanding and improve response accuracy over time, ultimately enhancing the overall user experience.

3 Literature Review

The article [1] delves into the realm of conversational question answering (CQA), focusing on the open-retrieval approach to tackle the challenge. Previous approaches often simplified the problem by selecting answers from candidate sets or extracting them from given passages. However, the authors propose a novel framework that integrates retriever and reader models, allowing for the effective retrieval of relevant passages from large-scale document collections to answer user questions. These approaches overlooked the critical role of retrieval in conversational search. To bridge this gap, the ORConvQA setting was introduced. In ORConvQA, evidence is first retrieved from a large collection before answers are extracted. The OR-QuAC dataset was created to facilitate research in this area. An end-to-end system for ORConvQA, including a retriever, reranker, and reader based on Transformers, demonstrated the importance of a learnable retriever and the benefits of history modeling. Additionally, the article investigates methods for fine-tuning pre-trained language models to improve their effectiveness in conversational contexts. Through extensive experiments and evaluations on benchmark datasets, the authors demonstrate the effectiveness of their proposed approach, achieving significant improvements in retrieval accuracy and conversational coherence. Overall, the article provides valuable insights and contributions to the field of conversational question answering, offering a comprehensive framework for open-retrieval CQA systems.

The article [2] presents a thorough exploration of various deep-learning techniques applied to the domain of question-answering (QA) systems. Beginning with an overview of QA systems and their significance in information retrieval and natural language processing (NLP), the authors delve into the core concepts of deep learning, illuminating neural network architectures such as convolution neural networks (CNN's), recurrent neural networks (RNNs), and attention mechanisms. They discuss how these architectures have been adapted and refined for QA tasks, including both extractive and generative approaches. The article provides an in-depth analysis of different QA datasets and benchmarks commonly used for evaluating deep learning models, highlighting their strengths and limitations. Furthermore, it surveys recent advancements in pre-trained language models such as BERT, GPT, and their variants, including their impact on QA tasks and the paradigm shift towards transfer learning in NLP. The authors also explore challenges and future directions in the field, such as handling multi-turn conversations, incorporating external knowledge sources, and improving model interpretability. Through comprehensive critical analysis, this article serves as a valuable resource in understanding the deep learning approaches for question-answering systems, offering insights into the evolving landscape of QA research and applications.

In the study [3], the researchers aimed to develop an advanced chatbot capable of engaging in meaningful conversations with users. To achieve this, they employed deep learning techniques, specifically Bidirectional Recurrent Neural Networks (RNN) and an attention mechanism. The methodology involved collecting a dataset comprising pairs of questions and corresponding answers. This dataset was crucial for training the chatbot to understand various user queries and generate appropriate responses. During the training phase, the chatbot learned to analyze the input questions using the Bidirectional RNN, which allowed it to consider both past and future contexts in understanding the meaning of each word or phrase. Additionally, the attention mechanism helped the chatbot focus on relevant parts of the input when generating responses, enhancing the coherence and relevance of its replies. After training, the researchers tested the chatbot's performance by inputting different questions and evaluating the quality of its responses. They found that the chatbot generally performed well, generating accurate and contextually appropriate answers

for various queries. However, despite its overall success, the chatbot exhibited some limitations. For instance, it struggled with complex or ambiguous questions, often providing generic or irrelevant responses. This limitation underscores the challenge of teaching chatbots to grasp nuances in language and understand context effectively. Moreover, the researchers encountered difficulties in fine-tuning the model and ensuring its robustness across diverse user inputs. These challenges highlight the ongoing need for refinement and optimization in chatbot development, especially when aiming for high levels of conversational sophistication.

In the article [4], the authors proposed a question-answering chatbot tailored for closed-domain applications, with a specific focus on the educational domain. The training utilized the TREC QA dataset, comprising factoid questions along with corresponding incorrect and correct answers. The chatbot operates through three primary phases: knowledge base creation, answer candidate pool creation, and best answer sentence selection. Initially, relevant documents in PDF format are uploaded, and their text is extracted, cleaned, and subjected to several NLP transformations. Subsequently, user queries undergo text cleaning and preprocessing similar to the PDF text, with IDF and BM25 used to assess the similarity between the question and sentences in the knowledge base. The resulting answer pool and query are then fed to a neural network for processing. Additional features were introduced to improve performance to establish relatedness between question and answer pairs, including word overlap measures accounting for stop words. This enhanced solution achieved higher Mean Reciprocal Rank and Mean Average Precision compared to related works, particularly when incorporating overlap features and abstaining from dropout.

In the study [5], the paper explores the challenge of users posing ambiguous questions, leading to potential confusion for systems due to multiple possible interpretations. To address this, the authors propose an intent clarification task centered on yes/no questions, aiming to identify the correct user intent with the fewest conversation turns. Leveraging negative feedback from previous questions in the conversation history, the system selects the next clarifying question by seeking dissimilarity to negative results while maintaining relevance to the query. Initially, an initial model is trained to select the first clarifying question based on the original query. Subsequently, a maximum-marginal-relevance (MMR)-based BERT model (MMR-BERT) is introduced to incorporate negative feedback for subsequent question selection. The system terminates the conversation and provides documents upon user confirmation with positive feedback. In evaluation, MMR-BERT identifies user intents in a significant portion of conversations, outperforming baselines across various turns 41.2%, 52.2%, and 59.2% conversations by asking at most 3, 4, and 5 clarifying questions, particularly in identifying user intents from ambiguous queries. The authors also observe that the performance of MMR-BERT is more pronounced for ambiguous queries compared to faceted queries, emphasizing the importance of differentiating semantic meanings and leveraging negative feedback effectively. Conversely, all methods perform better on faceted queries, indicating the influence of initial candidate questions on overall performance, with ambiguous queries exhibiting less word matching in the corpus.

4 Data Analysis

The dataset [6] under investigation pertains to a question-answering system designed to preserve chat history for analysis. Known as the Customer Support on Twitter dataset, it serves as a significant and contemporary corpus of tweets and responses. This dataset is instrumental in advancing natural language understanding and conversational models, as well as in studying modern customer support practices and their impact.

Structured as a CSV file, each row in the dataset represents a tweet, with various columns providing detailed information. For instance, the "tweet_id" serves as a unique, anonymized identifier for each tweet, which is referenced by other columns such as "response_tweet_id" and "in_response_to_tweet_id." Moreover, the dataset includes an "author_id" column, which contains unique, anonymized user IDs. The "inbound" column indicates whether the tweet is an inbound message to a company providing customer support on Twitter, implying that the company is responding to the tweet. The "created_at" column records the date and time when the tweet was sent, while the "text" column contains the content of the tweet. Sensitive information, such as phone numbers and email addresses, is replaced with mask values to protect user privacy.

Furthermore, the dataset includes columns like "response_tweet_id," which lists the IDs of tweets that are responses to a particular tweet, and "in_response_to_tweet_id," which indicates the ID of the tweet to which a tweet is responding, if applicable. It is important to note that each conversation in the dataset comprises at least one request from a consumer and at least one response from a company. To distinguish between consumer and company user IDs, the "inbound" field can be used for calculation.

However, the dataset under examination presents several notable limitations and insights that warrant further elucidation. Primarily, a key limitation arises from the categorization of companies into diverse domains, encompassing sectors such as technical support, hospitality, automotive, among others. This categorization requires specifying the relevant domain to the model before responding to inquiries. Such specification is essential to ensure the provision of accurate responses tailored to the specific domain context. Failing to do so may result in inaccurate or irrelevant responses. Therefore, it is imperative to integrate a mechanism that enables the model to discern the appropriate domain context and tailor its responses accordingly. By incorporating this feature, we can ensure that the chatbot delivers precise and pertinent information aligned with the user's query and the specific domain in question. This approach not only enhances the effectiveness of the chatbot but also underscores the importance of domain adaptation in optimizing its performance across diverse industries and sectors.

During our examination, a pronounced discrepancy in the volume of outbound tweets from companies across diverse domains has become evident. Specifically, there is a notable variation in the number of tweets responded to by companies within each domain. For instance, within the technical domain, Apple Support alone managed a substantial volume of tweets, totaling 106,860. In contrast, the combined number of tweets handled by hotel-related support accounts, such as HotelTonightCX, HiltonHelp, and Kimpton, amounted to only 2,324. This significant difference in tweet volumes highlights a bias towards technical support companies within the dataset. Such a bias may impact the generalizability of findings and necessitates careful consideration when drawing conclusions or making decisions based on the data. Additional investigation into the underlying factors contributing to this bias is warranted to ensure a more comprehensive understanding of the

dataset's implications.

Another noteworthy limitation of this dataset is the potential for one tweet to have multiple response tweet IDs. This occurs when both a company and an individual respond to the same tweet. While our primary focus lies on the responses provided by companies, as they typically contain the relevant answers to the tweet, this presents a challenge. The issue arises when a company's response encompasses answers to multiple questions within the same dialogue, resulting in multiple conversations occurring within a single dialogue thread. This complicates the task of accurately attributing responses to specific inquiries and may lead to ambiguity in the dataset's interpretation. As such, careful consideration and possibly additional preprocessing steps are necessary to mitigate the impact of this limitation and ensure the validity and reliability of the dataset for analysis.

An additional challenge presented by the dataset is the presence of swear words in tweets. While the model must comprehend the meaning of such words, as well as emojis and other linguistic nuances, it is equally important to ensure that the model does not respond with swear words or emojis. This necessitates a delicate balance between understanding the context and intent behind these expressions while adhering to appropriate language standards in responses. Therefore, the model must be trained not only to recognize and interpret these elements accurately but also to generate responses that maintain a professional and respectful tone. Implementing filters or constraints during the training process to discourage the model from generating inappropriate responses can help address this challenge effectively. Additionally, incorporating guidelines or rules within the model's response generation mechanism can further reinforce the importance of maintaining decorum and appropriateness in interactions with users.

Continuing in a similar vein, let's address the aspect concerning the maximum length of conversations within the dataset, particularly focusing on a specific thread where users express their problems, and the company responds with a general question followed by individual replies to each user. The length of this particular conversation thread spans 1392 characters, indicating a comprehensive exchange between the company and its customers. This thread exemplifies a common pattern observed in customer support interactions, where a company initiates a conversation with a broad inquiry and subsequently engages with individual users to address their specific concerns. To overcome this challenge, a meticulous examination of the conversation thread is necessary, focusing on identifying mentions of individual users at the beginning of texts from the company.

The dataset's entries include minimal text, some as short as 7 characters, comprising only one author mention without additional data. Despite their brevity, these entries are crucial for understanding dataset distribution and characteristics, shedding light on user engagement patterns, author mention frequency and potential data quality issues like incomplete messages. Incorporating these insights into model development ensures robustness and accuracy, addressing diverse real-world scenarios effectively. Additionally, the dataset boasts a substantial 794,335 dialogues, offering rich training and evaluation data for question-answering models. Analyzing conversation distribution and structure informs model architecture design, evaluation strategies, resource allocation decisions, scalability considerations, and targeted performance enhancements. Notably, 45% of the responses have the inbound variable set to false, indicating the company answering the conversations. This highlights companies' proactive engagement strategies, crucial for accurately modeling conversational dynamics. By incorporating these insights, we can adapt the model's behavior to various question-answering scenarios, ensuring robust performance across diverse interaction contexts. Incorporating these insights enhances the model's ability to respond effectively

to user queries, fostering a deeper understanding of customer-company interactions and advancing question-answering systems for more engaging customer service experiences.

References

[1] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer, "Open-retrieval conversational question answering," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 539–548, 2020.

[2] Y. Sharma and S. Gupta, "Deep learning approaches for question answering system," *Procedia Computer Science*, vol. 132, pp. 785–794, 2018. International Conference on Computational Intelligence and Data Science.

[3] M. Dhyani and R. Kumar, "An intelligent chatbot using deep learning with bidirectional rnn and attention model," *Materials Today: Proceedings*, vol. 34, pp. 817–824, 2021. 3rd International Conference on Science and Engineering in Materials.

[4] D. Singh, K. Suraksha, and S. Nirmala, "Question answering chatbot using deep learning with nlp," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6, 2021.

[5] K. Bi, Q. Ai, and W. B. Croft, "Asking clarifying questions based on negative feedback in conversational search," in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 157–166, 2021.

[6] S. Axelbrooke, "Customer support on twitter," 2017.