QA model with Follow-up Question Generation for Better Answering in Large Language Models

Chaitanya Chakka
chvskch@bu.edu

Muskandeep Jindal
mujindal@bu.edu

Abstract

In numerous scenarios, individuals encounter challenges in fully articulating their circumstances to AI models, resulting in incomplete or less-than-ideal answers due to the omission of critical details. This limitation constrains the contextual accuracy of the model's outputs. We propose a model that emulates the behavior of a human counselor by posing contextually relevant questions to capture missing information and enhance its understanding of the user's situation before delivering a final response. By employing this interactive methodology, the model not only augments the depth of its responses but also ensures that its advice is precisely tailored to the user's specific needs across various domains, such as visa applications, nutrition, and fitness.

1 Related Work

The task of generating suitable question-answering pairs has been extensively studied, with various architectures and approaches applied to improve model performance. One such approach is demonstrated by (5) in which a fine-tuned BERT model was employed for the direct classification of question-answer pairs. In this study, the authors utilized the Microsoft Network Dataset, where a question-answer pair is input into the model to predict whether the answer is the accepted one. The model was trained by concatenating the question and answer with a separator token (SEP) and then applying a softmax classification layer over the output of the [CLS] token. This approach yielded an accuracy of 77%, even when tested on highly imbalanced datasets.

In contrast, (1) addressed a similar task, though they focused on ranking the quality of posts on StackOverflow. Their methodology relied heavily on feature engineering, carefully curating over 20 distinct features. These features were then used to classify posts into four categories based on the StackOverflow score and the presence of an accepted answer. Notable features used in their classification process include readability metrics such as the SMOG Index, the Flesch-Kincaid Index, and the Dale-Chall Readability Score. By incorporating these features, they aimed to create a robust classification framework that captured both linguistic and structural aspects of the posts.

Furthermore, (6) explored a more complex model architecture by integrating convolutional neural networks (CNN), long short-term memory networks (LSTM), conditional random fields (CRF), and self-attention mechanisms to classify question-answer pairs. Their model aimed to assess whether a given answer to a question is good, potentially useful, or irrelevant. Using the SemEval-2015 Task A dataset, which contains multiple answers per question classified into the aforementioned categories, the authors achieved a macro F1 score of 58.29%. This marked a 1.77 percentage point improvement over the previous state-of-the-art, demonstrating the efficacy of their multi-layered model design.

The relevance of these studies lies in their success within open-domain environments, where the questions and answers are not restricted to a single subject area. These models have shown the capacity to generalize well across diverse topics, learning the underlying language characteristics of effective question-answering. However, applying these models to a focused domain could further enhance their effectiveness. By narrowing the model's scope to a specific domain, the model can better capture the semantic nuances and domain-specific language, potentially leading to improved performance.

2 Data

Our approach focuses on working with closed-domain QA datasets, the goal is to retrieve domain-specific answers for fields like immigration consul-

tancy, technical IT support, or medical QA. The following sections describe each of the datasets collected. The data obtained has question-answer pairs, answer scores for some records, tags for subcategories, and other metadata. We aim to test the proposed paradigm on multiple datasets, though this will depend on the time constraints of the project.

Visa and Immigration: These services are commonly provided by human counselors but have not been extensively explored as a closed-domain task in previous research. Consequently, well-defined datasets are lacking. To address this, we combined data from multiple sources to gather a healthy number of data points. We have collected a total of around 25k immigration-related data records from Stack Exchange(Query) and immigration help blog. The immigration help blog dataset was extracted by performing web scraping by using Beautifulsoup, a Python package for HTML parsing. The Stack Exchange dataset is a BigQuery forum database consisting of community questions and multiple answers across multiple domains.

Healthcare: The second dataset is related to Covid QA obtained from (3) which is human annotated by biomedical sciences experts. The dataset follows a hierarchical structure of QA pairs, question, id, answer start, answer text, and a boolean if it is impossible to answer.

Technical Support: The third dataset is related to technical IT support obtained from (2). It consists of questions posed in a technical forum by technical users, who had a specific information need, and the answers are in technical documents that another human had linked in the "accepted answer" to the post.

Multi-domain: The last obtained dataset is a conversational dataset obtained from IBM developers. This dataset consists of over 20k annotated multi-domain, task-oriented conversations between a human and a virtual assistant. The dataset was annotated by the crowd, where each row was annotated by 5 different annotators. The confidence level for each row is provided by the annotation platform (Appen) based on the annotator's inter-agreement and annotator level. (4)

2.1 Dataset Preprocessing

A lot of unnecessary tokens are generally involved in the datasets which add redundant information and can sometimes affect the performance of the model. In this project, we have employed standard pre-processing techniques including lowercase, stopword removal, and stemming. We have also removed items like URLs since they do not necessarily add any semantic information to the model. On the same lines, we have also removed punctuation and special characters for the same reason. For some special datasets like the Stack Exchange forum, the answers are mentioned in HTML format since these websites allow users to post with HTML and markdown syntaxes. To tackle this, we have employed an HTML parser like Beautiful Soup to resolve tags like <p>(paragraph) and <a>(anchor).

3 Evaluation Metrics

Question-answering systems demand some unique metrics since these tasks involve both syntactic and semantic inferencing for prediction. Tasks like Parts of Speech tagging which require predicting in classes would suffice with metrics like Balanced Accuracy but tasks like summarization and question-answering systems need metrics that take even the semantic correctness into account.

The Recall-Oriented Understudy for Gisting Evaluation (ROGUE for brevity) score measures the similarity of machine-generated sentences and reference summaries using overlapping n-grams. Internally, it calculates the precision, recall, and F1 score of the instances. Three variants of this score depend on how the sentence has been generated like overlap of n-grams(ROGUE-N), longest common subsequence(ROGUE-L), and skip-bigram overlap (ROGUE-S). We plan to use ROGUE-S for our model as it offers more flexibility and allows gaps between the words. Since for the baseline we have used the F1 score for evaluation, ROGUE adds a layer over the F1 score, accounting for the semantic information in the task, making it more suitable for the question-answering problem at hand.

Bilingual Evaluation Understudy (BLEU for brevity) is a well-known evaluation score for calculating the accuracy of machine translation although the calculation can be utilized for any sequence-to-sequence models. The score is calculated by calculating the precision of the n-gram model and adjusting it with a penalty for translations that are shorter than the reference translations.

4 Baseline

A suitable baseline we can compare the model is a standard direct question-answering model. For the

milestone, we have taken a baseline run adopted from (2) which involves running a pre-trained BERT large model with a whole word masking model. The input token length is 512 tokens with 24 transformer layers 16 attention heads and 1024 embedding dimensions. The sentence is appended with *CLS* in the beginning and *SEP* token to separate the answer and question followed by a final *SEP* token. The task head consists of two fully connected feed-forward layers along with softmax output activation for answer extraction. We have also considered the TechQA dataset for this experiment since we are dealing with close-domain question answering. This dataset has an average length of 52.1 tokens for questions and 48.1 tokens for answers. For the dataset, there is also a concatenation of the query title and body using a *SEP* token to include more context. We have observed a BEST F1 score to be around 47.87 which we are setting as the baseline of this project.

References

[1] Zeeshan Anwar, Hammad Afzal, Ali Ahsan, Naima Iltaf, and Ayesha Maqbool. 2023. A novel hybrid cnn-lstm approach for assessing stackoverflow post quality. *Journal of Intelligent Systems*, 32(1):20230057.

[2] Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, et al. 2019. The techqa dataset. *arXiv preprint arXiv:1911.02984*.

[3] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

[4] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*.

[5] Bhaskar Sen, Nikhil Gopal, and Xinwei Xue. 2020. Support-bert: predicting quality of question-answer pairs in msdn using deep bidirectional transformer. *arXiv preprint arXiv:2005.08294*.

[6] Yang Xiang, Xiaoqiang Zhou, Qingcai Chen, Zhihui Zheng, Buzhou Tang, Xiaolong Wang, and Yang Qin. 2016. Incorporating label dependency for answer quality tagging in community question answering via CNN-LSTM-CRF. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1231–1241, Osaka, Japan. The COLING 2016 Organizing Committee.