

Follow-up Question Generator Analysis – Based on Chatgpt API

Chencheng Yi

https://github.com/ChenchengY/cs6200_final

Section 1

- **What is the task, and why is it important to users?**

As conversational AI systems like chatbots continue to advance, evaluating their ability to ask relevant follow-up questions is critical. This capability directly enables more natural, dynamic dialogue that meets user needs.

Specifically, analyzing follow-up question generation performance for inputs of varying complexity provides key insights. My project mainly focus on the following part:

By distinguishing between simple and complex initial user queries, this analysis reveals how well chatbots can adapt their responses for contextual relevance.

Examine the precision differences across granularity levels for follow-up question generation.

- **In general, what do queries look like?**

For each user query, I would send the user query to the chatbot and it would generate 10 possible follow-up question for me.

For the granularity analysis, each user query would have 5 follow-up questions for each sentence, paragraph and document level.

Both result above is already ranked based on the judgement from the chatbot.

Sample query listed as below, covering 10 topics

:Technology in Sports:

'How has the use of wearable technology improved athlete performance?'

Evolution of Music Genres:

'What are the characteristics that define a particular music genre?'

Artificial Intelligence in Scientific Research:

'How can AI assist in analyzing large datasets in scientific research?'

Ethical Considerations of Genetic Engineering in Agriculture:

'What are the potential benefits and risks of genetically modified organisms (GMOs) in food ',

renewable energy

'What are the most promising renewable energy sources for a sustainable future?'

Virtual Reality in Education and Training:

'How can virtual reality enhance immersive learning experiences for students?'

Social Media and Political Activism:

'Can social media platforms be effective tools for organizing and mobilizing social movements?'

Advancements in Medical Technology:

'What ethical considerations arise from the use of emerging medical technologies?'

. Exoplanets and the Search for Extraterrestrial Life:

'Have we found any promising candidates for hosting extraterrestrial life?'

Climate Change and Biodiversity:

'What conservation efforts are being made to protect biodiversity in the face of climate change?'

- **What kinds of results would be relevant to these queries? How many relevant results should there be per query?**

For each user query, both simple and complex, ten follow-up questions were generated by the system. Of those ten questions, seven were deemed relevant and three irrelevant for each query based on relevance assessments. Relevance was evaluated using the following criteria:

For granularity analysis, five follow-up questions were generated at each of the defined granularity levels for every user query analyzed. Four of the five questions for each granularity level were considered relevant follow-ups, with one irrelevant question according to relevance judgement. Below is how I define the relevance of a generated question

Entity:

1. **Question Quality:** Assess how well the chatbot generates grammatically correct, coherent, and contextually relevant follow-up questions.
2. **Relevance:** Measure the alignment of generated questions with the user's input or ongoing conversation.

Relationship:

3. **Diversity:** Evaluate if the chatbot produces a variety of follow-up questions to engage users, rather than repetitive ones.
4. **Engagement:** Analyze the effectiveness of questions in encouraging deeper interaction and meaningful dialogue.

Here is an example,

For the User query 'How has the use of wearable technology improved athlete performance?' Below is the generated questions and corresponding relevance scores.

"How does athlete work?", # Less obvious as irrelevant (0)
"How does wearable technology help athletes prevent injuries and monitor their overall health?", (1)
"What specific types of wearable technology are commonly used by athletes?", (1)
"Can you provide examples of professional athletes who have benefited from using wearable technology?", (1)
"What's the performance technology?", # Less obvious as irrelevant (0)
"How is the technology evolve in Earth?", # Less obvious as irrelevant (0)
"How does wearable technology track and analyze athlete performance?", (1)
"What are some key metrics that wearable technology can measure during training or competition?", (1)
"Are there any potential drawbacks or limitations to using wearable technology in sports?", (1)
"Can you discuss any notable studies or research conducted on the impact of wearable technology on athlete performance?" (1)

● **How should the results be organized**

Organization of Results:

The results will be systematically organized into a ranked list. This ranking will be predicated on the evaluative judgments made by the chatbot, thereby ensuring that the most relevant and contextually appropriate follow-up questions are prioritized.

Evaluation Metrics:

To effectively assess the performance of the generated results, we will deploy a suite of metrics designed to capture various dimensions of quality and relevance:

- Precision across diversified topics.
- Comparative analysis of MAP (Mean Average Precision) and F1 scores between different result sets, including original questions and those enhanced with nuances and details.
- Jaccard Similarity and Cosine Similarity measures to evaluate the overlap and directional similarity of the result sets, respectively.
- Semantic analysis leveraging the BLEU (Bilingual Evaluation Understudy) score to gauge the translation quality of the chatbot's responses in comparison to a set of high-quality reference translations.

Implementation Description and Performance Analysis:

Our implementation entails the development of a sophisticated follow-up question generator powered by the ChatGPT API. The process includes the following steps:

1. Generate ten follow-up questions for each of ten distinct topics using the ChatGPT API.
2. Assign relevance scores to these follow-up questions based on empirical observations and expert assessment.
3. Compute MAP and F1 scores for the resulting sets to quantify the accuracy and precision of the chatbot's responses.
4. Calculate Jaccard Similarity, Cosine Similarity, and BLEU scores to further dissect the linguistic and contextual alignment of the generated questions with the expected outcomes.

This multi-faceted analysis will not only validate the functional efficacy of the chatbot but also provide insights into its linguistic acumen and contextual responsiveness.

Executive Summary of Chatbot Performance Analysis

Key Findings:

- The chatbot demonstrated a robust capacity for generating follow-up questions, as evidenced by high MAP (Mean Average Precision) and F1 scores across both simple and complex queries.
- Performance metrics indicate a slight edge for complex queries over simple ones in terms of MAP, suggesting a nuanced understanding and processing of detailed inquiries.

Performance Breakdown:

Simple Queries:

- Mean Average Precision (MAP): 0.751
- Mean F1 Score: 0.823

Complex Queries:

- Mean Average Precision (MAP): 0.815
- Mean F1 Score: 0.824

The marginal disparity in F1 scores between simple and complex queries suggests consistency in accuracy. However, a higher MAP for complex queries implies that detailed questions may elicit more precise and relevant follow-up questions, potentially due to the chatbot's ability to dissect and engage with intricate details and subtopics.

Question Similarity Analysis:

An in-depth analysis of question similarity was conducted across 10 diverse topics. This involved evaluating both simple and complex datasets, aggregating cosine and Jaccard similarity metrics for each user query and its corresponding follow-up question and computing the average. Notably, question similarity exhibited greater variability, which could be attributed to the chatbot's exploration of varied subtopics. Despite this, precision remained unaffected, underscoring the chatbot's adeptness at maintaining relevance.

Detailed Similarity and Diversity Metrics:

topic 1

Simple Question Avg Similarity: 0.38140804537442113

Complex Question Avg Similarity: 0.3084478834843163

topic 2

Simple Question Avg Similarity: 0.2481236947683568

Complex Question Avg Similarity: 0.19039577922333667

topic 3

Simple Question Avg Similarity: 0.3403562282512874

Complex Question Avg Similarity: 0.33466841776564926

topic 4

Simple Question Avg Similarity: 0.38903610375550524

Complex Question Avg Similarity: 0.3979798495215102

topic 5

Simple Question Avg Similarity: 0.40412910942738434

Complex Question Avg Similarity: 0.18129720007253508

topic 6

Simple Question Avg Similarity: 0.3158706720178279

Complex Question Avg Similarity: 0.37994012664012233

topic 7

Simple Question Avg Similarity: 0.564789606360631
Complex Question Avg Similarity: 0.34473827301380516
topic 8
Simple Question Avg Similarity: 0.4808796398722094
Complex Question Avg Similarity: 0.30747819987248637
topic 9
Simple Question Avg Similarity: 0.3359185431110457
Complex Question Avg Similarity: 0.3979798495215102
topic 10
Simple Question Avg Similarity: 0.6530104766406845
Complex Question Avg Similarity: 0.31775415560916953

topic 1
Simple Question Diversity: 0.616
Complex Question Diversity: 0.621
topic 2
Simple Question Diversity: 0.675
Complex Question Diversity: 0.737
topic 3
Simple Question Diversity: 0.638
Complex Question Diversity: 0.683
topic 4
Simple Question Diversity: 0.756
Complex Question Diversity: 0.624
topic 5
Simple Question Diversity: 0.556
Complex Question Diversity: 0.748
topic 6
Simple Question Diversity: 0.625
Complex Question Diversity: 0.612
topic 7
Simple Question Diversity: 0.543
Complex Question Diversity: 0.644
topic 8
Simple Question Diversity: 0.593
Complex Question Diversity: 0.617
topic 9
Simple Question Diversity: 0.628
Complex Question Diversity: 0.624
topic 10
Simple Question Diversity: 0.546
Complex Question Diversity: 0.639

Granularity Analysis:

A granularity analysis underscored the chatbot's improved performance when scaling from paragraph to sentence to document level:

- **Sentence Level:**
 - MAP: 0.854
 - F1: 0.889
- **Paragraph Level:**
 - MAP: 0.808
 - F1: 0.889
- **Document Level:**
 - MAP: 0.928
 - F1: 0.913

This trend indicates an enhanced ability to extract and utilize context as the unit of analysis increases in size, culminating in superior performance at the document level.

Conclusion:

The chatbot's performance metrics, particularly in the context of complex queries, attest to its advanced capabilities in generating relevant, nuanced follow-up questions. The consistency in precision across varying degrees of query complexity and the scalability of performance with the increase in data granularity are indicative of a well-designed and effective chatbot system.

● Performance Analysis

This analysis serves as an initial prototype to assess the follow-up question generation capabilities of a developed chatbot system. The project enables analysis of similarity, user preference, and consistency metrics for the bot's automatically generated follow-up questions. The following project components have been implemented thus far:

B Level Criteria Implementation:

- Chatbot system development complete with multi-turn conversations
- Generated follow-up question pipelines integrated
- Relevance evaluation of follow-up question responses

B+ Level Criteria:

- Similarity analysis - Jaccard and cosine similarity metrics calculated to quantify effect of query complexity

A- Level Analysis:

- Semantic analysis conducted and relevance evaluation performed to assess granularity differences in generated follow-up questions

This project has strong foundational components for analysis of chatbot follow-up question generation. As a basic prototype without machine learning for more automated and robust relevance classification, there are opportunities in future work for enhancing analysis with additional models and improved annotator agreement quantification.

Section 2 – Sample queries

Simple Queries:

1. How has the use of wearable technology improved athlete performance?

Narrative: The user is seeking information on how wearable technology has positively impacted the performance of athletes, looking for examples and benefits.

"How does athlete work?", # Less obvious as irrelevant (0)

"How does wearable technology help athletes prevent injuries and monitor their overall health?", (1)

"What specific types of wearable technology are commonly used by athletes?", (1)

"Can you provide examples of professional athletes who have benefited from using wearable technology?", (1)

"What's the performance technology?", # Less obvious as irrelevant (0)

"How is the technology evolve in Earth?", # Less obvious as irrelevant (0)

"How does wearable technology track and analyze athlete performance?", (1)

"What are some key metrics that wearable technology can measure during training or competition?", (1)

"Are there any potential drawbacks or limitations to using wearable technology in sports?", (1)

"Can you discuss any notable studies or research conducted on the impact of wearable technology on athlete performance?" (1)

2. What are the characteristics that define a particular music genre? Narrative:

The user is interested in understanding the defining characteristics that distinguish one music genre from another, possibly for music appreciation or analysis.

How do music genres differ from one another in terms of sound and style?", (1)

"What's your favorite color?", # Less obvious as irrelevant (0)

"How do music genres evolve and change over time?", (1)

"How do cultural and historical factors influence the development of music genres?", (1)

"What are the key elements that contribute to the identity of a specific music genre?", (1)

"How does musician work?", # Less obvious as irrelevant (0)

"How do music genres influence fashion, lifestyle, and subcultures?", (1)

"How do music genres impact audience preferences and consumption patterns?", (1)

"Can music genres be blended or mixed together, and if so, what are the results?", (1)

"Are there any common themes or lyrical content associated with specific music performer?" # Less obvious as irrelevant (0),

3. **How can AI assist in analyzing large datasets in scientific research?**

Narrative: The user wants to know how artificial intelligence is utilized to analyze extensive datasets in scientific research, aiming to grasp the practical applications and advantages of AI in this context.

"What specific techniques or algorithms does AI employ to analyze large scientific datasets?", (1)

"Can you provide examples of scientific research areas where AI has been successfully used for data analysis?", (1)

"What's the latest technology gadget on the market?", # Irrelevant (0)

"What are the advantages of using AI for data analysis compared to traditional statistical methods?", (1)

"Are there any limitations or potential biases that researchers should be aware of when using AI for dataset analysis?", (1)

"How does AI assist in identifying patterns, correlations, or anomalies in large scientific datasets?", (1)

"What are some of the computational requirements and resources needed to implement AI for data analysis?", (1)

"How many hours does AI assistant work?", # Irrelevant (0)

"Can you share your favorite recipe when you are doing research?", # Irrelevant (0)

"What ethical considerations should be taken into account when using AI in scientific research to analyze large datasets?" (1)

Complex Queries:

4. How has the use of wearable technology improved athlete performance across different sports when it comes to quantifiable metrics around endurance, motor control, injury prevention, and Return on Investment for professional training?" Narrative: The user is seeking information on how wearable technology has positively impacted the performance of athletes, looking for examples and benefits.

"Can you provide specific examples of how wearable technology has improved endurance in different sports?", (1)

"How do wearables help athletes enhance their motor control in sports that require precise movements?", (1)

"Are there any studies or research that have quantified the impact of wearable technology on injury prevention in athletes?", (1)

"Can you explain how wearables assist in optimizing recovery for athletes, and what metrics they track to determine the recovery process?", (1)

"Have there been any notable cases where the return on investment for professional training significantly improved due to the use of wearable technology?", (1)

"Can you share a fun fact about a random element on the wearable technology?", # Irrelevant (0)

"How do wearable guitar provide actionable insights and recommendations based on the data collected?", # Irrelevant (0)

"Are there any regulations or guidelines in place regarding the use of wearable technology in professional sports?", (1)

"What are some potential risks or downsides associated with relying heavily on wearable technology for athlete performance improvement?", (1)

"What's your opinion on the best way to play basketball?", # Irrelevant (0)

5. "What are the key musical qualities like rhythm, tonality, instrumentation, melody, and lyrical themes that define and distinguish between genres as varied as jazz, hip-hop, rock, country, and electronic dance? Narrative: The user is interested in understanding the defining characteristics that distinguish one music genre from another, possibly for music appreciation or analysis.

"Can you provide examples of specific jazz artists or songs that showcase the improvisational nature and complex harmonies you mentioned?", (1)

"How has hip-hop evolved over time in terms of its rhythmic emphasis and lyrical themes?", (1)

"Are there any notable move that deviate from the typical characteristics you described?", # Irrelevant (0)

"What's your preferred brand of toothpaste?", # Irrelevant (0)

"What are some popular subgenres within dance , and how do they differ in terms of move and history?", # Irrelevant (0)

"Are there any specific artists or songs that have successfully blended multiple genres, creating a unique musical style?", (1)

"How do the lyrical themes in rock music differ from those in hip-hop or country music?", (1)

"Can you elaborate on the role of technology in shaping the sound and production of electronic dance music?", (1)

"Are there any regional or cultural influences that have contributed to the distinct characteristics of these genres?", (1)

"How have the advancements in music production and recording technology influenced the sound and evolution of these genres over time?" (1)

Granularity level Query

User input: "What are the most promising renewable energy sources for a sustainable future?":

Sentence level:

"What are renewable energy sources, and why are they important for sustainability?",

"What are some common examples of renewable energy sources?",

"How do renewable energy sources differ from fossil fuels in terms of environmental impact?",

"Can you explain the concept of energy efficiency in renewable technologies?",

"What are the basic principles of harnessing energy from renewable sources?"

,

Paragraph level:

"Are there specific countries or regions that have made significant strides in adopting renewable energy?",

"What are the economic implications of transitioning to renewable energy sources?",

"Can you discuss the role of government policies and incentives in promoting renewable energy?",

"Are there any emerging technologies or innovations in renewable energy production?",

"How do renewable energy sources contribute to reducing greenhouse gas emissions?"

,

Document level:

"What are the global trends in renewable energy adoption and investment?",

"How do different renewable sources, such as solar, wind, and hydroelectric power, compare in terms of scalability?",

"Can you provide insights into the challenges and barriers faced by the renewable energy sector?",

"What is the outlook for renewable energy as a primary source of power in the coming decades?",

"How does the integration of renewable energy sources impact grid stability and energy storage?"