Asking Clarifying Questions in Open-Domain Information-Seeking Conversations

Mohammad Aliannejadi\textsuperscript{(1)}, Hamed Zamani\textsuperscript{(2)}, Fabio Crestani\textsuperscript{(1)}, and W. Bruce Croft\textsuperscript{(2)}

(1) Università della Svizzera italiana (USI), Switzerland
(2) University of Massachusetts Amherst, USA
Choosing Clarifying Questions

- Evaluation: User, NPL, OASIS, S, 32
- Model:

SIGIR 2022
Dinosaur

Reptile

Dinosaurs are a diverse group of reptiles of the clade Dinosauria. They first appeared during the Triassic period, between 243 and 233.23 million years ago, although the exact origin and timing of the evolution of dinosaurs is the subject of active research. Wikipedia

Scientific name: Dinosauria

Higher classification: Dinosauriformes

Phylum: Chordata

Subclass: Diapsid

Lower classifications

Tyrannosaurus rex
Bird
Velociraptor
Spinosaurus
Diplodocus
dinosaur discovery channel
dinosaur

???
Can we ask questions to clarify the user information needs?

Johannes Kiesel et al. Toward Voice Query Clarification. SIGIR 2018
Radlinski and Craswell. A Theoretical Framework for Conversational Search. CHIIR 2017
<table>
<thead>
<tr>
<th>Information Need (Facet)</th>
<th>Information Need (Facet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm looking for the Discovery Channel's dinosaur site, which has pictures of dinosaurs and games.</td>
<td>I'm looking for a list of all (or many of) the different kinds of dinosaurs, with pictures.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Are you looking for dinosaur books?</th>
<th>Are you looking for dinosaur books?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No, just the discovery channel website.</td>
<td>Yes, if they contain pictures of all the different kinds of dinosaurs.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Are you looking for meat-eating or plant-eating dinosaurs?</th>
<th>Which dinosaurs are you interested in?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm not sure.</td>
<td>I'm interested in any and all dinosaurs.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Would you like to see pictures or videos of dinosaurs?</th>
<th>Do you want a list of dinosaurs names?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'd like to see pictures of dinosaurs on the discovery channels website.</td>
<td>Yes, I would also like the list to include pictures of the dinosaurs.</td>
</tr>
</tbody>
</table>
How to evaluate?
ClueWeb Collection

• A part of the Lemur Project
• A common web crawl (English) with 50M documents
• TREC Web Track 2009 – 2012
• Ad-hoc retrieval and **diversification** tasks

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># topics</td>
<td>198</td>
</tr>
<tr>
<td># faceted topics</td>
<td>141</td>
</tr>
<tr>
<td># ambiguous topics</td>
<td>57</td>
</tr>
<tr>
<td># facets</td>
<td>762</td>
</tr>
<tr>
<td>Average facet per topic</td>
<td>3.85 ± 1.05</td>
</tr>
<tr>
<td>Median facet per topic</td>
<td>4</td>
</tr>
<tr>
<td># informational facets</td>
<td>576</td>
</tr>
<tr>
<td># navigational facets</td>
<td>185</td>
</tr>
</tbody>
</table>
TREC facets

- Discovery channel
- Dinosaur pictures
- Dinosaur books

Search for dinosaur
An offline evaluation methodology

• We assume that each user is interested in one facet per topic.
An offline evaluation methodology

Let \( T = \{t_1, t_2, ..., t_n\} \) be the set of topics (queries).

A collection of facet sets: \( F = \{f_1, f_2, ..., f_n\} \)

\( f_i = \{f_1^i, f_2^i, ..., f_{m_i}^i\} \) includes all defined facets for topic \( t_i \).

A collection of clarifying question sets: \( Q = \{q_1, q_2, ..., q_n\} \)

With \( q_i = \{q_1^i, q_2^i, ..., q_{m_i}^i\} \) including all clarifying questions relevant to topic \( t_i \).

An offline evaluation requires defining \( A(t, f, q) \rightarrow a. \)
Task Description

Welcome and thanks for your interest. In this HIT, we would like to find out about different ways a search engine can ask questions from users to clarify their intention of a search. Therefore, here we show you a query which a user has submitted to a search engine (e.g., Google) and expect you to submit 6 clarifying questions. We highly recommend you to submit the same query to a search engine (Google, Bing, etc.), quickly scan the results and and see what questions can help you distinguish between results.

Below you can see an example of clarifying question. The user has asked for "history of america", and the system asked the clarifying question: "Which historical period are you interested in?"

So, in this HIT, you should play the role of system and ask clarifying questions for the given queries.
If you find it difficult to figure out different facets of a query by scanning the results, you can use the following trick for a better understanding of different facets:

- Open Google and type the given query in the search input box.
- As soon as you enter the last letter of the query and hit the space key, you will see Google’s suggestions for the next word in the query.
- Those words usually show how the same query can be completed and cover different facets of the same query.
- Take the following figure as an example:

![Google Search Suggestion Example](image)

- As you see, the suggestions made by Google can give you a good idea on various facets of the query "history of america". For example, one might be interested in buying a book on this topic, while someone might be looking for a documentary.
- This observation, together with scanning the results can give you an overview of different facets which should be helpful for generating clarifying questions.
Query:
dangers of asbestos

Questions

Write six questions that would help you clarify the intention behind the given query:

Question 1:
Write a question to clarify user’s intention behind the given query...

Question 2:
Write a question to clarify user’s intention behind the given query...

Question 3:
Write a question to clarify user’s intention behind the given query...

Question 4:
Write a question to clarify user’s intention behind the given query...

Question 5:
Write a question to clarify user’s intention behind the given query...

Question 6:
Write a question to clarify user’s intention behind the given query...
Question Verification and Facet Linking

• Two main concerns:
  • Precision: how is the quality of the collected clarifying questions?
  • Recall: are all facets addressed by at least one clarifying question?

• Two expert annotators:
  • Marked invalid and duplicate questions.
  • Linked questions to the facets they found relevant.
  • Facets with no questions: generated new questions relevant to them.
Would you like to hear about President Obama's family tree?
How many generations would you like me to list from President Obama's family tree?
Are you interested in seeing Barack Obama's family?
Would you like to know who is currently alive from President Obama's family tree?
What would you like to know about Obama's heritage?
Would you like to know about Obama's ancestors?
Which family members of Obama are you interested in learning more about?
Which specific details would you like to learn about Obama's heritage?
Do you want to know more about Obama's parents?
Would you like to know Barack Obama's genealogy?
An offline evaluation methodology

• Let \( T = \{t_1, t_2, \ldots, t_n\} \) be the set of topics (queries).

• A collection of facet sets: \( F = \{f_1, f_2, \ldots, f_n\} \)

• \( f_i = \{f_1^i, f_2^i, \ldots, f_{m_i}^i\} \) includes all defined facets for topic \( t_i \).

• A collection of clarifying question sets: \( Q = \{q_1, q_2, \ldots, q_n\} \)

• With \( q_i = \{q_1^i, q_2^i, \ldots, q_{m_i}^i\} \) including all clarifying questions relevant to topic \( t_i \).

• An offline evaluation requires defining \( A(t, f, q) \rightarrow a. \)
Query
website design hosting

User's Actual Information Need
Find sites that offer free DNS hosting.

IMPORTANT! Please make sure you have read the instructions carefully before submitting the HIT. You can expand the instructions on top of the page.

Question & Answer

Answer the following question in relation to the query and given information need:

Question: Are you interested in free websites or would you like your own domain?

Answer:
Write your answer to the above question in relation to the query and information need...

☐ I cannot answer the question based on the given information need.

Comment (optional)
Please leave a comment if you have any suggestions, questions or remarks
Quality Check

• Regular quality checks on the collected answers.
• Manual checks on 10% of submissions per worker.
  • If any invalid answer was observed, we then checked all the submissions of the corresponding worker.
• Invalid answers were removed and workers banned from future tasks.
• Disabled copy/paste feature.
• Monitored keystrokes.
Qulac: Questions for Lack of Clarity

Qulac has two meanings in Persian:
- blizzard
- wonderful or masterpiece

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># topics</td>
<td>198</td>
</tr>
<tr>
<td># faceted topics</td>
<td>141</td>
</tr>
<tr>
<td># ambiguous topics</td>
<td>57</td>
</tr>
<tr>
<td># facets</td>
<td>762</td>
</tr>
<tr>
<td>Average facet per topic</td>
<td>3.85 ± 1.05</td>
</tr>
<tr>
<td>Median facet per topic</td>
<td>4</td>
</tr>
<tr>
<td># informational facets</td>
<td>576</td>
</tr>
<tr>
<td># navigational facets</td>
<td>185</td>
</tr>
<tr>
<td># questions</td>
<td>2,639</td>
</tr>
<tr>
<td># question-answer pairs</td>
<td>10,277</td>
</tr>
<tr>
<td>Average terms per question</td>
<td>9.49 ± 2.53</td>
</tr>
<tr>
<td>Average terms per answer</td>
<td>8.21 ± 4.42</td>
</tr>
</tbody>
</table>
Learning to ask clarifying questions
Question Retrieval

• Task: Given a topic and a context (question-answer history), retrieve clarifying questions.

• Desired objective: high recall

• Approaches:
  • Term matching retrieval models: language models, BM25, RM3 (query expansion)
  • Learning to rank: LambdaMART, RankNet, neural ranking models (e.g., BERT)
## Question Retrieval

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>Recall@10</th>
<th>Recall@20</th>
<th>Recall@30</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>0.6714</td>
<td>0.5917</td>
<td>0.6946</td>
<td>0.7076</td>
</tr>
<tr>
<td>BM25</td>
<td>0.6715</td>
<td>0.5938</td>
<td>0.6848</td>
<td>0.7076</td>
</tr>
<tr>
<td>RM3</td>
<td>0.6858</td>
<td>0.5970</td>
<td>0.7091</td>
<td>0.7244</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.7218</td>
<td>0.6220</td>
<td>0.7234</td>
<td>0.7336</td>
</tr>
<tr>
<td>RankNet</td>
<td>0.7304</td>
<td>0.6233</td>
<td>0.7314</td>
<td>0.7500</td>
</tr>
<tr>
<td>BERT-LeaQuR</td>
<td>0.8349*</td>
<td>0.6775*</td>
<td>0.8310*</td>
<td>0.8630*</td>
</tr>
</tbody>
</table>
Question Selection

• Task: selecting a clarifying question that leads to retrieval improvement
• Objective: high precision (in retrieval)
• Approaches:
  • Query performance prediction (QPP): predicting the retrieval performance after asking each question (without answer) and selecting the one with the highest QPP.
  • Learning to rank: defining a set of features for ranking questions. The features include QPP, similarity to the topic, similarity to the context, etc.
  • Neural ranking models: learning to rank with representation learning (e.g., BERT)
Question Selection

<table>
<thead>
<tr>
<th>Method</th>
<th>Qulac-T Dataset</th>
<th></th>
<th>Qulac-F Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR</td>
<td>P@1</td>
<td>nDCG@1</td>
<td>nDCG@5</td>
</tr>
<tr>
<td>OriginalQuery</td>
<td>0.2715</td>
<td>0.1842</td>
<td>0.1381</td>
<td>0.1451</td>
</tr>
<tr>
<td>σ-QPP</td>
<td>0.3570</td>
<td>0.2548</td>
<td>0.1960</td>
<td>0.1938</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.3558</td>
<td>0.2537</td>
<td>0.1945</td>
<td>0.1940</td>
</tr>
<tr>
<td>RankNet</td>
<td>0.3573</td>
<td>0.2562</td>
<td>0.1979</td>
<td>0.1943</td>
</tr>
<tr>
<td>NeuQS</td>
<td><strong>0.3625</strong>*</td>
<td><strong>0.2664</strong>*</td>
<td><strong>0.2064</strong>*</td>
<td><strong>0.2013</strong>*</td>
</tr>
</tbody>
</table>

Asking only one good question improves the performance by over 100%.
## Case Study

<table>
<thead>
<tr>
<th>Query</th>
<th>Facet Description</th>
<th>Selected Question</th>
<th>User’s Answer</th>
<th>ΔMRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog heat</td>
<td>What is the effect of excessive heat on dogs?</td>
<td>Would you like to know how to care for your dog during heat?</td>
<td>No, I want to know what happens when a dog is too hot.</td>
<td>−0.86</td>
</tr>
<tr>
<td>sit and reach test</td>
<td>How is the sit and reach test properly done?</td>
<td>Do you want to know how to perform this test?</td>
<td>Yes, I do.</td>
<td>−0.75</td>
</tr>
<tr>
<td>alexian brothers hospital</td>
<td>Find Alexian Brothers hospitals.</td>
<td>Are you looking for our schedule of classes or events?</td>
<td>No, I don’t need that.</td>
<td>−0.54</td>
</tr>
<tr>
<td>east ridge high school</td>
<td>Information about the sports program at East Ridge High School in Clermont, Florida</td>
<td>What information about East Ridge High School are you looking for?</td>
<td>I’m looking for information about their sports program.</td>
<td>+0.96</td>
</tr>
<tr>
<td>euclid</td>
<td>Find information on the Greek mathematician Euclid.</td>
<td>Do you want a biography?</td>
<td>Yes.</td>
<td>+0.93</td>
</tr>
<tr>
<td>rocky mountain news</td>
<td>Who are the sports reporters for the Rocky Mountain News?</td>
<td>Would you like to read recent news about the Rocky Mountain News?</td>
<td>No, I just want a list of the reporters who write the sports for the Rocky Mountain News.</td>
<td>+0.88</td>
</tr>
</tbody>
</table>

Negative answer; new information. Retrieval model fails.
<table>
<thead>
<tr>
<th>Query</th>
<th>Facet Description</th>
<th>Selected Question</th>
<th>User’s Answer</th>
<th>ΔMRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog heat</td>
<td>What is the effect of excessive heat on dogs?</td>
<td>Would you like to know how to care for your dog during heat?</td>
<td>No, I want to know what happens when a dog is too hot.</td>
<td>-0.86</td>
</tr>
<tr>
<td>sit and reach test</td>
<td>How is the sit and reach test properly done?</td>
<td>Do you want to know how to perform this test?</td>
<td>Yes, I do.</td>
<td>-0.75</td>
</tr>
<tr>
<td>alexian brothers hospital</td>
<td>Find Alexian Brothers hospitals.</td>
<td>Are you looking for our schedule of classes or events?</td>
<td>No, I don’t need that.</td>
<td>-0.54</td>
</tr>
<tr>
<td>east ridge high school</td>
<td>Information about the sports program at East Ridge High School in Clermont, Florida</td>
<td>What information about East Ridge High School are you looking for?</td>
<td>I’m looking for information about their sports program.</td>
<td>+0.96</td>
</tr>
<tr>
<td>euclid</td>
<td>Find information on the Greek mathematician Euclid.</td>
<td>Do you want a biography?</td>
<td>Yes.</td>
<td>+0.93</td>
</tr>
<tr>
<td>rocky mountain news</td>
<td>Who are the sports reporters for the Rocky Mountain News?</td>
<td>Would you like to read recent news about the Rocky Mountain News?</td>
<td>No, I just want a list of the reporters who write the sports for the Rocky Mountain News.</td>
<td>+0.88</td>
</tr>
</tbody>
</table>
Future Directions

• Utilizing positive and negative feedback for document retrieval.
• Joint modeling of question retrieval and selection.
• Question generation.
• Determining the number of questions to ask based on the system’s confidence.
• Explore other ways of evaluating a system:
  • Conversation turns;
  • Retrieval performance.
Conclusions

• Asking clarifying questions in open-domain information-seeking conversations.

• **Qulac**: a collection for automatic offline evaluation of asking clarifying questions for conversational IR.

• A simple yet effective retrieval framework.

• Asking only **one** good question improves the performance by over 100%!

• More improvement for:
  • Shorter queries;
  • Ambiguous queries.
Questions?


Thanks to SIGIR for the student travel grant!