TopiOCQA: Open-domain Conversational Question Answering with Topic Switching

Vaibhav Adlakha[1,4] **Shehzaad Dhuliawala**[2]
Kaheer Suleman[3] **Harm de Vries**[4] **Siva Reddy**[1,5]
[1]Mila, McGill University, Canada [2]ETH Zürich, Switzerland [3]Microsoft Montréal, Canada
[4]ServiceNow Research, Canada [5]Facebook CIFAR AI Chair, Canada
{vaibhav.adlakha,siva.reddy}@mila.quebec

Abstract

In a conversational question answering scenario, a questioner seeks to extract information about a topic through a series of interdependent questions and answers. As the conversation progresses, they may switch to related topics, a phenomenon commonly observed in information-seeking search sessions. However, current datasets for conversational question answering are limiting in two ways: 1) they do not contain topic switches; and 2) they assume the reference text for the conversation is given, that is, the setting is not open-domain. We introduce TopiOCQA (pronounced Tapioca), an open-domain conversational dataset with topic switches based on Wikipedia. TopiOCQA contains 3,920 conversations with information-seeking questions and free-form answers. On average, a conversation in our dataset spans 13 question-answer turns and involves four topics (documents). TopiOCQA poses a challenging test-bed for models, where efficient retrieval is required on multiple turns of the same conversation, in conjunction with constructing valid responses using conversational history. We evaluate several baselines, by combining state-of-the-art document retrieval methods with neural reader models. Our best model achieves F1 of 55.8, falling short of human performance by 14.2 points, indicating the difficulty of our dataset. Our dataset and code are available at https://mcgill-nlp.github.io/topiocqa.

1 Introduction

People often engage in information-seeking conversations to discover new knowledge (Walton, 2019). In such conversations, a questioner (the seeker) asks multiple rounds of questions to an answerer (the expert). As the conversation proceeds, the questioner becomes inquisitive of new but related topics based on the information provided in the answers (Stede and Schlangen, 2004). Such topic switching behaviour is natural in information-seeking conversations and is commonly observed when people seek information through search engines (Spink et al., 2002).

According to Spink et al., people switch from one to ten topics with a mean of 2.11 topic switches per search session. For example, a person can start a search session about *tennis*, and then land on *Roger Federer*, and after learning a bit about him may land on his country *Switzerland*, and spend more time learning about other *Swiss athletes*. Thanks to tremendous progress in question answering research (Rogers et al., 2021), we are coming close to enabling information-seeking conversations with machines (as opposed to just using keywords-based search). In order to realize this goal further, it is crucial to construct datasets that contain information-seeking conversations with topic switching, and measure progress of conversational models on this task, the two primary contributions of this work.

In the literature, a simplified setting of information-seeking conversation known as conversational question answering (CQA) has been deeply explored (Choi et al., 2018; Reddy et al., 2019). In this task, the entire conversation is based on a given reference text of a topic/entity. While the CQA task is challenging, it still falls short of the real-world setting, where the reference text is not known beforehand (first limitation) and the conversation is not restricted to a single topic (second limitation).

Qu et al. (2020) and Anantha et al. (2021) have attempted to overcome the first limitation by adapting existing CQA datasets to the open-domain setting. They do so by obtaining context-independent rewrites of the first question to make the question independent of the reference text. For example, if the reference text is about *Augusto*

Transactions of the Association for Computational Linguistics, vol. 10, pp. 468–483, 2022. https://doi.org/10.1162/tacl_a_00471
Action Editor: Hua Wu. Submission batch: 11/2021; Revision batch: 12/2021; Published 4/2022.

Pinochet and the conversation starts with a question *"Was he known for being intelligent?"*, the question is re-written to *"Was Augusto Pinochet known for being intelligent?"*. However, as the entire question sequence in the conversation was collected with a given reference text of a topic, all the turns still revolve around a single topic.

In this work, we present **TopiOCQA**[1]—**Topi**c switching in **O**pen-domain **C**onversational **Q**uestion **A**nswering—a large-scale dataset for information-seeking conversations in open-domain based on the Wikipedia corpus. We consider each Wikipedia document to be a separate topic. The conversations in TopiOCQA start with a real information-seeking question from Natural Questions (Kwiatkowski et al., 2019) in order to determine a seed topic (document), and then the questioner may shift to other related topics (documents) as the conversation progresses.[2] Throughout the conversation, the questioner is never shown the content of the documents (but only the main title and section titles) to simulate an information-seeking scenario, whereas the answerer has full access to the content along with the hyperlink structure for navigation. In each turn, both questioner and answerer use free-form text to converse (as opposed to extractive text spans as is common for an answerer in many existing datasets).

Figure 1 shows an example of a conversation from our dataset. The first question leads to the seed topic *Byzantine Empire*, and after two turns switches to *Mehmed the Conqueror* in Q_4, based on part of the answer (A_3) that contains reference to *Mehmed*. Note that the answers A_1, A_3, and A_4 are free-form answers that do not occur as spans in either the seed document or the follow up document. The topic then switches to *Anatolia* based on part of the previous answer (A_4). The topics change in further turns to *Turkey* and *Ankara*. Because of the conversational nature, TopiOCQA contains questions rife with complex coreference phenomena, for instance, Q_9 relies on entities mentioned in A_7, A_8 and Q_1.

TopiOCQA contains 3,920 conversations and 50,574 QA pairs, based on Wikipedia corpus of 5.9 million documents. On average, a conversation has 13 question-answer turns and involves 4 topics. Twenty-eight percent of turns in our dataset require retrieving a document different from the previous turn. To the best of our knowledge, TopiOCQA is the first open-domain information-seeking CQA dataset that incorporates topical changes, along with other desirable properties (see Table 1).

To investigate the difficulty of the TopiOCQA dataset, we benchmark several strong retriever-reader neural baselines, considering both sparse and dense retrievers, as well as extractive and generative readers (Karpukhin et al., 2020; Izacard and Grave, 2021). Inspired by previous work, we explore two ways to represent the question: (1) concatenating the entire conversation history (Qu et al., 2020), and (2) self-contained rewrites of the conversational question (Anantha et al., 2021). The best performing model—Fusion-in-

Q_1: **when was the byzantine empire born what was it originally called?**
A_1: 5th century AD and was called Eastern Roman Empire, or Byzantium
Topic: Byzantine Empire

. . . . .

Q_3: **which battle or event marked the fall of this empire?**
A_3: A six-year-long civil war followed by attack from Sultan Mehmed's army
Topic: Byzantine Empire

Q_4: **did he conquer other territories as well?**
A_4: Yes. Anatolia and in Southeast Europe as far west as Bosnia
Topic: Mehmed the Conqueror

Q_5: **where is the first area located in present day terms?**
A_5: Turkey
Topic: Anatolia

. . . . .

Q_7: **what is the present day capital of the country?**
A_7: Ankara
Topic: Turkey

Q_8: **can you name some of the other major cities here?**
A_8: Istanbul
Topic: Turkey

Q_9: **were any of these cities associated with the first empire you were discussing?**
A_9: The Ottomans made the city of Ankara the capital first of the Anatolia Eyalet and then the Angora Vilayet
Topic: Ankara

Figure 1: A conversation from TopiOCQA. Our dataset has information-seeking questions with free-form answers across multiple topics (documents). The consecutive turns from the same topic (document) have been excluded for brevity.

[1] TopiOCQA is pronounced as Tapioca.

[2] A portion of the training data also contains conversations where the questioner asks the first question given a seed topic.

Dataset	Multi–turn	Open–domain	Free–form answers	Information–seeking questions	Topic Switching
TopiOCQA (ours)	✓	✓	✓	✓	✓
QReCC (Anantha et al., 2021)	✓	✓	✓	⚠	⚠
OR-QuAC (Qu et al., 2020)	✓	✓	✗	✓	✗
CoQA (Reddy et al., 2019)	✓	✗	✓	✗	✗
QuAC (Choi et al., 2018)	✓	✗	✗	✓	✗
NarrativeQA (Kočiský et al., 2018)	✗	✓	✓	✓	✗
Natural Questions (Kwiatkowski et al., 2019)	✗	✓	✗	✓	✗
SQuAD 2.0 (Rajpurkar et al., 2018)	✗	✗	✗	✗	✗

Table 1: Comparison of TopiOCQA with other QA datasets. TopiOCQA incorporates topical changes, along with several best practices of previous datasets. ⚠ represents that only a proportion of dataset satisfies the property.

Decoder (Izacard and Grave, 2021) trained on concatenated conversation history—is 14.2 F1 points short of human performance, indicating significant room for improvement. We also evaluate GPT-3 to estimate the performance in a closed-book zero-shot setting, and its performance is 38.2 F1 points below the human performance.

2 Related Work

2.1 Open-Domain Question Answering

In open-domain question answering, a model has to answer natural language questions by retrieving relevant documents. This can be considered as a simplified setting of open-domain CQA, where the conversation is limited to just one turn. Several datasets have been proposed for this task. On one hand, reading comprehension datasets like SQuAD (Rajpurkar et al., 2016, 2018), which consist of (question, document, answer) triplets, have been adapted for the task by withholding access to the document (Chen et al., 2017). While these datasets have been helpful in spurring modelling advances, they suffer from an annotator bias because they were not collected in an information-seeking setup. That is, annotators had access to the target answer and its surrounding context and therefore formulated questions that had a high lexical overlap with the answer (Jia and Liang, 2017). On the other hand, Web-search based datasets do not suffer from such artefacts because they are curated from real search engine queries. The WikiQA (Yang et al., 2015) and MS Marco (Nguyen et al., 2016) datasets contain queries from the Bing search engine, whereas Natural Questions (Kwiatkowski et al., 2019) contain queries from the Google search engine.

Models for open-domain QA often follow a two-stage process: (1) A retriever selects a small collection of documents relevant to the question from a big corpus (e.g., Wikipedia), (2) a reader extracts or generates an answer from the selected documents. While classical approaches rely on counting-based bag-of-words representations like TF-IDF or BM25 (Chen et al., 2017; Wang et al., 2018; Yang et al., 2019), more recent deep learning approaches learn dense representations of the questions and document through a dual-encoder framework (Lee et al., 2019; Karpukhin et al., 2020). In such learned retriever setups, document retrieval is done efficiently using Maximum Inner Product Search (MIPS, Shrivastava and Li, 2014).

2.2 Conversational Question Answering (CQA)

CQA extends the reading comprehension task from a single turn to multiple turns. Given a reference document, a system is tasked with interactively answering a sequence of information-seeking questions about the corresponding document. This conversational extension leads to novel challenges in modeling linguistic phenomena such as anaphora (referencing previous turns) and ellipsis (omitting words from questions), as well as in performing pragmatic reasoning. Large-scale conversational datasets such as CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) have facilitated much of the research in this area. These datasets differ along several dimensions, two of which are (1) CoQA has short free-form answers, whereas QuAC has long extractive span-based answers, and (2) unlike CoQA, QuAC is collected in a simulated information-seeking scenario.

Models for CQA have used simple concatenation of the question-answer history (Zhu et al., 2019), history turn selection (Qu et al., 2019a,b), and question-rewrites (Vakulenko et al., 2021). For question-rewriting, a different module is trained on self-contained rewrites of context-dependent questions. For example, a plausible rewrite of Q_8 (Figure 1) is *''can you name some of the major cities in Turkey apart from*

Ankara?''. The re-written question is then answered using open-domain QA systems. Two popular question-rewriting datasets for training this module are (1) CANARD (Elgohary et al., 2019), which contains re-writes of 50% of QuAC, and (2) QReCC (Anantha et al., 2021), which contains rewrites of the entire QuAC dataset and a small portion from other sources.

2.3 Open-Domain CQA

In this work, we focus on constructing a challenging benchmark for open-domain CQA. The open-domain aspect requires systems to answer questions *without* access to a reference document. The conversational aspect enables users to ask multiple related questions, which can, in principle, span several different topics. With TopiOCQA, we introduce the first open-domain CQA dataset that explicitly covers such topical switches.

Previous datasets for this task re-purpose existing CQA datasets. The OR-QuAC dataset (Qu et al., 2020) is automatically constructed from QuAC (Choi et al., 2018) and CANARD (Elgohary et al., 2019) by replacing the first question in QuAC with context-independent rewrites from CANARD. QReCC (Anantha et al., 2021) is a large-scale open-domain CQA and question rewriting dataset that contains conversations from QuAC, TREC CAsT (Dalton et al., 2020), and Natural Questions (NQ; Kwiatkowski et al., 2019). All the questions in OR-QuAC and 78% of questions in QReCC are based on QuAC. As conversations in QuAC were collected with a given reference document, the question sequences of these conversations revolve around the topic or entity corresponding to that document. Twenty-one percent of questions in QReCC are from NQ-based conversations. As NQ is not a conversational dataset, the annotators of QReCC use NQ to start a conversation. A single annotator is tasked with providing both follow-up questions and answers for a given NQ question. In contrast to QReCC, conversations in our dataset are collected in a simulated information-seeking scenario using two annotators (Section 3.3).

Deep learning models for this task have followed a similar retriever-reader setup as open-domain QA. Instead of a single question, previous works have explored feeding the entire conversation history (Qu et al., 2020), or a context independent re-written question (Anantha et al., 2021).

3 Dataset Collection

Each conversation in TopiOCQA is an interaction between two annotators—a *questioner* and an *answerer*. The details about the annotator selection are provided in Appendix A.

3.1 Seed Topics and Document Collection

The seed topics essentially drive the conversation. In order to make them interesting for annotators, we select the *good*[3] articles of Wikipedia as seed topics (around 35k) for the first turn, but use entire Wikipedia for later turns. We used the Wikipedia dump from 10/20/2020, which consists of 5.9 million documents. We used Wikiextractor[4] to extract the text. While pre-processing the Wikipedia documents, we retain the hyperlinks that refer to other Wikipedia documents, thus ensuring that we can provide all the documents requested by annotators (via hyperlinks) during the conversation.

3.2 Simulating Information-seeking Scenario

Information-seeking conversations are closer to the real-world if an information need can be simulated via the data collection interface. In TopiOCQA, we achieve this by withholding questioner's access to the full reference text of the document. The questioner can only see the metadata (main title and the section titles) of the Wikipedia documents, whereas the answerer can access the entire text of the documents. On finding the answer, the answerer highlights a contiguous span of text as rationale, and generates a free-form answer. The answerer also has the option to mark the question as *unanswerable*. The conversation history is visible to both the annotators.

As a conversation starting point, the first question is sampled from a subset of NQ (Kwiatkowski et al., 2019) since NQ contains genuine information-seeking questions asked on Google. We only sample those questions for which the answer is in our seed document pool. To increase the diversity of our dataset, we also allow the questioner to formulate the first question based on the provided seed topic entity for 28% of the conversations.

[3] Wikipedia Good articles.
[4] github:wikiextractor.

3.3 Enabling Topic-switching

The key feature of the interface is enabling topic switching via hyperlinks. For the answerer, the text of the document includes clickable hyperlinks to other documents. On clicking these links, the current document in the answerer's interface changes to the requested (clicked) document. This enables the answerer to search for answers in documents beyond the current one. The questioner can access the metadata of documents visited by the answerer and documents present in the rationale of the answers. For example, let us assume that given the seed document *Daniel Radcliffe* and the first question *''Where was Daniel Radcliffe born?''*, the answerer selects the *''Daniel Jacob Radcliffe was born in London on 23 July 1989''* span as rationale and provides *''London''* as the answer. If *London* is a hyperlink in the rationale span, then the metadata of both *Daniel Radcliffe* and *London* is available to the questioner to form the next question. If the next question is *''What is its population?''*, the answerer can switch the current document from *Daniel Radcliffe* to *London* by clicking on the hyperlink, and can then find and provide the answer. The conversation up till this point involves two topics: *Daniel Radcliffe* and *London*. We also provide easy navigation to previously visited documents for both the annotators. This interface design (Figure 8) ensures that information about the new topic is semantically connected to topics of the previous turns, similar to natural human-human conversations (Sacks and Jefferson, 1995).

3.4 Additional Annotations

To account for multiple valid answers, we collected three additional annotations for answers of conversations in evaluation sets (development and test splits). For this task, at any turn, the annotator can see all the previous questions and original answers. Showing original answers of previous turns is important in a conversational setting as the subsequent questions can potentially depend on them. We also provide the list of documents corresponding to previous turns of the original conversation. This ensures that the current annotator has all the information the original answerer had while providing the answer. Similar to the answerer, the annotator then provides the rationale and the answer, or marks the question as *unanswerable*.

Dataset	Train	Dev	Test	Overall
# Turns	45,450	2,514	2,502	50,466
# Conversations	3,509	205	206	3920
# Tokens / Question	6.91	6.89	7.11	6.92
# Tokens / Answer	11.71	11.96	12.27	11.75
# Turns / conversation	13	12	12	13
# Topics / conversation	4	4	4	4

Table 2: Dataset statistics of TopiOCQA.

4 Dataset Analysis

We collected a total of 3,920 conversations, consisting of 50,466 turns. The annotators were encouraged to complete a minimum of 10 turns. Conversations with fewer than 5 turns were discarded. We split the data into train, development, and test splits.

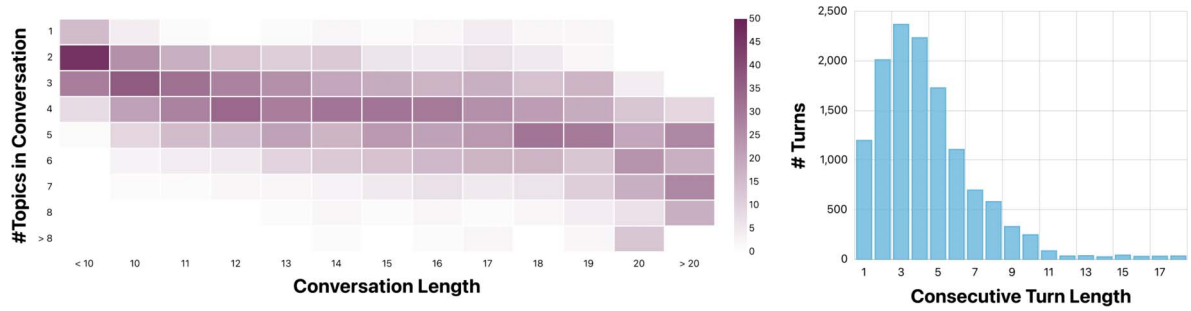
Table 2 reports simple statistics of the dataset splits. On average, a conversation in TopiOCQA has 13 question-answer turns and is based on 4 documents. Our dataset differs from other conversational question-answering datasets by incorporating topic switches in the conversation.

4.1 Topic Switching

Before we start our analysis, let us first define the notion of a topic switch in TopiOCQA. Recall that answers are based on Wikipedia articles, where each document consists of several sections. While one can argue that a topic switch occurs when the answer is based on a different section of the *same* document, we opt for a more conservative notion and define a switch of topic if the answer is based on a *different* Wikipedia document.

Number of Topics vs Conversation Length We begin our analysis by investigating how the number of topics varies with the conversation length. In Figure 2(a) we show a heat-map of the number of topics for each conversation length, where each column is normalized by the number of conversations of that length. We observe that longer conversations usually include more topics. Most 10-turn conversations include 3 topics, 14-turn conversations include 4 topics, and 18-turn conversations include 5 topics. The conversations with fewer than 10 turns mostly include just 2 topics.

Topic Flow in Conversation Next, we examine how often consecutive questions stay within the



(a) (b)

Figure 2: Analysis of the topic switches in TopiOCQA. In (a) we show the distribution of the number of topics (in percentage) for each conversation length. Longer conversations typically include more topics. In (b) we show a histogram of the topic lengths, illustrating that usually 3–4 consecutive questions stay within the same topic.

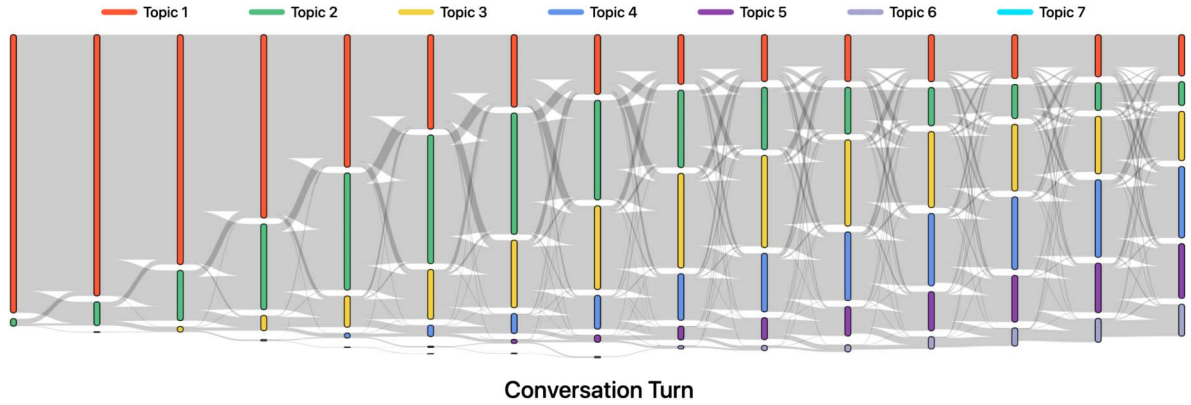


Figure 3: A flow diagram of topic switches over conversations up to 15 turns. There are complex interactions between the topics, especially later in the conversation.

same topic. To do so, we first cluster conversations into sequences of turns for which all answers are from the same document. Then, we count how many turns belong to topic clusters of a particular length. Figure 2(b) shows the distribution of topic lengths. The mode of the distribution is 3, signifying that annotators usually ask 3 questions about the same topic before switching. Asking 2 or 4 consecutive questions on the same topic is also frequently observed. However, we rarely see more than 10 consecutive turns on the same topic.

We also analyze the flow of topics throughout the conversation. Do annotators always introduce new topics or do they also go back to old ones? Figure 3 depicts a flow diagram of topics in conversations up to 15 turns. Note that we have indexed topics according to their first occurrence in the conversation. We can see that the majority of switches introduce new topics, but also that more complex topic switching emerges in later turns. Specifically, we see that, from sixth turn onwards, questioners frequently go back one or two topics in the conversation. Overall, this diagram suggests that there are complex interactions among the topics in the conversation.

Qualitative Assessment of Topic Switching In order to understand the nature of a topic switch, inspired from Stede and Schlangen (2004), we classify questions into three types: `ask-generic` refers to general open-ended questions, `ask-specific` questions ask about a specific attribute or detail of a topic, and `ask-further` is a question type that seeks additional details of an attribute discussed in one of the previous turns. Table 4 shows examples of each type for questions in the same conversation. We consider three types of turns for our evaluation. If the answer document of the turn is same as the previous turn, we refer to it as `no-switch`. If a topic switch has happened, and the answer document is present in one of the previous turns, it is considered to be `switch-to-old`. The final category, `switch-to-new` refers to turns where current

Question Type	Avg Answer length
ask-generic	22.43
ask-specific	11.38
ask-further	11.23

Table 3: Average answer length of different question types. Generic questions tend to have longer answers.

Turn type	Question type	Conversation turn
no-switch	ask-generic	Q: who is mariah carey? A: An American singer songwriter and actress Topic: Mariah Carey
no-switch	ask-specific	Q: name one of her famous songs. A: Oh Santa! Topic: Mariah Carey
switch-to-new	ask-specific	Q: how was it received? A: There were mixed reviews Topic: Oh Santa!
switch-to-old	ask-specific	Q: is she married? A: Yes Topic: Mariah Carey
no-switch	ask-further	Q: to whom? A: Tommy Mottola Topic: Mariah Carey

Table 4: Examples of various turn types and question types in a conversation. Random samples of each turn type are manually annotated with one of the question types.

answer document has not been seen in the conversation before. These different types of topic switches are also illustrated in Table 4.

We sample 50 turns of each type, and manually label them with one of the three question types. Figure 4 shows the results of our evaluation. ask-specific is the most common question type across all types of turns, indicating that most of the questions in the dataset focus on specific attributes of a topic. ask-generic has a much higher proportion in switch-to-new turn types, indicating that it is more likely to see generic questions in turns that introduce a new topic in the conversation, compared to other turn types. ask-further has almost equal proportion in no-switch and switch-to-old, with switch-to-old being slightly higher. ask-further is not observed in switch-to-new as follow-up questions are generally not possible without the topic being discussed in any of the previous turns.

We also look at average answer length of answers of all three question types (Table 3). Unsurprisingly, ask-generic has a much higher answer length compared to other types, presumably due to the open-ended nature of the question.

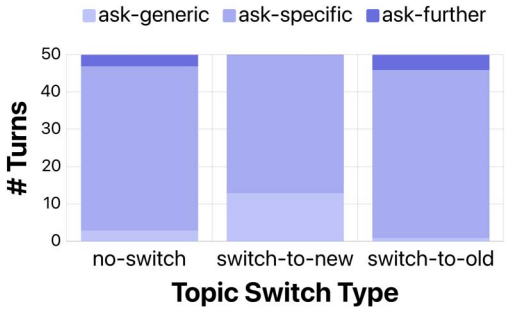


Figure 4: Distribution of various question types for each turn type. Questions asking about specific attributes are most common. Generic questions are likely to be observed when switching to a new topic.

5 Experimental Setup

The task of open-domain information-seeking conversation can be framed as follows. Given previous questions and ground truth answers $\{q_1, a_1, q_2, a_2, \ldots, q_{i-1}, a_{i-1}\}$ and current question q_i, the model has to provide the answer a_i. This can be considered as an *oracle* setting, as the gold answers of previous questions are provided. The models can optionally use a corpus of documents $C = \{d_1, d_2, \ldots, d_N\}$.

5.1 Models

We consider models from two categories, based on whether they use the document corpus or not. The *closed-book* models use just the question-answer pairs, whereas *open-book* models use the document corpus, along with question-answer pairs. We now describe the implementation and technical details of both classes of models.

5.1.1 Closed-book

Large-scale language models often capture a lot of world knowledge during unsupervised pre-training (Petroni et al., 2019; Roberts et al., 2020). These models, in principle, can answer questions without access to any external corpus. We consider GPT-3 (Brown et al., 2020)—an autoregressive language model with 175 billion parameters, and evaluate it on TopiOCQA. The input to GPT-3 is a prompt[5] followed by previous question-answer pairs and the current question. Because GPT-3 is

[5] beta.openai.com/examples/default-qa.

Q_1: who is lead singer of rage against the machine? A_1: Zack de la Rocha Q_2: when was it formed? A_2: 1991 Q_3: was it nominated for any award?
ORIGINAL: was it nominated for any award
ALLHISTORY: who is lead singer of rage against the machine [SEP] Zack de la Rocha [SEP] when was it formed? [SEP] 1991 [SEP] was it nominated for any award
REWRITES: was rage against the machine nominated for any award

Figure 5: A partial conversation and different question representations of Q_3. The REWRITES representation is an example, not the output of our *QR* module.

never explicitly exposed to any training examples, this can be considered as a *zero-shot* setting.

5.1.2 Open-book

We build on state-of-the-art QA models that adapt a two step retriever-reader approach. For the retriever, we consider BM25 (Robertson et al., 1995) and DPR Retriever (Karpukhin et al., 2020). Given a query, BM25 ranks the documents based on a bag-of-words scoring function. On the other hand, DPR learns dense vector representations of document and query, and uses the dot product between them as a ranking function.

We consider two types of neural readers. (1) DPR Reader (Karpukhin et al., 2020), which reranks the retrieved passages and selects a span from each document independently. The span with highest span score is chosen as the answer. (2) Fusion-in-Decoder (FiD; Izacard and Grave, 2021), which encodes all retrieved passages independently, and then jointly attends over all of them in the decoder to generate the answer.

For these models, we consider three different question representations for question at n^{th} turn of the conversation (q_n). Figure 5 shows an example of different question representations for the third question (Q_3) of a conversation.

- **ORIGINAL**: This serves as a naive baseline where just the current question q_n is passed to the model.
- **ALLHISTORY**: The question is represented as q_1 [SEP] a_1 [SEP] q_2 [SEP] a_2 [SEP] ... [SEP] q_{n-1} [SEP] a_{n-1} [SEP] q_n. When constrained by the encoder input sequence length, we retain the first turn and as many turns prior to the current turn as possible, that is, k is chosen such that q_1 [SEP] a_1 [SEP] q_{n-k} [SEP] a_{n-k} [SEP] ... [SEP] q_{n-1} [SEP] a_{n-1} [SEP] q_n satisfies encoder input limits.
- **REWRITES**: Given a query-rewriting module QR, let $q'_n = QR(q_1, a_1, \ldots, q_{n-1}, a_{n-1}, q_n)$ denote the decontextualized question, conditioned on the conversation history. q'_n is then passed to the model.

Wang et al. (2019) observed that fixed-length text segments from documents are more useful than full documents in both retrieval and final QA accuracy. Hence, we split a Wikipedia document into multiple text blocks of at least 100 words, while preserving section and sentence boundaries. These text blocks, augmented with the metadata (main title and section title) are referred to as *passages*. This resulted in 25.7 million passages, which act as basic units of retrieval. To form question-passage pairs for training DPR Retriever, we select the passage from gold answer document that contains the majority of rationale span.

Following the original works, we use BERT (Devlin et al., 2019) for DPR (both Retriever and Reader) and T5 (Raffel et al., 2020) for FiD as base models. Because DPR Reader requires a span from passage for each training example, we heuristically select the span from the gold passage that has the highest lexical overlap (F1 score) with the gold answer. For the query-rewriting module QR, we fine-tune T5 model on rewrites of QReCC (Anantha et al., 2021), and use that to generate the rewrites for TOPIOCQA. We refer the reader to Appendix B for more details. The hyperparameters for all models are mentioned in Appendix C.

5.2 Evaluation Metrics

Following Choi et al. (2018) and Reddy et al. (2019), we use *exact match (EM)* and *F1* as evaluation metrics for TOPIOCQA.

To compute human and system performance in the presence of multiple gold annotations, we follow the evaluation process similar to Choi et al. (2018) and Reddy et al. (2019). Given n human answers, human performance on the task is determined by considering each answer as prediction

Model	Question Rep	Dev		Test	
		EM	F1	EM	F1
Human		**40.2**	**70.1**	**40.3**	**70.0**
GPT-3		12.4	33.4	10.4	31.8
BM25 + DPR Reader	ORIGINAL	7.1	12.8	7.2	13.0
	ALLHISTORY	13.6	25.0	13.8	25.2
	REWRITES	15.4	32.5	15.7	31.7
BM25 + FiD	ORIGINAL	10.1	21.8	10.5	22.6
	ALLHISTORY	24.1	37.2	23.4	36.1
	REWRITES	24.0	41.6	24.9	41.4
DPR Retriever + DPR Reader	ORIGINAL	4.9	14.9	4.3	14.9
	ALLHISTORY	21.0	43.4	19.4	41.1
	REWRITES	17.2	36.4	16.5	35.2
DPR Retriever + FiD	ORIGINAL	7.9	21.6	7.8	21.4
	ALLHISTORY	**33.0**	**55.3**	**33.4**	**55.8**
	REWRITES	23.5	44.2	24.0	44.7

Table 5: Overall performance of all model variants on TOPIOCQA development and test set.

and other human answers as the reference set. This results in n scores, which are averaged to give the final human performance score. The system prediction is also compared with n distinct reference sets, each containing $n-1$ human answers, and then averaged. For TOPIOCQA, $n = 4$ (the original answer and three additional annotations). Note that human performance is not necessarily an upper bound for the task, as document retrieval can potentially be performed better by the systems.

6 Results and Discussion

We report the end-to-end performance of all systems in Table 5. For open-book models, we also look at the performance of its constituents (retriever and reader). Table 6 reports the retrieval performance and Table 7 reports the reading comprehension performance of the readers, given the gold passage. Based on these results, we answer the following research questions.

How do the models compare against humans for TOPIOCQA*?*

We report model and human performance on development and test set in Table 5. Overall, model performance in all settings is significantly lower than the human performance. The best performing model (DPR Retriever + FiD using ALLHISTORY question representation) achieves 33.4 points EM and 55.8 points F1 on the test set, which falls short of human performance by 6.9 points and 14.2 points, respectively, indicating room for further improvement.

Model	Question Rep	Dev		Test	
		Top-20	Top-100	Top-20	Top-100
BM25	ORIGINAL	5.2	9.1	6.0	10.1
	ALLHISTORY	23.1	36.8	22.5	35.6
	REWRITES	32.5	49.2	33.0	47.4
DPR Retriever	ORIGINAL	9.9	16.5	10.0	15.3
	ALLHISTORY	**70.4**	**82.4**	**67.0**	**80.8**
	REWRITES	49.9	62.4	49.3	61.1

Table 6: Retrieval performance of all model variants on TOPIOCQA development and test set.

Model	Question Rep	Dev		Test	
		EM	F1	EM	F1
Extractive Bound		**47.7**	**81.1**	**47.3**	**81.0**
DPR Reader	ORIGINAL	27.1	51.4	25.5	50.4
	ALLHISTORY	29.7	54.2	28.0	52.6
	REWRITES	29.8	53.8	28.1	52.1
FiD	ORIGINAL	34.4	60.5	33.7	61.0
	ALLHISTORY	**38.3**	**65.5**	**37.2**	**64.1**
	REWRITES	34.5	61.9	35.3	62.8

Table 7: Reader performance of all model variants on TOPIOCQA development and test set when provided with the gold passage.

Which class of models perform better—Closed book or Open book?
GPT-3 is directly comparable to AllHistory variant of open-book models as it takes the entire conversation history as input. Apart from BM25 + DPR Reader, GPT-3 performs worse than all other AllHistory variants of open-book models. It achieves an F1 score of 31.8 on the test set, which is 24 points behind the best performing open-book model (DPR Retriever + FiD). We observe that GPT-3 often hallucinates many answers, a phenomenon commonly observed in literature (Shuster et al., 2021).

How does the performance of open-book models vary with various question representations?
For all open-book models, we fine-tune on three different question representations (Section 5). From the results in Table 5, we observe that the Original representation is consistently worse than others for all models. This highlights the importance of encoding the conversational context for TopiOCQA. Between AllHistory and Rewrites, we observe that AllHistory performs better with dense retriever (DPR Retriever), whereas Rewrites performs better with sparse retriever (BM25). To confirm that this performance difference in end-to-end systems stems from the retriever, we look at Top-20 and Top-100 retrieval accuracy of BM25 and DPR Retriever in Table 6. Indeed, AllHistory representation performs better than Rewrites for DPR Retriever but worse for BM25. As DPR Retriever is trained on TopiOCQA, it can probably learn how to select relevant information from the AllHistory representation, whereas for BM25, the non-relevant keywords in the representation act as distractors. The better performance of DPR Retriever over BM25 indicates that TopiOCQA requires learning task-specific dense semantic encoding for superior retrieval performance.

How much are the readers constrained due to retrieved results?
Table 6 shows retrieval results. In an end-to-end system, the reader takes as input the retrieved passages, which may or may not contain the gold passage. To get an estimate of reader performance independently from the retriever, we experiment with directly providing only the gold passage to the readers, instead of the retrieved ones. Table 7 shows the results. This can be seen as an ''Ideal Retriever'' setting, where the retriever always retrieves the correct passage as the top one. Although we observe significant gains over end-to-end systems for all models across all variants, the best model (FiD with AllHistory) still falls short of human performance by 3.1 points EM and 5.9 points F1 on the test set. These experiments indicate that while passage retrieval is a significant bottleneck for the task, technical advancements are needed for the readers as well.

While it is plausible to assume that DPR Reader is restricted in its performance due to its extractive nature, we show that this is not the case. We calculate the extractive upper bound for TopiOCQA (reported in Table 7) by selecting the span from the gold document with best F1 overlap with the ground truth answer. This bound is 47.3 points EM and 81.0 points F1, which essentially represents the best that any extractive model can do on this task. DPR Reader falls short of this upper bound by 19.2 points EM and 28.4 points F1.

7 Conclusion

We introduced TopiOCQA, a novel open-domain conversational question answering dataset with topic switching. In this work, we described our data collection effort, analyzed its topic switching behavior, and established strong neural baselines. The best performing model (DPR Retriever + FiD) is 6.9 points EM and 14.2 points F1 below human performance, suggesting that advances in modeling are needed. We hope our dataset will be an important resource to enable more research on conversational agents that support topic switches in information-seeking scenarios.

Acknowledgments

We would like to thank TAKT's annotators and Jai Thirani (Chief Data Officer) for contributing to TopiOCQA. We are grateful for constructive and insightful feedback from the anonymous reviewers. This work is supported by the following grants: MSR-Mila Grant, NSERC Discovery Grant on *Robust conversational models for accessing the world's knowledge*, and the Facebook CIFAR AI Chair. Shehzaad is partly supported by the IBM PhD Fellowship. We thank Compute Canada for the computing resources.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. https://doi.org/10.18653/v1/2021.naacl-main.44

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1985–1988, New York, NY, USA. https://doi.org/10.1145/3397271.3401206

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? Learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. https://doi.org/10.18653/v1/D19-1605

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. https://doi.org/10.18653/v1/2021.eacl-main.74

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1215

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. https://doi.org/10.18653/v1/2020.emnlp-main.550

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328. https://doi.org/10.1162/tacl_a_00023

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *CoCo@NIPS*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. https://doi.org/10.18653/v1/D19-1250

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. SIGIR '20, pages 539–548, New York, NY, USA.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1133–1136, New York, NY, USA.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 1391–1400, New York, NY, USA.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. https://doi.org/10.18653/v1/P18-2124

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. https://doi.org/10.18653/v1/D16-1264

Siva Reddy, Danqi Chen, and Christopher Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7(0):249–266. https://doi.org/10.1162/tacl_a_00266

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. https://doi.org/10.18653/v1/2020.emnlp-main.437

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Overview of the Third Text Retrieval Conference (TREC-3)*, pages 109–126.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Harvey Sacks and Gail Jefferson. 1995. Lectures on conversation, Volume 1. pages 289–331.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2321–2329, Montreal, Canada.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Amanda Spink, H. Cenk Ozmutlu, and Seda Ozmutlu. 2002. Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology*, 53(8):639–652. https://doi.org/10.1002/asi.10124

Manfred Stede and David Schlangen. 2004. Information-seeking chat: Dialogue management by topic structure.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 355–363, New York, NY, USA. https://doi.org/10.1145/3437963.3441748

Douglas Walton. 2019. *The New Dialectic: Conversational Contexts of Argument*. University of Toronto Press.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R3: Reinforced ranker-reader for open-domain question answering. In *AAAI*, pages 5981–5988.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. https://doi.org/10.18653/v1/D19-1599

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-4013

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1237

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. SDNet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

A Annotators Details

Each conversation in TopiOCQA is an interaction between two annotators, a *questioner* and an *answerer*. The annotators were selected from TAKT's in-house workforce, based on their English language proficiency and trained for the role of both questioner and answerer. The annotators are provided with the following guidelines.

Guidelines for the Questioner:

- The first question should be unambiguous and about the seed entity.
- The follow-up questions be contextualized and dependent on the conversation history whenever possible.
- Avoid using same words as in section titles of the document. For example, if the section title is *''Awards''*, a plausible question can be *''What accolades did she receive for her work?''*.
- The conversation should involve multiple documents (topics).

Guidelines for the Answerer:

- Based on the question, identify the relevant document and section.
- The answer should be based on the contents of the identified document.

- The rationale should be selected such that it justifies the answer.
- The answer should be a sub-string in rationale whenever possible. However, answers should be edited to fit the conversational context (adding *yes*, *no*), perform reasoning (e.g., counting), and so forth.
- Personal opinions should never be included.

After providing the guidelines and a few examples, the initial annotated conversations were manually inspected by the authors. The workers who provided low-quality annotations during this inspection phase were disqualified. The final workforce consisted of 15 workers, who provided annotations for the dataset over a period of two months. Random quality checks were performed by the authors and periodic feedback was given to the annotators throughout the data collection to maintain high quality of data. Figure 8 shows annotation interfaces for questioner and answerer. Figure 6 shows an example from the dataset.

We also implemented several real-time checks in the questioner's interface to encourage topic switching and use of co-reference, and to reduce the lexical overlap with the metadata of the document while forming the question.

B Query Rewriting

A query-rewriting module, QR, takes the current question and the conversation history as input $(q_1, a_1, \ldots, q_{n-1}, a_{n-1}, q_n)$ and provides a decontextualized rewritten question, q'_n, as the output. As we don't collect rewrites in TopiOCQA, we rely on other datasets to train our QR model. Two datasets that provide rewrites for information-seeking conversations are CANARD (Elgohary et al., 2019) and QReCC (Anantha et al., 2021). Due to its large-scale and diverse nature, we use QReCC to train our T5 model based QR module.

To rewrite the n^{th} question, the conversation history and the current question is given to model as q_1 `[SEP]` a_1 `[SEP]` q_2 `[SEP]` a_2 `[SEP]` ... `[SEP]` q_{n-1} `[SEP]` a_{n-1} `[SEP]` q_n. We train this model on QReCC dataset. On the test split of QReCC, our model achieves a BLEU score of 62.74 points. We use this model to generate rewrites for TopiOCQA in our experiments. Figure 7 shows a conversation from the dataset along with rewrites from this T5-based QR module.

Q_1: **when was the byzantine empire born what was it originally called?**
A_1: 5th century AD and was called Eastern Roman Empire, or Byzantium
Topic: Byzantine Empire

Q_2: **and when did it fall?**
A_2: 1453
Topic: Byzantine Empire

Q_3: **which battle or event marked the fall of this empire?**
A_3: A six-year-long civil war followed by attack from Sultan Mehmed's army
Topic: Byzantine Empire

Q_4: **did he conquer other territories as well?**
A_4: Yes. Anatolia and in Southeast Europe as far west as Bosnia
Topic: Mehmed the Conqueror

Q_5: **where is the first area located in present day terms?**
A_5: Turkey
Topic: Anatolia

Q_6: **who were the oldest known inhabitants of this region?**
A_6: Mesopotamian-based Akkadian Empire
Topic: Anatolia

Q_7: **what is the present day capital of the country?**
A_7: Ankara
Topic: Turkey

Q_8: **can you name some of the other major cities here?**
A_8: Istanbul
Topic: Turkey

Q_9: **were any of these cities associated with the first empire you were discussing?**
A_9: The Ottomans made the city of Ankara the capital first of the Anatolia Eyalet and then the Angora Vilayet
Topic: Ankara

Q_{10}: **what are some of the most famous landmarks in the second city?**
A_{10}: The obelisk, Valens Aqueduct, Column of Constantine, Church of the Saints Sergius and Bacchus
Topic: Istanbul

Q_{11}: **who was the first monument you mentioned dedicated to?**
A_{11}: UNANSWERABLE
Topic:

Q_{12}: **and who was the third monument name after?**
A_{12}: Roman emperor Constantine the Great
Topic: Column of Constantine

Q_{12}: **what is it made of?**
A_{12}: Porphyry and white marble
Topic: Column of Constantine

Q_{12}: **how tall is it?**
A_{12}: The column's top is 34.8 m above the present-day ground level but the original height of the monument as a whole would have been nearly 50 m tall
Topic: Column of Constantine

Figure 6: A full conversation from TopiOCQA.

We observe that while this QR module can resolve simple coreferences (Q_2 and Q_5), it struggles later in the conversation in the presence of multiple entities (*he* is resolved to *albus dumbledore*

$\mathtt{Q}_1$: **harry potter and the chamber of secrets full book summary**
$\mathtt{Q}'_1$: *harry potter and the chamber of secrets full book summary*
$\mathtt{A}_1$: The plot follows Harry's second year at Hogwarts School ... investigate the attacks.

$\mathtt{Q}_2$: **when was the book published?**
$\mathtt{Q}'_2$: *when was harry potter and the chamber of secrets published?*
$\mathtt{A}_2$: 2 June 1999

. . . . .

$\mathtt{Q}_4$: **does it have any special connection with any other book in this series?**
$\mathtt{Q}'_4$: *does the book have any special connection with any other book in the harry potter series?*
$\mathtt{A}_4$: Yes, "Chamber of Secrets" has many links with the sixth book of the series, "Harry Potter and the Half-Blood Prince".

$\mathtt{Q}_5$: **what is the storyline of that book?**
$\mathtt{Q}'_5$: *what is the storyline of the sixth book of the series, harry potter and the half-blood prince?*
$\mathtt{A}_5$: The novel explores the past of the boy wizard's nemesis, Lord Voldemort ... alongside his headmaster and mentor Albus Dumbledore.

$\mathtt{Q}_6$: **what does the headmaster look like?**
$\mathtt{Q}'_6$: *what does the headmaster of harry potter and the chamber of secrets look like?*
$\mathtt{A}_6$: Albus Dumbledore is tall and thin ... twinkled with kindness and mischief.
. . . . .

$\mathtt{Q}_9$: **what is the real name of the boy wizard's nemesis mentioned above?**
$\mathtt{Q}'_9$: *what is the real name of the boy wizard's nemesis mentioned above?*
$\mathtt{A}_9$: Tom Marvolo Riddle

$\mathtt{Q}_{10}$: **what are his magical strengths?**
$\mathtt{Q}'_{10}$: *what are albus dumbledore's magical strengths?*
$\mathtt{A}_{10}$: He is known as one of the greatest Legilimens in the world ... can read minds and shield his own from penetration.

$\mathtt{Q}_{11}$: **is he related to any other family apart from the riddle one?**
$\mathtt{Q}'_{11}$: *is tom marvolo riddle related to any other family apart from the riddle one?*
$\mathtt{A}_{11}$: Yes, Gaunts

$\mathtt{Q}_{12}$: **how is he related to them?**
$\mathtt{Q}'_{12}$: *how is albus dumbledore related to the gaunts?*
$\mathtt{A}_{12}$: They are the last known descendants of Salazar Slytherin.

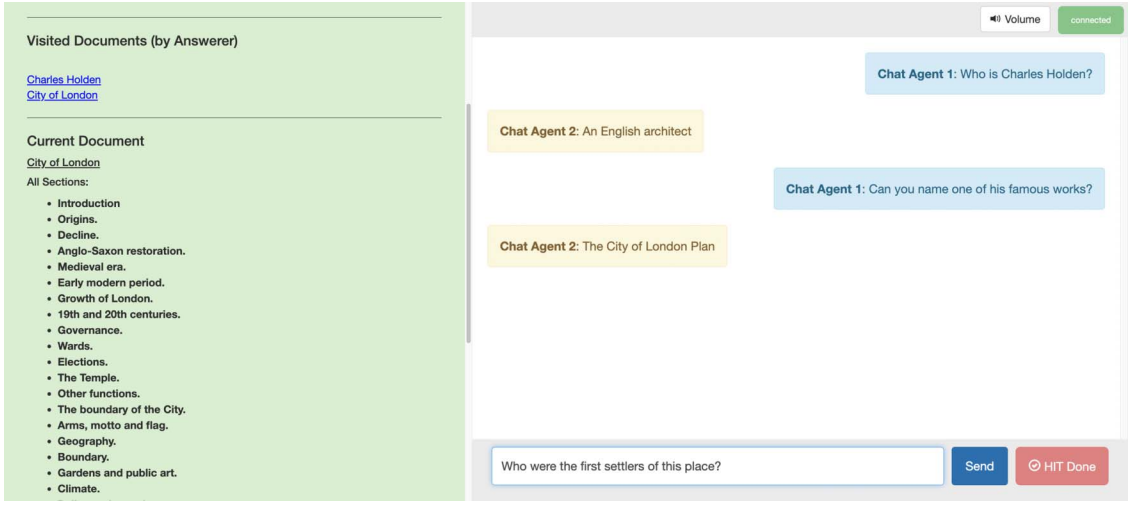
$\mathtt{Q}_{13}$: **in which city did the alchemist who worked with the headmaster live?**
$\mathtt{Q}'_{13}$: *in which city did the alchemist who worked with the headmaster live?*
$\mathtt{A}_{13}$: Paris

Figure 7: An example of a conversation from TopiOCQA along with rewrites from the QR module. Few turns are excluded and some answers are shorted for brevity.

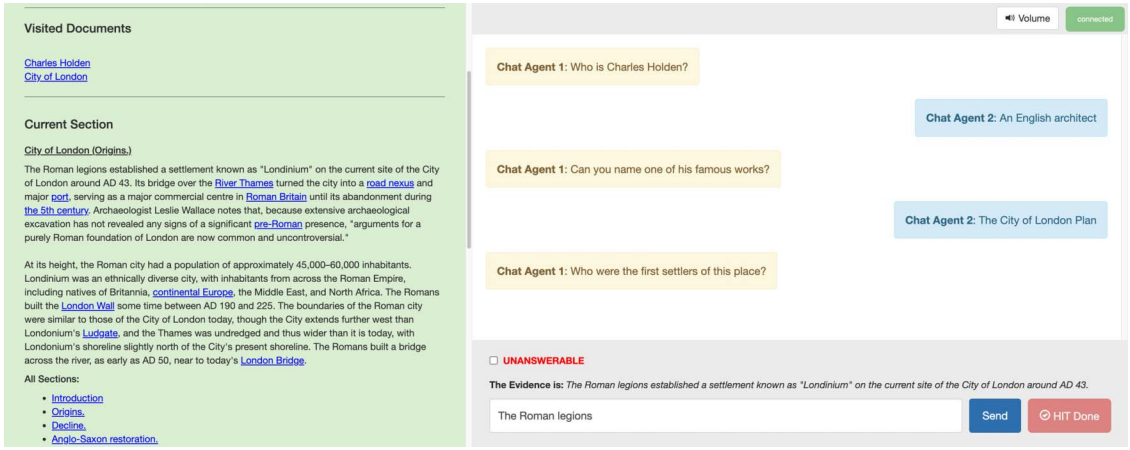
instead of *tom marvolo riddle* in $\mathtt{Q}_{10}$ and $\mathtt{Q}_{12}$). The QR module also fails to perform reasoning required for correct rewrites, for example, *boy wizard's nemesis* is not rewritten to *Lord Voldemort* in $\mathtt{Q}_9$, even though this information is present in $\mathtt{A}_5$).

C Hyperparameter Details

We use Lucene BM25 with $k1 = 0.9$ (term frequency scaling) and $b = 0.4$ (document length normalization). For both DPR and FiD, apart from the batch size, we use the hyperparameters suggested in their codebases. We use the maximum batch size that fits in the GPU cluster. DPR Retriever is trained on four 40GB A100 GPUs, whereas DPR Reader and FiD are trained on 8 32GB V100 GPUs. We use `base` model size for all systems. Following original implementations, DPR Retriever is trained for 40 epochs, DPR Reader for 20 epochs, and FiD for 15,000 gradient steps. The model checkpoint with best EM score on development set is selected as the final model.



(a) Questioner Interface



(b) Answerer Interface

Figure 8: Annotation interface for questioners and answerers.