HYBRIDIALOGUE: An Information-Seeking Dialogue Dataset Grounded on Tabular and Textual Data

Anonymous ACL submission

Abstract

A pressing challenge in current dialogue systems is to successfully converse with users on topics with information distributed across different modalities. Previous work in multiturn dialogue systems has primarily focused on either text or table information. In more realistic scenarios, having a joint understanding of both is critical as knowledge is typically distributed over both unstructured and structured forms. We present a new dialogue dataset, HYBRIDIALOGUE, which consists of crowdsourced natural conversations grounded on both Wikipedia text and tables. The conversations are created through the decomposition of complex multihop questions into simple, realistic multiturn dialogue interactions. We conduct several baseline experiments, including retrieval, system state tracking, and dialogue response generation. Our results show that there is still ample opportunity for improvement, demonstrating the importance of building stronger dialogue systems that can reason over the complex setting of information-seeking dialogue grounded on tables and text.

1 Introduction

When creating dialogue systems, researchers strive to enable fluent free-text interactions with users on a number of topics. In many cases, these systems can be utilized to navigate users over the vast amount of online content to answer the user's question. Current systems may search for information within text passages on sites such as Wikipedia (Dinan et al., 2018). However, knowledge comes in many forms other than text. The ability to understand multiple knowledge forms is critical in developing more general-purpose and realistic conversational models. Tables may be used to convey a different type of information that cannot be captured via text, such as structured relational representations between multiple entities across different categories (Chen et al., 2019, 2020b; Herzig et al., 2020). On the other hand, text may provide contextual or more fine-grained information regarding a specific entity. Thus, dialogue systems must be able to effectively incorporate and reason across both modalities to yield the best performance in the real world.

While there are several existing datasets targeted at dialogue systems (Dinan et al., 2018; Budzianowski et al., 2018; Eric et al., 2017; Zhou et al., 2018b), these are limited to either table-only or text-only information-sources. As a result, current dialogue systems may fail to respond correctly in situations that require combined tabular and textual knowledge.

To advance the current state of dialogue systems, we create HYBRIDIALOGUE. Our dataset is an information-seeking dialogue dataset grounded on structured and unstructured knowledge from tables and text. HYBRIDIALOGUE, or HYDI, is constructed by decomposing the complex and artificial questions in OTT-QA (Chen et al., 2020a) into a series of simple and more realistic intermediate questions regarding tables and text. HYBRIDIALOGUE contains conversations written by crowdsourced workers in a free-flowing and natural dialogue structure that answer these simpler questions and the complex question as well. We provide an example dialogue from our dataset in Figure 1. We also propose several tasks for HYBRIDIALOGUE that illustrate the usage of an information-seeking dialogue system trained on the dataset. These tasks include retrieval, system state tracking, and dialogue generation. Together, they demonstrate the challenges with respect to the dialogue system and the necessity for a dataset such as HYBRIDIALOGUE to further research in this space.

Our contributions are as follows:

- We create a novel dialogue dataset consisting of 4800+ samples of conversations that require reasoning over both tables and text.

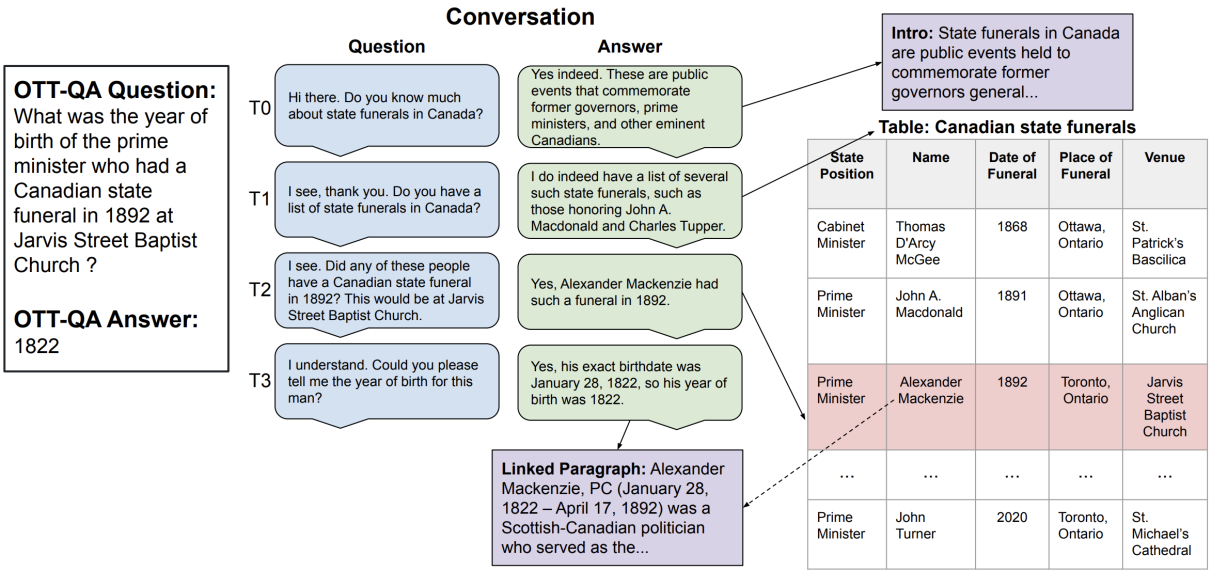

Figure 1: Overview of a sample from HYBRIDIALOGUE, where each conversation is created from a decomposed multihop question-answer pair. T0,...,T3 represent turns in the dialogue and consist of a single question and answer pair. The solid arrows represent the reference (e.g., row or intro paragraph) utilized to retrieve the correct answer in each turn. The dashed arrow represents a paragraph linked from a table cell.

- We decompose the overly-complex multihop questions from an existing dataset into more realistic intermediate question-answer pairs and formulate these in the dialogue setting.
- We propose system state tracking, dialogue generation, and retrieval tasks for our dataset. Our baseline experiments demonstrate opportunities to improve current state-of-the-art models in these various tasks and the overall information-seeking dialogue setting.

2 Related Work

Related work in the space of dialogue-based question-answering can be split into two areas: question-answering systems and information-grounded dialogue. We provide a comparison of the related datasets in Table 1 and analyze these datasets below.

Dataset	Dialogue	QA	Modality
CoQA	8K	127K	Text
Natural Questions	0	323K	Text
Hybrid-QA	0	7k	Table/Text
OTT-QA	0	45K	Table/Text
SQA	6.6K	17.5K	Table
ShARC	948	32K	Text
DoQA	2.4K	10.9K	Text
RecipeQA	0	36K	Image/Text
KVRET	3K	12.7K	Table
MultiWOZ	10.4K	113.6K	Table
WoW	22.3K	202K	Text
Topical-Chat	10.8K	235.4K	Text
CMU_DoG	4.2K	130K	Text
HYBRIDIALOGUE	4.8K	22.5K	Table/Text

Table 1: Comparison of HYBRIDIALOGUE and other dialogue and question-answering datasets.

Question-Answering As question-answering is one of the long-established NLP tasks, there are numerous existing datasets related to this task. Recently, other modalities have been incorporated into question-answering datasets. The Recipe-QA (Yagcioglu et al., 2018) dataset is comprised of question-answer pairs targeted at both image and text. OTT-QA (Chen et al., 2020a) and Hybrid-QA (Chen et al., 2020b) both contain complex multihop questions with answers appearing in both text and tabular formats. Several datasets are also targeted at the open-domain question-answering task such as TriviaQA, HotPotQA, and Natural Questions (Joshi et al., 2017; Yang et al., 2018; Kwiatkowski et al., 2019). While single-turn question-answering is valuable, the dialogue setting is interesting as it proposes many new challenges, such as requiring conversational context, reasoning, and coreference resolution.

Conversational Question-Answering Several question-answering datasets contain question and answer pairs within a conversational structure. CoQA (Reddy et al., 2019) and DoQA (Campos et al., 2020) both contain dialogues grounded with

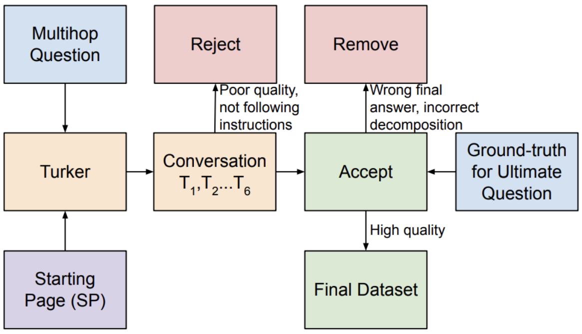

Figure 2: Overview of the dataset collection process, including the validation steps.

knowledge from Wikipedia pages, FAQ pairs, and other domains. ShARC (Saeidi et al., 2018) employs a decomposition strategy where the task is to ask follow-up questions to understand the user's background when answering the original question. However, ShARC is limited to rule-based reasoning and 'yes' or 'no' answer types. SQA (Iyyer et al., 2017) provides a tabular-type dataset, consisting of the decomposition of WikiTable questions. Each decomposed answer is related to a cell in a particular table. We utilize a similar strategy: we decompose the complex multihop questions from OTT-QA into a sequence of single-hop practical questions as described in Section 3. However, knowledge is limited to either the text or table within the reference page in the previous datasets. Thus, multimodality in the dialogue setting is limited, especially in the space of tables and text.

Compared to the previous datasets, our dataset poses a more challenging yet realistic setting, where knowledge over structured tables and unstructured text is required for providing reasonable answers to the conversational questions, and understanding the interaction between different types is necessary. In addition to cell locations, we also provide several other selection types including row, table, and text paragraph selection to provide more freedom in the way of answering questions. While the previous datasets contain samples written in a conversational structure, the answers are not necessarily presented in this way; they will instead formulate simple answers that do not emulate a human dialogue. In comparison, our dataset contains human-written questions and answers that produce an engaging dialogue.

Dialogue Generation Recent work actively constructs information-grounded dialogue datasets. The information sources are mainly from structured knowledge (e.g., tables and knowledge graphs) and unstructured ones (i.e., text). Among the dialogue datasets that leverage structured knowledge, some (Ghazvininejad et al., 2018; Zhou et al., 2018a) use conversational data from Twitter or Reddit and contain dialogues relying on external knowledge graphs such as Freebase (Bollacker et al., 2008) or ConceptNet (Speer et al., 2017). On the other hand, OpenDialKG (Moon et al., 2019) and DyKGChat (Tuan et al., 2019) collect conversations that are explicitly related to the paired external knowledge graphs.

Other related work revolves around task-oriented dialogues that are grounded on tables. For example, KVRET (Eric et al., 2017) and MultiWOZ (Budzianowski et al., 2018; Ramadan et al., 2018; Eric et al., 2019; Zang et al., 2020) provide tables that require an assistant to interact with users and complete a task. Dialogue datasets that are grounded on unstructured knowledge include CMU_DoG (Zhou et al., 2018b), which is composed of conversations regarding popular movies using their Wikipedia articles. On the other hand, Wizard-of-Wikipedia (WoW) (Dinan et al., 2018) and Topical-Chat (Gopalakrishnan et al., 2019) simulate the human-human conversations through Wizard-Apprentice, in which the apprentice tries to learn information from the wizard. Our proposed task shares a similar idea with Wizard-of-Wikipedia and Topical-Chat. However, we focus more on information-seeking dialogues grounded on both structured and unstructured knowledge, which provides abundant and heterogeneous information, and requires joint reasoning capabilities using both modalities.

3 Dataset Creation

3.1 Crowdsourcing Instructions

Instead of requiring users to create entire dialogues on various topics from scratch, we employed the usage of OTT-QA question-answer pairs as guidance, thereby increasing efficiency in the dataset construction. Given a multihop question from OTT-QA, crowdsourced workers (Turkers) from Amazon Mechanical Turk[1] (Crowston, 2012) were asked to decompose it into a series of simpler intermediate questions and answers to formulate a simulated conversation between a seeker and a knowledge expert similar to the Wizard of Wikipedia dataset collection process (Dinan et al., 2018). We

[1]https://www.mturk.com/

refer to the multihop question from OTT-QA as the "ultimate question". Turkers are instructed as follows: "In this task, you will engage in a dialogue with yourself. You will act as two characters: the seeker and the expert. At the top of the page, you are given the Ultimate Question. The seeker wants to know the answer to the ultimate question. However, directly asking this ultimate question is too complex. Thus, the seeker needs to decompose (break down) this complex question into a sequence of simple questions, which the expert will answer using a database."

To emphasize the conversational aspect of the dataset, Turkers were encouraged to ask questions that required understanding the conversation history context, such as through co-referencing. For example, Turkers used proper nouns with pronouns and indirect references such that they logically refer to their antecedents. In addition, Turkers were asked to provide questions that required understanding table logic to make the conversation more interesting and challenging. An example conversation is demonstrated in Figure 1 and an overview of the dataset collection process is shown in Figure 2.

3.2 Task Definitions

We obtain the ultimate question UQ, a starting table ST, and a gold answer to the ultimate question GA from the OTT-QA dataset. The starting table is the table that the Turker should use. We present this information to the Turker. A starting table is associated with the page SP it came from. For example, the ST table about herbariums in North America is located on the SP page about herbariums.

A conversation is composed of a sequence of turns. Each conversation consists of a minimum of 4 turns and a maximum of 6 turns. Each turn T acts as a piece of the decomposition of the ultimate question. The i-th turn T_i consists of a natural language question Q_i, a natural language answer A_i, a reference R_i, and an available reference pool set RP_i. The Turker provides Q_i, A_i, and selects a particular R_i from the set RP_i. R_i can be considered the evidence required to generate A_i given the question Q_i. The reference pool RP_i contains different types of references including the (linked) paragraph, a (whole) table, a single inner table row, multiple inner table rows, or a single cell. Note that the whole table refers to the table as a whole – a whole table reference would be used as opposed to an inner table row if the question asked about

Dataset Statistics	
# Dialogues	4844
# Turns (QA pairs)	21070
Avg Turns per Dialogue	4.34
# Wikipedia Pages	2919
Avg # words per question	10
Avg # words per answer	12.9
# Table selections	4975
# Row selections	6769
# Cell selections	1830
# (Linked) paragraph selections	3337
# Intro selections	7131
# Unique decompositions	267

Table 2: HYBRIDIALOGUE dataset statistics.

Decomposition	Count
$I \to T \to R \to P$	1419
$I \to T \to R \to C$	733
$I \to T \to R \to R$	290
$I \to T \to R \to C \to P$	218
$T \to R \to R \to P \to P$	136
$T \to R \to P \to P$	116
$T \to R \to C \to P$	116

Table 3: Top 7 most frequent decompositions. A decomposition is defined to be the sequence of references in a given conversation. I = Intro, T = Table. R = Row, P = Linked Paragraph, C = Cell

the summary about the table. In order to enforce the naturalness and moderate the difficulty of questions, we restricted RP_i based on RP_{i-1} and R_{i-1}. In other words, the questions that the Turker could ask were restricted based upon the selections made at previous turns.

In the Turker interface, RP_0 is restricted to the intro paragraph and any whole table references in SP. In addition, to help the Turker, we avoid the selection of any table references that are not ST to guide the Turker correctly.

The interface was built, tested, and refined multiple times to ensure maximum Turker productivity and a high-quality dataset. The interface evolved to a single page solution — all the tables, start page, and linked pages were fed ahead of time to the interface. [2]

[2] https://confident-jennings-6a2f67.netlify.app/plaid_interfaces/examples/1a_example_1.html

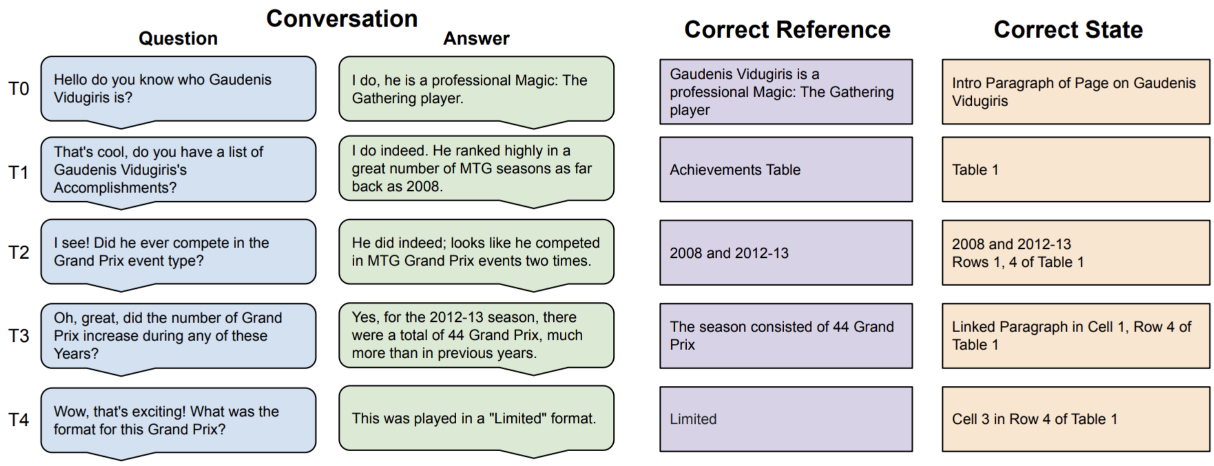

Figure 3: Overview of the state-tracking experiment. For each question in a conversation turn, there is a correct reference and corresponding state (e.g., row, linked paragraph) to select when answering the question.

3.3 Validation

When validating the Mechanical Turk samples, we underwent various filtering and verification steps. Rejections were made due to the Turker not following the instructions at all or having poor-quality conversations. Turkers were paid an average of $1.1 per conversation. Completing a conversation took the worker an average of 5 minutes, which translates to an average of $13.2 per hour. In some cases, we gave bonuses to Turkers who consistently submitted high-quality results. After final verification of the accepted HITs, we obtained a final dataset consisting of 4,844 conversations. The statistics of the dataset are shown in Table 2. From these conversations, we counted the number and frequency of unique decompositions, which is the selected reference sequence in a conversation. The most frequent decompositions are shown in Table 3.

We conducted additional filtering to further enhance the dataset quality. Utilizing gold answers obtained from the source OTT-QA dataset, we checked if the final answer appeared as a substring in Turker's conversation. If it did, we auto-approved the conversation. For the remaining questions, we manually reviewed them. We approved conversations that had the correct answer but in a different format (e.g., September 1, 2021, instead of 9/1/21). In some cases, Turkers provided their own decomposition or their own ultimate question and decomposition, so they did not obtain the final answer provided by OTT-QA. In these cases, if the conversation had high-quality and accuracy, we accepted it. We additionally removed any conversations that had a single type of reference used throughout the entire conversation (e.g., all intros).

4 Tasks and Baseline Models

We outline three different tasks in the following sections: retrieval, system state tracking, and dialogue generation. Together, these tasks formulate a pipeline dialogue system grounded on both structured and unstructured knowledge from tables and text. The first step of the system is to **retrieve** the correct Wikipedia reference given the first question in the dialogue. As the conversation continues, the system must be able to **track the state** of the conversation in order to obtain the correct information from the Wikipedia reference for the user. Finally, the system will need to **generate a natural conversational response** to communicate with the user at each turn. Thus, following each of these tasks in order simulates the pipeline system with our dataset. We describe each of these tasks and their respective models in detail below.

4.1 Retrieval

The retrieval experiment is run for each T_0 of each conversation. Given the first question of the conversation Q_0, the model must predict the correct reference R_0 from the set containing all intro paragraph and table reference candidates in the dataset. For our baseline, we run the Okapi BM25 retriever (Brown, 2020) on the training set and candidates. BM25 is a standard document retrieval model that uses keyword-matching techniques to rank relevant documents.

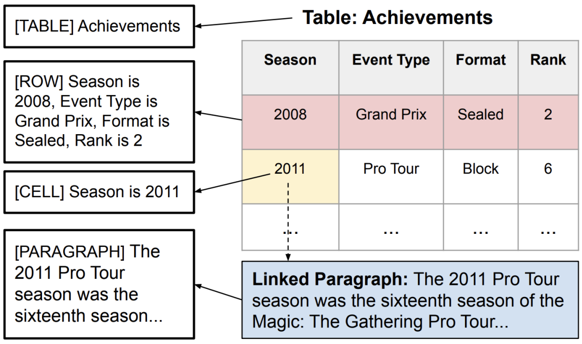

Figure 4: Table, row, cell, and paragraph flattening for input to the SentenceBERT and DialoGPT models.

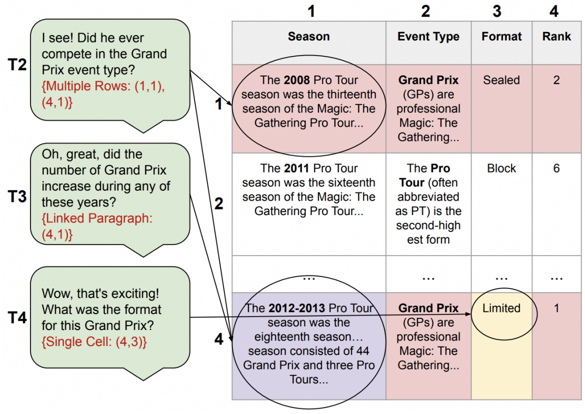

Figure 5: System state tracking with the TaPas model. Single rows and multiple rows are mapped to single cells and linked paragraphs are mapped to their respective cells in the original table in order to adapt to TaPas.

4.2 System State Tracking

Previous work in dialogue systems focuses on the task of belief state tracking, which aims to determine the user's goal or the current state of the conversation at each turn in the dialogue (Mrkšić et al., 2017; Ren et al., 2018). Inspired by work in belief state tracking, we propose the task of system state tracking in an information-seeking dialogue system. The task is framed similarly to belief state tracking, where a model attempts to classify the current state in the conversation at each turn. However, the "state" in our proposed task is modeled as a reference location from the current reference pool. As such, the task is formulated as using the information from the existing conversation and current question to determine the state of the conversation and choose which reference to utilize to create an answer. The reference types considered in this experiment are single cell, linked paragraph, inner table row, and multiple inner table rows. The implementation of system state tracking increases the interpretability and explainability of the system by determining the understanding of the user's question and discovering the point in the conversation in which the model is incorrectly interpreting the user's question. This, in turn, can help us understand the types of errors the model is prone to and allow us to work towards increasing the robustness of the model regarding these errors.

The system state tracking process is visualized in Figure 3. We perform system state tracking for all turns in each dialogue except the first turn. Given the history of the conversation H_i, we predict the correct reference R_i. H_i consists of turns $T_1 ... T_{i-1}$, the current query Q_i, and the candidate references RP_i. Thus, the goal is to determine the correct reference R_i at the specific turn in the dialogue, given the dialogue history. We utilize SentenceBERT (Reimers and Gurevych, 2019a) and TaPas (Herzig et al., 2020) as baselines for the experiment.

SentenceBERT We utilize the sentence transformer [3] and the triplet-loss configuration as described in equation 1. We minimize the difference between the correct candidate R_i and context H_i while maximizing the difference between every incorrect candidate W and H_i. We create samples for each $W \in RP_i$ where $W \neq R_i$. (RP_i is the reference pool). k is some fixed margin.

$$loss = max(||H_i - R_i|| - ||H_i - W|| + k, 0) \quad (1)$$

To allow SentenceBERT to process the data, we flatten the references and prepend a special token to provide information about the type of candidate it is. This process is visualized in Figure 4.

TaPas We additionally utilize the TaPas model for system state tracking. TaPas is a BERT-based question-answering model for tabular data. We use the TaPas model that has been fine-tuned on the SQA dataset, which enables sequential question-answering in a conversational nature. As the model performs only cell selection, we adapt TaPas towards this setting. We do not need to pre-process the data differently for cell selection as TaPas already performs the cell selection task. We place linked paragraphs in their respective cells within a table to accommodate cell selection in this setting. For row and multi-row selection, we pre-process the data by choosing one cell from the row as the

[3] We utilized paraphrase-distilroberta-base-v1 weights provided by the SBERT library (Reimers and Gurevych, 2019b).

Model	MRR@10	MAP
SentenceBERT	0.626	0.625
TaPas (Pre-processed)	0.455	0.427
TaPas (All)	**0.689**	**0.634**

Table 4: The results of the system state tracking experiments with the SentenceBERT and TaPas models.

Reference	MRR@10	MAP	Count
Cell	0.384	0.395	108
Paragraph	0.599	0.606	124
Row	0.782	**0.786**	338
Multi-row	**0.881**	0.292	66

Table 5: System state tracking results split by reference type for the TaPas All model.

correct answer. This is done by finding the cell with the highest text similarity to the ground truth answer at that turn. Therefore, each row will have a single cell associated with it during fine-tuning. We visualize the state tracking experiment with TaPas in Figure 5. For our experiments, we fine-tuned the TaPas model with our pre-processed training set.

4.3 Dialogue Generation

We conduct experiments on dialogue response generation to look into the dataset's expressivity for real-world dialogue scenarios. We fine-tuned a pre-trained DialoGPT model (Zhang et al., 2020) by minimizing the negative log-likelihood with two input settings. Q_i, A_i, and R_i are defined as the question, answer, and reference at the i-th turn, respectively. First, we only take the dialogue history as the input without knowledge content and predict the following natural language response. The format is described as:

$$\{Q_1, A_1, ..., Q_i, A_i, Q_{i+1}\} \mapsto A_{i+1} \quad (2)$$

The model trained with this setting is called DialoGPT-noR. Second, we flatten the references and concatenate the dialogue history as the input and predict the following natural language response. The references are flattened in the process seen in Figure 4. The format is:

$$\{R_1, Q_1, A_1, ..., R_i, Q_i, A_i, R_{i+1}, Q_{i+1}\} \mapsto A_{i+1} \quad (3)$$

The two settings enable us to validate whether the dataset provides valuable information for response

Method	SacreBLEU	BERTscore
DialoGPT-noR	14.72	0.8875
DialoGPT	21.63	0.8901

Table 6: The results of dialogue generation experiments on HybriDialogue dataset.

construction and how much information the references provide.

5 Experiments

5.1 Retrieval

As retrieval is the first step in the information-seeking dialogue pipeline, we need to ensure that information from the correct Wikipedia page is retrieved to determine whether the first question and any following questions will be answerable. We evaluate our retrieval model with MRR@1 (Mean Reciprocal Rank @1). Our results show that the model achieves an MRR@1 score of 0.37 (1619/4359) for retrieving the correct candidate.

5.2 System State Tracking

Evaluation To evaluate the SentenceBERT and TaPas predictions, we calculate MRR@10 (Mean Reciprocal Rank @10) and MAP (Mean Average Precision). Each model produces scores for the candidate references for a question. These scores are sorted into a ranked list, and the correct references are identified in this list. We then calculate MRR and MAP values with respect to the ranking of the correct reference in the ranked list.

We evaluate the TaPas model in two settings. In the first (Pre-processed), we only consider pre-processed ground truth selected cells as correct for row and multi-row states. In the second setting (All), we consider the highest-ranking cell from the ground truth row correct during test time. While both settings consider only a single cell within a row as correct for row and multi-row states, the first is limited to the pre-processed cell, while the second simulates a more realistic setting by allowing any cell within the row to be correct.

Results The results of our experiments with TaPas and SentenceBERT are shown in Table 4. Our results show that the first and more limited setting of TaPas (Pre-processed) drastically underperforms compared to SentenceBERT. Meanwhile, the second setting (All) is more comparable to SentenceBERT. This can be due to the fact that during

Error Type	Input	Output
Incoherent	[TABLE] Best-selling physical singles – 7–7.9 million copies ; [QUERY] Can you give me a list ... [PARAGRAPH] .... Known for her emotive mezzo-soprano voice, Morissette began her career in Canada . . . ; [QUERY] What is the vocal range of this singer?	Dialog GPT:Alanis Nadine Morissette is a Canadian-American singer, songwriter, and actress. GT: Alanis Morissette has a vocal range of mezzo-soprano.
Non-fluent	[ROW] Year is 1985 ; Song is La det swinge ; Artist is Bobbysocks ; Position is 1st ; Points is 123... [QUERY] Do you know what song they performed to win?	DialoGPT: Yes, they performed to win the song La det swinge. GT: Yeah, they did La det swinge.
Unfaithful	[PARAGRAPH] Immigration to Spain Immigration to Spain increased significantly in the beginning of the 21st century. In 1998, immigrants accounted for 1.6% of the population, and by 2009, that number had jumped to above 12% - one of the highest in Europe at the time. ... This trend is far from being reversed at the present moment and, in 2005 alone, the immigrant population of Spain increased by 700,000 people.; [QUERY] when did the immigrant population of spain increase by 700,000 people?	DialoGPT: The immigrant population increased by 700,000 in 1998. GT: the immigrant population increased by that number in 2005

Table 7: The error types observed in dialogue generation on HYBRIDIALOGUE. (GT: ground-truth)

row selection, more information is needed to answer the question than simply one cell in the row. The flexibility of the All setting eliminates this issue and still allows a single cell to be correct.

We further analyze the results of TaPas in the All setting by breaking down the MRR and MAP scores based on the four reference types: cell, linked paragraph, row, and multi-row. These results are shown in Table 5, along with the number of samples for each reference type in the test set. We find that TaPas achieves the best overall results for row states, which also comprise the largest fraction of samples. Meanwhile, multi-row achieves a high MRR score but a low MAP score, indicating that TaPas ranks some of the correct row candidates very low. Cell and linked paragraph states are limited to a single cell within the table, but linked paragraph samples achieve noticeably better results. This is likely because the paragraph text will contain more information than a cell's text, making it easier to determine the correct reference.

5.3 Dialogue Generation

We adopted SacreBLEU (Post, 2018) and BERTscore (Zhang et al., 2019) as the automatic evaluation metrics. As shown in Table 6, concatenating references can consistently improve both metrics. This shows that the collected references are necessary for generating dialogue. It can be seen that differences are more noticeable for SacreBLEU as opposed to BERTscore. This is due to the naturally similar outputs of BERTscore, where the ranking of the scores is a more reliable view of the metric.

We conduct further error analysis and find three main types of errors as listed in Table 7: *incoherent*, *non-fluent*, and *unfaithful*. As shown in Table 7, the generated response "Alanis Nadine Morissette is a Canadian-American singer, songwriter, and actress." is not an appropriate response to "What is the vocal range of this singer?". In this case, the generated response is incoherent based on the dialogue. Sometimes the response has the correct information, but it is not a fluent sentence. One example is the generated statement "Yes, they performed to win the song La det swinge". The final primary error type is that the generated response may be unfaithful to the perceived knowledge. For example, given a paragraph mentioning several years and events in history, the generated response mentions "1998", while the answer should be "2005".

6 Conclusion

In this paper, we presented a novel dataset, HYBRIDIALOGUE, for information-seeking dialogue where knowledge is grounded in both tables and text. While previous work has combined table and text modality in the question-answering space, this has not been utilized in the dialogue setting. Our results in the various tasks demonstrate that there is still significant room for improvement and illustrate the need to build models that can adapt well to this hybrid format. In addition to the baseline tasks, future research can utilize HYBRIDIALOGUE to explore automatic multihop question decomposition.

Ethical Considerations

We will be providing open access to our dataset for use in future research. This includes the samples of dialogues written by Mechanical Turk workers, the references that each dialogue turn is associated with, and the Wikipedia pages in which the references are located. The dataset will be open-sourced under the MIT License.

For the dataset collection task, we required Turkers to have a HIT Approval Rate of greater than 96% and be located in AU, CA, IE, NZ, GB, or the US. We also required workers to have had 500 HITs approved previously. Workers were shown an interface containing text input fields and navigation tools. Turkers were also given an instruction page containing a video demo and a completed example. The time to complete the task is around 5 minutes, and Turkers were paid $1.1 per conversation, which translates to an hourly wage of $13.2 per hour.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W Cohen. 2020a. Open question answering over tables and text. In *International Conference on Learning Representations*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of* the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-Yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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)*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.