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. 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. 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 often convey information that cannot be efficiently 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 contain more detailed 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 multihop questions in OTT-QA (Chen et al., 2020a) which may not reflect real-life queries. We transform these into a series of simple and more realistic intermediate questions regarding tables and text that lead to and eventually answer the multihop question. 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 dialogue systems 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.

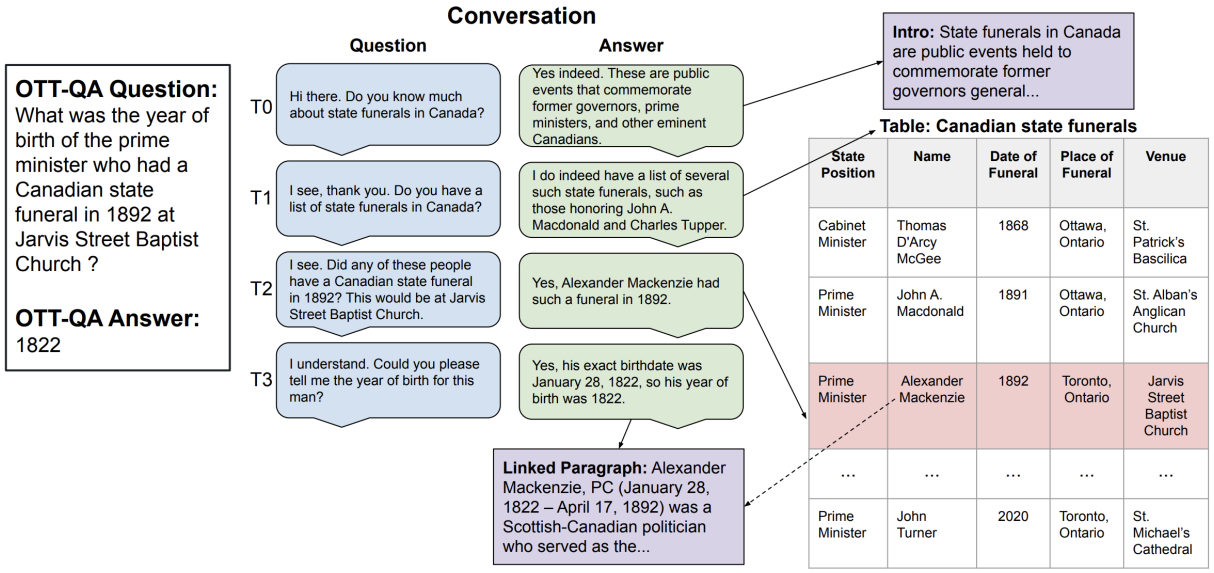

Figure 1: Overview of a sample from HYBRIDIALOGUE, where each conversation is created from a decomposed multihop question-answer pair. T0,...,T3 represent turns in the dialogue and consist of a single question and answer pair. The solid arrows represent the reference (e.g., row or intro paragraph) utilized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

Question-Answering As question-answering (QA) is one of the long-established NLP tasks, there are numerous existing datasets related to this task. Recently, QA datasets have been incorporating new modalities. The RecipeQA (Yagcioglu et al., 2018) dataset is comprised of question-answer pairs targeted at both image and text. OTT-QA (Chen et al., 2020a) and HybridQA (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 more interesting as it proposes many new challenges, such as requiring conversational context, reasoning, and naturalness.

Dataset	Dialogue	Turns	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	101K	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. For question-answering datasets, turns refers to question-answer pairs. For ShARC, dialogues refers to dialogue trees.

Conversational Question-Answering Several question-answering datasets contain question and answer pairs within a conversational structure.

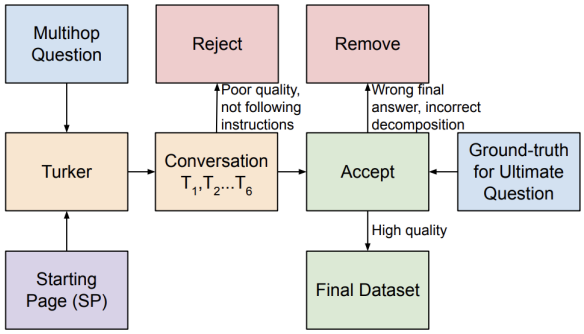

Figure 2: Overview of the dataset collection process, including the validation steps.

CoQA (Reddy et al., 2019) and DoQA (Campos et al., 2020) both contain dialogues grounded with 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 or column of cells in a particular table. In these datasets, knowledge is limited to a single modality.

In comparison, our dataset poses a more challenging yet realistic setting, where knowledge over structured tables and unstructured text is required to provide reasonable answers to the conversational 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 and short answers that do not emulate a human dialogue. Our dataset, therefore, extends conversational question-answering and falls into the dialogue space. HYBRIDIALOGUE contains natural dialogues with strongly related question-answer pair interactions whose answers are longer than the exact answer string. This models real-world occurrences in which a person wants to ask follow-up questions after their initial question has been answered.

Dialogue Generation Among the dialogue datasets that leverage structured (tables and knowledge graphs) 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), DuConv (Wu et al., 2019), DyKGChat (Tuan et al., 2019), and KdConv (Zhou et al., 2020) 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 and their corresponding simplified 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 in terms of asymmetric information among participants. 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

Given a multihop question from OTT-QA, crowdsourced workers (Turkers) from Amazon Mechanical Turk (Crowston, 2012) were asked to decompose it into a series of simpler intermediate questions and answers to formulate a simulated conversation in English. [1] As opposed to datasets such as Wizard of Wikipedia (Dinan et al., 2018) that are more open-ended, our annotators have a specific goal in mind: to answer an original complex question. By utilizing a single annotator to represent both sides, we keep the flow of the dialogue consistent and natural as it converges to the final answer. The usage of two annotators for our specific task comes with the risk of having one user diverge and reduce the chance of reaching the correct final

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

answer.

We 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 further emphasize the naturalness 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. 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

A conversation is composed of a sequence of turns. Each conversation consists of a minimum of 4 turns and a maximum of 6 turns. This limitation is specified to ensure that Turkers are thoroughly decomposing each complex question and the conversations do not go off on tangents. 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 from an English Wikipedia page, 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.

We differentiate between multiple rows and the whole table in order to obtain a more specific source for the information. For example, the question "Do you have a list of Steve's accomplishments?" requires a Table response as the answer contains a summary of the table. On the other hand, the question "Did he ever compete in the Grand Prix event type?" requires a selection of specific rows of some 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 type of questions that the Turker could ask were restricted to the references enabled from previous selections. In the Turker interface, RP_0 is restricted to the intro paragraph and any whole table references in a provided starting page.

Dataset Statistics	
# Train Dialogues	4359
# Development Dialogues	242
# Test Dialogues	243
# 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.

3.3 Validation

To ensure high-quality samples, we conducted various filtering steps. Rejections were made due to the Turker not following the instructions at all or having poor-quality conversations. For example, if the Turker purposefully copy and pasted unrelated paragraphs of texts, repeated the same questions multiple times, used unrelated references, or utilized a single reference throughout the entire conversation, we automatically rejected it. 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.

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 ques-

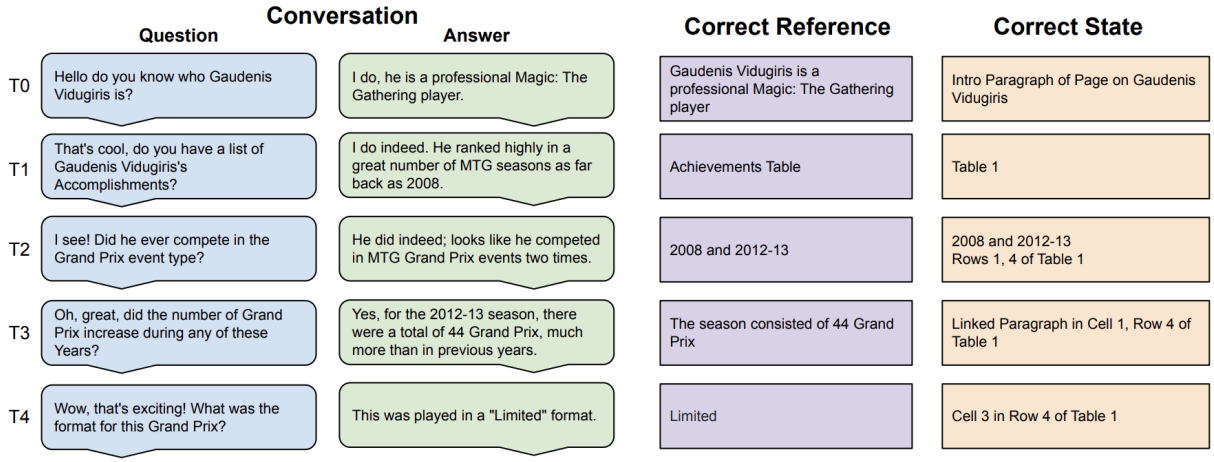

Figure 3: Overview of the state-tracking experiment. For each question in a conversation turn, there is a correct reference and corresponding state (e.g., row, linked paragraph) to select when answering the question.

tions, 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 was both accurate and had good quality, we accepted it.

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. First questions discuss information that is either in a table or an intro paragraph; so

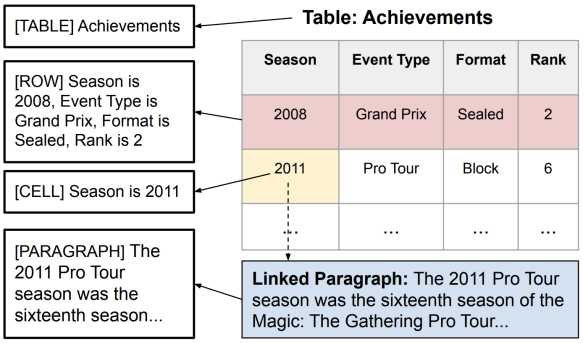

Figure 4: Table, row, cell, and paragraph flattening for input to the SentenceBERT and DialoGPT models.

the candidate space contains all intro paragraphs and tables in the dataset. The purpose of the retrieval experiment is to get a baseline of how well we are able to predict the table or page the subsequent conversation will be based upon, given the first query. The references that are utilized in the subsequent conversation are on the same page as the selected intro paragraph or table. 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 documents.

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

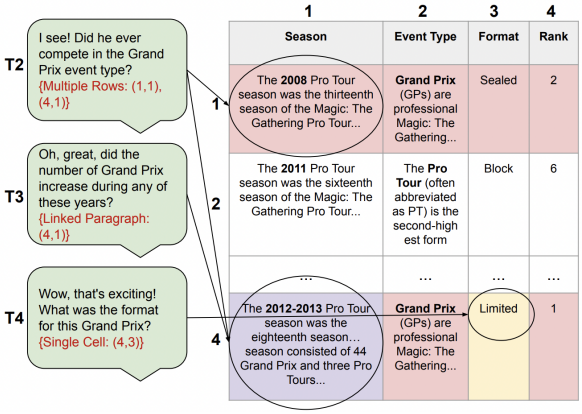

Figure 5: System state tracking with the TaPas model. Single rows and multiple rows are mapped to single cells and linked paragraphs are mapped to their respective cells in the original table in order to adapt to TaPas.

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 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.

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

Table 3: The results of the system state tracking experiments with the SentenceBERT and TaPas models.

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 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

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

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 (DialoGPT-noR) is described as:

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

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 (DialoGPT) is:

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

The two settings enable us to validate how much information the references provide for response construction.

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.

Method	SacreBLEU	BERTscore
DialoGPT-noR	14.72	0.8875
DialoGPT	21.63	0.8901

Table 5: The results of dialogue generation experiments on HYBRIDIALOGUE dataset.

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 3. 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 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 4, 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

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?	DialoGPT: 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...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 6: The error types observed in dialogue generation on HYBRIDIALOGUE. (GT: ground truth)

evaluation metrics. As shown in Table 5, concatenating references can consistently improve both metrics and 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 6: *incoherent*, *non-fluent*, and *unfaithful*. As shown in Table 6, the generated response "Alanis Nadine Morissette is a Canadian-American singer, songwriter, and actress." is not an appropriate response to the question. 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".

5.4 Human Evaluation

In addition, we conduct a human evaluation. We randomly sample 200 test samples containing previous conversation histories, human-written answers, and machine-generated answers from DialoGPT. For each sample, we have two Turkers provide ratings. We ask the Turker to evaluate the machine-generated response on three criteria: coherence, fluency, and informativeness from a scale of 1 to 5. Coherence measures how well the response is connected to the question and prior conversation history. Fluency measures the use of proper English. Informativeness measures how accurate the machine-generated response is against the human-provided ground truth response. We provide the average ratings for each model in Table 7. The model that utilizes the state tracking references achieves a better "informativeness" rating as it is able to utilize the extra information to provide a more correct response. It is notable however that the model with no references achieves better coherence and fluency scores. Thus, the human evaluation demonstrates the importance and challenge for models to provide both an accurate and articulate response.

Method	C	F	I
DialoGPT-noR	**3.88**	**3.98**	3.13
DialoGPT	3.59	3.68	**3.49**

Table 7: The results of human evaluation on dialogue generation model outputs. C = Coherence, F = Fluency, I = Informativeness.

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

While the dialogues in our dataset are grounded on both structured and unstructured data, they are limited to tables and text and do not cover other forms such as knowledge graphs. Additionally, the conversations are limited to discussions on single Wikipedia pages. We believe future research can expand on this for the creation of more open-ended information-seeking dialogues.

Wikipedia has extensive measures of risks and employs staff and volunteer editors to make sure Wikipedia articles meet the requirement and quality of the Wikimedia Foundation. Our data is based on Wikipedia pages, and we contain our dialogues to Wikipedia knowledge. We carefully validate the dataset collection process, and the quality of our data is carefully controlled.

The HybriDialogue dataset was built from the OTT-QA dataset, which is under MIT license. The authors of the OTT-QA dataset paper have allowed us to utilize the dataset within our use case.

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. For the human evaluation task, Turkers were paid $0.1 per task with an estimated time of less than 30 seconds per task. The dataset collection protocol was approved by the IRB. We follow the user agreement on Mechanical Turk for our dataset creation, which gives us explicit consent to receive users' service in the form of data annotation in return for monetary compensation. Given our settings, the Turkers understand that their data will be utilized in machine learning research.

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.

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.

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. Association for Computational Linguistics.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

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*.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108.

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.

A Appendix

A.1 Conversation Decompositions

We counted the number and frequency of unique decompositions in our dataset, which is the selected reference sequence in a conversation. The most frequent decompositions are shown in Table 8.

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

Table 8: 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

A.2 Experimental Details

We utilized paraphrase-distilroberta-base-v1 model with 82 million parameters provided by the SBERT library (Reimers and Gurevych, 2019b) for the SentenceBERT system state tracking experiment.

The TaPas model is built on the BERT model (Devlin et al., 2019). We utilize the TaPas-base model, which correlates to the BERT-base model that contains 110 million parameters. For system state tracking evaluation, we utilize average_precision_score from sklearn (Pedregosa et al., 2011). For retrieval experiments, we utilized the BM25Okapi algorithm from the Rank-BM25 library (Brown, 2020). Our experiments on dialogue generation utilize DialoGPT-small in the Huggingface transformers library (Wolf et al., 2020), which contains 124 million parameters.