MuTual: A Dataset for Multi-Turn Dialogue Reasoning

Leyang Cui[†‡*]**, Yu Wu**[◇]**, Shujie Liu**[◇]**, Yue Zhang**[‡]**, Ming Zhou**[◇]

[†]Zhejiang University
[◇]Microsoft Research Asia
[‡]School of Engineering, Westlake University
[‡]{cuileyang,zhangyue}@westlake.edu.cn [◇]{Wu.Yu,shujliu,mingzhou}@microsoft.com

Abstract

Non-task oriented dialogue systems have achieved great success in recent years due to largely accessible conversation data and the development of deep learning techniques. Given a context, current systems are able to yield a relevant and fluent response, but sometimes make logical mistakes because of weak reasoning capabilities. To facilitate the conversation reasoning research, we introduce MuTual, a novel dataset for **Mu**lti-**Tu**rn di**al**ogue Reasoning, consisting of 8,860 manually annotated dialogues based on Chinese student English listening comprehension exams. Compared to previous benchmarks for non-task oriented dialogue systems, MuTual is much more challenging since it requires a model that can handle various reasoning problems. Empirical results show that state-of-the-art methods only reach 71%, which is far behind the human performance of 94%, indicating that there is ample room for improving reasoning ability. MuTual is available at `https://github.com/Nealcly/MuTual`.

1 Introduction

Building an intelligent conversational agent is one of the longest running goals in AI. Existing conversational agents can be categorized into task-oriented dialogue systems (Kannan et al., 2016) and non-task-oriented chatbot systems (Shum et al., 2018; Wu et al., 2019). Owing to the rise of deep learning techniques and the large amount of conversation data for training (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018b), we are now witnessing promising results of chatbots both in academia and industry (Pan et al., 2019; Tao et al., 2019).

Neural dialogue systems are trained over a large dialogue corpus and used to predict responses given a context. There are two lines of methods. Retrieve-based methods and generation based methods rely

---

[*]Contribution during internship at MSRA.

M: Ma'am, you forgot your phone.
F: Oh, thanks, I couldn't live without this little thing.
M: I know what you mean. It is of great significance to you. So did you enjoy your dinner?
F: Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect. Johnny had his own place to play in and I had time to talk with my sisters and their husbands.

✓ (A) M: Thanks for your compliment for the restaurant.
✗ (B) M: I'm sorry that you don't have a good time.
✗ (C) M: Goodbye brother! Love you.
✗ (D) M: Hurry up honey, or we will be late for the dinner.

Figure 1: B is incorrect because there is no reason to apologize. C and D can be excluded because the relationship between two speakers are waiter and customer based on the context.

on matching scores and perplexity scores, respectively. Due to the development of text matching and pre-training models (Devlin et al., 2019; Liu et al., 2019), a machine is able to achieve highly competitive results on these datasets, even close to human performance. For instance, ESIM (Chen et al., 2017) achieves 88% on the Dialogue NLI (Welleck et al., 2019), and BERT achieves 85.8%, 93.1% and 98.5% in terms of $R_{10}@1$, $R_{10}@2$ and $R_{10}@5$ on the Ubuntu Corpus (Whang et al., 2019).

However, there is still a huge gap between high performance on the leader-board and poor practical user experience. Chatbot engines often generate responses that are logically incorrect or violate commonsense knowledge (Shum et al., 2018). A likely reason is that current dialogue systems do not have strong reasoning skills, and most of the cases in previous benchmarks can be tackled by linguistic information matching. Previous work has demonstrated that neural encoders capture a rich hierarchy of syntactic and semantic information (Jawahar et al., 2019; Clark et al., 2019). However, reasoning capability and commonsense knowledge are not captured sufficiently (Young et al., 2018).

One important research question is how we can evaluate reasoning ability in chatbots, which can potentially allow us to bridge the gap between high performance on leader-board and unsatisfactory practical performance. To this end, we develop

dataset	Task	Reasoning	Domain	Manually
Ubuntu (Lowe et al., 2015)	**Next Utterances Prediction**	✘	Technique	✘
PERSONA-CHAT (Zhang et al., 2018a)	**Next Utterances Prediction**	✘	Persona	✔
Dialogue NLI (Welleck et al., 2019)	**Next Utterances Prediction**	✘	Persona	✘
CoQA (Reddy et al., 2019)	Conversational QA	✔	Diverse	✔
Douban (Wu et al., 2017)	**Next Utterances Prediction**	✘	Open	✘
DREAM (Sun et al., 2019)	Reading Comprehension	✔	Open	✔
WSC (Levesque et al., 2012)	Coreference Resolution	✔	Open	✘
SWAG (Zellers et al., 2018)	Plausible Inference	✔	Movie	✘
CommonsenseQA (Talmor et al., 2019)	Reading Comprehension	✔	Open	✔
RACE (Lai et al., 2017)	Reading Comprehension	✔	Open	✘
ARC (Clark et al., 2018)	Reading Comprehension	✔	Science	✘
DROP (Dua et al., 2019)	Reading Comprehension	✔	Open	✘
Cosmos (Huang et al., 2019)	Reading Comprehension	✔	Narrative	✔
MuTual	**Next Utterances Prediction**	✔	Open	✔

Table 1: Comparison between our dataset and other datasets. "Manually" indicates that human writing of the question or answers is involved in the data annotation process, rather than mere manual selection of data.

an open domain **Mu**lti-**Tu**rn di**al**ogue reasoning dataset (MuTual) to facilitate conversation model reasoning capabilities. In particular, given a context, we prepare four response candidates, each of which is relevant to the context, but only one of them is logically correct. As shown in Figure 1, all responses follow the same topic, but only the first one is appropriated. It requires reasoning ability on social etiquette and relationship to make the correct choice, which is not considered by existing dialogue benchmarks.

We build our dataset based on Chinese high school English listening comprehension test data, where students are excepted to select the best answer from three candidate options, given a multi-turn dialogue and a question. The original data is formatted as ⟨dialogue, question, answer⟩, which is not directly suitable for our goal since chatbots only concern about how to respond contexts instead of answering an additional question. Therefore, we ask human annotators to rewrite the question and answer candidates as response candidates. Then our dataset follows the traditional response selection setting (Lowe et al., 2015), where a model should recognize a correct response from others for a multi-turn dialogue.

The resulting dataset, MuTual, consists of 8,860 challenge questions, in terms of almost all questions involving reasoning, which are designed by linguist experts and high-quality annotators. We evaluate state-of-the-art retrieval-based models and pre-training models on MuTual. The best method gives a R@1 of 71%, which significantly underperforms human performance (94%). To the best of our knowledge, MuTual is the first human-labeled reasoning-based dataset for multi-turn dialogue. We provide detailed analysis to provide insights into developing potentially reasoning-based chit-chat dialogue systems.

2 Related work

Table 1 compares our dataset with prior dialogue and reasoning related benchmarks.

Dialogue: The Ubuntu Dialogue Corpus is a large retrieval-based dataset (Lowe et al., 2015), extracted from Ubuntu chat logs. PERSONA-CHAT (Zhang et al., 2018a) considers consistent personality in dialogue. Crowd workers are required to act the part of a given provided persona, and chat naturally. Dialogue NLI (Welleck et al., 2019) is a natural language inference dataset modified from PERSONA-CHAT. It demonstrates that NLI can be used to improve the consistency of dialogue models. CoQA (Reddy et al., 2019) is collected by pairing two annotators to chat about a passage in the form of questions and answers. Each question is dependent on the conversation history. There are also several large-scale datasets in Chinese, such as Sina Weibo (Shang et al., 2015), Douban Conversation Corpus (Wu et al., 2017) and E-commerce Dialogue Corpus (Zhang et al., 2018b).

As shown in Table 1, most of the existing conversation benchmarks do not focus on testing reasoning ability. One exception is CoQA, which considers pragmatic reasoning. The difference is that CoQA is a machine comprehension dataset, in which conversations are based on a given passage. Another related reading comprehension dataset is DREAM (Sun et al., 2019), which is designed specifically for challenging dialogue-based reading
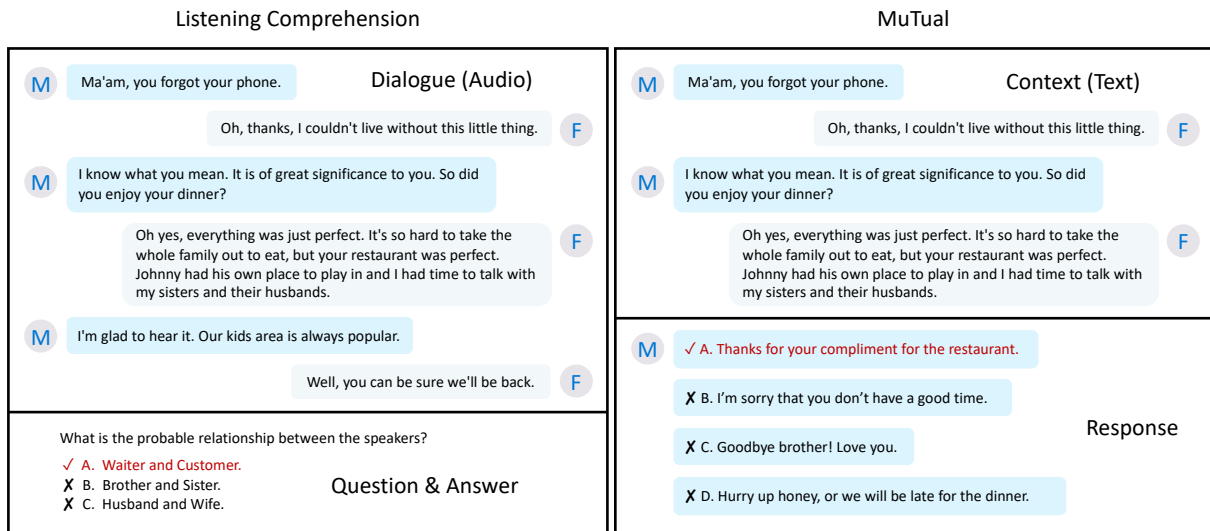
Figure 2: The process of modifying the listening comprehension test data.

comprehension. It relies on an external question to test the model's understanding capability. In contrast to the above dataset, our dataset is a next utterance prediction task, which is the fundamental problem in retrieval-based chatbots. In addition, our dataset requires various specific reasoning abilities, such as algebraic reasoning, intention prediction and so on, which is the main characteristic of our dataset.

Reasoning: Recently, efforts have been made to develop benchmarks and tasks to address reasoning for language understanding. Winograd Schema Challenge (Levesque et al., 2012) is a reasoning-based coreference resolution task. Each pair of sentences differs by only one phrase. SWAG (Zellers et al., 2018) is derived from pairs of consecutive video captions, including 113k short context each with four candidates endings. CommonsenseQA (Talmor et al., 2019) is a question answering dataset extracted from CONCEPTNET (Speer et al., 2016). Utilizing CONCEPTNET to construct the dataset ensures that questions directly target commonsense reasoning. RACE is a machine reading comprehension dataset collected from English exams for Chinese students. AI2 Reasoning Challenge (Clark et al., 2018) contains 7,787 genuine grade-school level science questions with a corpus of 14M science reference sentences. DROP (Dua et al., 2019) and COSMOS (Huang et al., 2019) focus on factual understanding and commonsense comprehension, respectively.

Despite their success, these datasets can hardly help chatbots directly. Following the traditional dia-

logue response selection setting, we deeply modify English listening comprehension conversation to form an utterance prediction task.

3 Dataset

3.1 Collection

The original listening comprehension materials and question-answer pairs are designed by linguist experts. Students are required to choose the best answer from three options for a question based on a piece of audio. To ensure students fully understand the audio, most of the questions need to be answered with reasoning capability.

We crawled the listening exams from public websites[1]. Since the audio is either a conversation between two people or a simple passage, we only crawled data in the conversation format. The raw data is formatted as triples ⟨Conversation (audio), Question and Choices (text), Answer (image)⟩. The following data pre-processing methods are applied to convert raw data to data in Figure 2.

Step 1 Pre-processing: If question and candidate choices in two problems are the same, we consider them as duplicates and delete one of them. If there are more than three candidate options in one problem, we randomly drop incorrect options until three candidates are left.

The answers are stored as images. We apply a commercial OCR system to convert images to text. It is easy to recognize the printed alphabet answer for the OCR system. We manually correct all OCR

---

[1] All the problems in our dataset are freely accessible online without copyright by consulting the legal adviser.

outputs to ensure quality. In the original listening comprehension test, the conversation is stored as audio. We adopt a commercial ASR system to convert speech to text, and further recruit experienced annotators to correct the transcription errors. To further ensure the quality of the transcripts, they are double-checked by annotators in the next step.

Step 2 Candidate Response Creation: Figure 2 illustrates the process of modifying the listening comprehension problem. At first, an annotator is required to segment the original conversation, after clues to answer the question have appeared. Then, they construct positive response (Response A in Figure 2) and negative responses (Response C and Response D) by consulting correct choice (Choice A) and incorrect choices (Choice B and Choice C), respectively. To make MuTual more challenging, we further ask the annotator to construct one more negative response (Response B) based on the correct choice. Through these steps, MuTual not only keeps the reasoning test designed by experts, but also introduces one more another type of reasoning for each instance. As shown in Figure 2, Response C and D can be excluded based on the relationship between two speakers. But B is incorrect due to the attitude reasoning.

It is worth noting that all negative responses are logically correct if the context is not considered, but they are not appropriated responses if the context is taken into account. Therefore, our dataset focuses on multi-turn conversation reasoning rather than the logic of a sentence. When framing a *negative response*, we encourage annotators to copy some phrases in the context to discourage a model that can solve the problem by text matching. We further calculate the *lexical overlap* between response and context. There are 9.98% (10.63%) words in the positive (negative) response that occur in the corresponding context, suggesting that MuTual is hard to solve by plain text matching.

Annotators in Step 2 are all English-major graduate students in Chinese, who are familiar with English language exams in China and fluent in English (pass the TEM-8[2]). Annotators are required to draft annotate 170 instances repeatedly, until their labeling is sufficiently accurate to provide useful annotation. Because not all conversations are adapted to construct a reasoning-based response problem, the annotator has the right to skip the con-

---

[2]The highest level test for English majors as a foreign language in China.

	MuTual
# Context-Response Pairs	8,860
# Avg. Turns per Dialogue	4.73
# Avg. Words per Utterance	19.57
Vocabulary Size (Context)	8,809
Vocabulary Size (Response)	8,943
Vocabulary Size	11,343
# Original Dialogues	6,371
# Original Questions	11,323
# Response Candidates	4

Table 2: Data statistics of MuTual.

versation. We employ five annotators to construct the response, and two quality inspectors to check it. We discard the instance when inspectors doubt the uniqueness or correctness of the answer.

3.2 Analysis

The detailed statistics of MuTual are summarized in Table 2. MuTual has an average of 4.73 turns. The vocabulary size is 11,343, which is smaller than other dialogue datasets (Lowe et al., 2015; Wu et al., 2017). Because MuTual is modified from listening tests of English as a foreign language, the complexity of morphology and grammar is much simpler than other datasets.

For human-annotated datasets, there is always a trade-off between the number of instances being annotated and the quality of annotations (Kryciski et al., 2019). Our dataset is smaller than the previous crawling-based dialogue dataset (Lowe et al., 2015; Wu et al., 2017) due to the collection method. But it is comparable with high-quality reasoning based dataset (Clark et al., 2018; Khashabi et al., 2018; Talmor et al., 2019) and human-designed dialogue dataset (Zhang et al., 2018a). Moreover, around 10k is sufficient to train a discriminative model (Nivre et al., 2019) or fine-tuning the pre-training model (Wang et al., 2019).

To assess the distribution of different reasoning types, we annotate the specific types of reasoning that are involved for instance, sampled from the test set and categorize them into six groups. The definition and ratio of each group are shown as follows.

Attitude Reasoning: This type of instance tests if a model knows the speaker's attitude towards an object.

Algebraic Reasoning: This type of instances tests whether a model is equipped with algebraic abilities when it chooses a response.

Intention Prediction: This type tests whether a model can predict what the speaker is going to do next.

Context	Candidates Responses	Reasoning Type
M: Hi, Della. How long are you going to stay here? F: Only 4 days. I have to go to London after the concert here at the weekend. M: I'm looking forward to that concert very much. Can you tell us where you <u>sing in public for the first time</u>? F: Hmm...at my <u>high school concert</u>, <u>my legs shook uncontrollably</u> and <u>I almost fell</u>.	✓ M: Haha, I can imagine how nervous you were then. ✗ M: Why were you so nervous at that time? It wasn't your first singing at your high school concert. ✗ M: Yeah, if I had been you, I would have been happy too. ✗ M: Why did you feel disappointed?	Attitude Reasoning (13%)
F: I'd like <u>2 tickets</u> for the 5:50 concert. M: That's <u>all be $9</u>.	✗ F: Please give me $9 refund. ✓ F: It's $4.5 for each ticket, right? ✗ F: Shouldn't it be $4.5 in total? ✗ F: I will pay you $2 more.	Algebraic Reasoning (7%)
F: I heard you were <u>having problems meeting your school fees</u> and <u>may not be able to study next term</u>. M: I was having some difficulties, but I have <u>received the scholarship</u> and <u>things are finally looking up.</u>	✗ F: Why are you going to drop out of school? ✗ F: You mean you'll try to get a scholarship? ✓ F: I am glad to hear that you will continue your studies. ✗ F: Why you have not received the scholarship?	Intention Prediction (31%)
F: Excuse me, sir. <u>This is a non smoking area</u>. M: Oh, sorry. I will move to the smoking area. F: I'm afraid <u>no table in the smoking area</u> is available now.	✗ M: Sorry. I won't smoke in the hospital again. ✓ M: OK. I won't smoke. Could you please give me a menu? ✗ M: Could you please tell the customer over there not to smoke? We can't stand the smell. ✗ M: Sorry. I will smoke when I get off the bus.	Situation Reasoning (16%)
M: This <u>painting</u> is one of the most valuable in the museum's collection. F: It is amazing. I'm glad I <u>spent $30 on my ticket</u> to the exhibit today. M: <u>The museum purchased it in 1935 for $2000</u>. But it is <u>now worth $2,000,000</u>.	✗ M: I heard the museum purchased it in 1678 for $2000. ✗ M: I heard the museum purchased it in 1678 for $30. ✗ M: So the sculpture worth $2,000,000 now. ✓ M: So the painting worth $2,000,000 now.	Multi-fact Reasoning (24%)
M: Good evening, ma'am. Do you have a <u>reservation</u>? F: No, I don't. M: Awfully sorry, but there are <u>no empty tables left now</u>.	✓ F: The restaurant is too popular. ✗ F: The restaurant is not crowded at all. ✗ F: So I have to eat in a bad table in the restaurant. ✗ F: Show me the way to the table.	Others (9%)

Figure 3: Examples from the MuTual dataset. All choices are relevant to context, but only one of them is logic correct. Some negative choices might be reasonable in extreme cases, but the positive one is the most appropriate. Clue words are purple and underline.

Situational Reasoning: Situation information (e.g., Location, Relationship between two speakers) is considered in this type of instance. A model should mine the implicit information from the previous context.

Multi-fact Reasoning: In this type of instance, the correct response is related to multiple facts in context, which requires the model to deeply understand the context rather than simply text matching.

Others:. There are 9% of instances that require other commonsense knowledge. For example, at the bottom of Figure 3, the model should know that a fully reserved restaurant is usually very popular.

The six types of reasoning are considered the most relevant to real chatbots. For example, it enables chatbots to make personal recommendations if a machine knows the user's attitude. The ability of intention prediction allows chatbots to respond more intelligently in a long conversation session.

3.3 MuTual^{plus}

To further increase the difficulty, we use *safe response* to replace one of the candidate responses for each instance in MuTual. To guarantee diversity, the safe response is sampled from a list including "I'm afraid I didn't quite catch what you were saying.", "Could you repeat that?", "I'm really sorry, I didn't catch that.", etc. In particular, once the

instance is chosen, we randomly select a response to replace. If the positive response is replaced, the correct one is the safe response. If the negative response is replaced, the original positive response is still the best one.

The motivation to build MuTual^{plus} is to evaluate whether a model is able to select a safe response when the other candidates are inappropriate. When we replace the positive response with a safe response, it simulates a scenario in which all the other candidates are incorrect. The phenomenon is common in retrieval-based chatbots, because limited candidate responses cannot handle all cases in practice. Similarly, we can evaluate if the model can choose the correct response instead of a safe response when a correct response exists.

4 Experiments

We split the data into training, development and test sets, with an 80%, 10% and 10% ratio. We pack instances constructed from the same conversation during splitting to avoid data leakage. Following the standard dialogue setting (Lowe et al., 2015; Wu et al., 2017), we consider our task as a response selection task and employ traditional information retrieval evaluation methods, including recall at position 1 in 4 candidates (R@1), recall at position 2 in 4 candidates (R@2) and Mean Reciprocal Rank

(MRR) (Voorhees, 2000). We compare the performance of several response selection models as well as pre-training models. We simply introduce these works as follows:

4.1 Baselines

We evaluate individual scoring methods, multi-choice methods and human performance in our experiment. Given a context c and four candidates (r_1, r_2, r_3, r_4), the individual scoring method computes a score for each choice independently with a score $g(c, r_i)$, and selects the individual with the highest score among four candidates. On the contrary, the multi-choice method selects the best one by classification over all choices, formulated as $h(c, r_1, r_2, r_3, r_4)$.

TF-IDF: The correct response tends to share more words with the context than the incorrect ones. Following Lowe et al. (2015), we calculate the TF-IDF vectors for the context and each of the candidate responses, respectively, and then select the highest cosine similarity between the context and the candidate response as the model output. The "IDF" is calculated only on the training set.

Dual LSTM (Lowe et al., 2015): Two LSTMs are used to encode context and response, respectively. The relevance between context and response is calculated by the similarity of the final hidden state from both LSTMs.

Sequential Matching Network (Wu et al., 2017): To avoid losing information in the context, SMN constructs a word-word and a sequence-sequence similarity matrix, instead of utilizing the last hidden state only, and then aggregates similarity matrix as a matching score.

Deep Attention Matching Network: Zhou et al. (2018) adopt self attention module (Vaswani et al., 2017) to encode response and each utterance, respectively. To match utterance and response, DAM further applies cross-attention module and 3D matching to obtain final score.

BERT (Devlin et al., 2019): Pre-training models have shown promising results on various multi-choice and reasoning tasks (Whang et al., 2019; Xu et al., 2019). Following Devlin et al. (2019), we concatenate the context (sentence A), and a candidate response (sentence B) as BERT input. On the top of BERT, a fully-connected layer is used for transforming the [CLS] token representation to the matching score.

RoBERTa: Liu et al. (2019) re-establish

BERT's masked language model training objective by using more data and different hyper-parameters. We fine-tune RoBERTa in the same way as BERT.

GPT-2 (Radford et al., 2019): Given a context, the positive response has a higher probability compared with negative responses. Motivated by this, we concatenate context and response as a sequence, and calculate the joint probability of an entire sequence. The response in the lowest perplexity sequence is considered as the positive response. Moreover, we fine-tune the GPT-2 on [Context, Positive Response] pairs in MuTual training set, denoted as **GPT-2-FT**.

Multi-choice Method: Inspired by BERT for multiple choice (Devlin et al., 2019), the task is considered as picking the most suitable response by comparing four candidates responses. In particular, we concatenate each candidate response with the corresponding context. Each input sequence is subsequently encoded to produce a [CLS] representation. The positive response is predicted based on the concatenation of all [CLS] representations, on which a fully connected layer with softmax is used. The method is denoted as **BERT-MC**. Similarly, we implement **RoBERTa-MC** as another multi-choice method.

Human Performance: To obtain the human performance, we employ 3 NLP experts to measure the ceiling performance on the test set.

4.2 Experiment Results

We report the performance of approaches introduced in 4.1, and human performance. Implementation details are shown in Appendix B.

4.2.1 Results on MuTual

All models perform significantly worse than on other popular conversation datasets, such as the Ubuntu Corpus (Lowe et al., 2015) and the Dialogue NLI dataset (Welleck et al., 2019), while human can address the reasoning problems easily. For example, BERT gives 85.8 % $R_{10}@1$ on the Ubuntu Corpus, but RoBERTa only gives 71.3% $R_4@1$ on MuTual.

TF-IDF only slightly better than randomly guessing, which indicates that there is no obvious statistic clue between context and positive response. In contrast, TF-IDF achieves 54.98% R@1 score on the Ubuntu Corpus, showing our dataset is more difficult to get the correct answer by text overlap. We evaluate typical retrieved-based dialogue models' performance on MuTual. From Table 3, we

Baseline category	Baseline method	Dev			Test		
		R@1	R@2	MRR	R@1	R@2	MRR
Baseline	Human	-	-	-	0.938	0.971	0.964
	Random	0.250	0.500	0.604	0.250	0.500	0.604
Individual scoring method (discrimination)	TF-IDF	0.276	0.541	0.541	0.279	0.536	0.542
	Dual LSTM (Lowe et al., 2015)	0.266	0.528	0.538	0.260	0.491	0.743
	SMN (Wu et al., 2017)	0.274	0.524	0.575	0.299	0.585	0.595
	DAM (Zhou et al., 2018)	0.239	0.463	0.575	0.241	0.465	0.518
	BERT (Devlin et al., 2019)	0.657	0.867	0.803	0.648	0.847	0.795
	RoBERTa (Liu et al., 2019)	**0.695**	0.878	0.824	**0.713**	**0.892**	**0.836**
Individual scoring method (generation)	GPT-2 (Radford et al., 2019)	0.335	0.595	0.586	0.332	0.602	0.584
	GPT-2-FT (Radford et al., 2019)	0.398	0.646	0.628	0.392	0.670	0.629
Multi-choice method	BERT-MC (Devlin et al., 2019)	0.661	0.871	0.806	0.667	0.878	0.810
	RoBERTa-MC (Liu et al., 2019)	0.693	**0.887**	**0.825**	0.686	0.887	0.822

Table 3: Comparison of varying approaches on MuTual.

Baseline category	Baseline method	Dev			Test		
		R@1	R@2	MRR	R@1	R@2	MRR
Baseline	Human	-	-	-	0.930	0.972	0.961
	Random	0.250	0.500	0.604	0.250	0.500	0.604
Individual scoring method (discrimination)	TF-IDF	0.283	0.530	0.763	0.278	0.529	0.764
	SMN (Wu et al., 2017)	0.264	0.524	0.578	0.265	0.516	0.627
	DAM (Zhou et al., 2018)	0.261	0.520	0.645	0.272	0.523	0.695
	BERT (Devlin et al., 2019)	0.514	0.787	0.715	0.514	0.787	0.715
	RoBERTa (Liu et al., 2019)	0.622	0.853	0.782	0.626	0.866	0.787
Individual scoring method (generation)	GPT-2 (Radford et al., 2019)	0.305	0.565	0.562	0.316	0.574	0.568
	GPT-2-FT (Radford et al., 2019)	0.226	0.577	0.528	0.226	0.611	0.535
Multi-choice method	BERT-MC (Devlin et al., 2019)	0.586	0.791	0.751	0.580	0.792	0.749
	RoBERTa-MC (Liu et al., 2019)	**0.621**	**0.830**	**0.778**	**0.643**	**0.845**	**0.792**
Transfer method	RoBERTa (Liu et al., 2019)	0.559	0.827	0.746	0.558	0.827	0.746
	RoBERTa-MC (Liu et al., 2019)	0.384	0.815	0.656	0.402	0.845	0.673

Table 4: Results on MuTual[plus]. Transfer method denotes that we train it on MuTual and test on MuTual[plus].

can see that well-designed matching models do not give better performance compared with simple dual LSTM, moreover, they drop by more than 50 absolute R@1 points compared to their performance on the Ubuntu Corpus, indicating that text matching models cannot handle reasoning problem well.

Both BERT and RoBERTa outperform other models in MuTual, which is consistent with results in other literatures (Talmor et al., 2019). This is mainly because models learn reasoning capability during the pre-training on a large corpus. Although RoBERTa only gets 71.3% on R@1, it achieves a surprising number, 89.2 %, on R@2, indicating that the model is able to rank the correct response to the top-2 position. BERT-MC and RoBERTa-MC obtain similar results with BERT and RoBERTa, respectively. However, even RoBERTa is far behind human performance 23 points on R@1, indicating that MuTual is indeed a challenging dataset, which opens the door for tackling new and complex reasoning problems in multi-turn conversations.

GPT-2 and GPT-2-FT also perform undesirably on MuTual, even if the averaged perplexity on MuTual testset is 10.40. This phenomenon illustrates that 1) sentences in MuTual are fluent; and

2) current generative models still have plenty of room to improve their reasoning ability.

4.2.2 Results on MuTual[plus]

As shown in Table 4, all models perform worse on MuTual[plus], indicating the dataset is more difficult than MuTual, which is consistent with our assumption. We find that the performance of multi-choice method is significantly better than individual scoring method. One possible explanation is that multi-choice methods consider candidates together, so they can distinguish whether or not the safe response is the best one. In contrast, individual scoring methods are not robust, and safe responses are easy to confuse methods in the training stage. Moreover, RoBERTa-MC outperforms others by a large margin, showing its outstanding performance on reasoning problems.

Furthermore, we conduct a transfer experiment, in which models are trained on MuTual but tested on MuTual[plus] without fine-tuning. The experiment investigates whether the model handles safe responses well if they have never seen them in training corpus. As shown in Table 4, RoBERTa-MC and RoBERTa drops 24.1% and 6.8%, respectively,
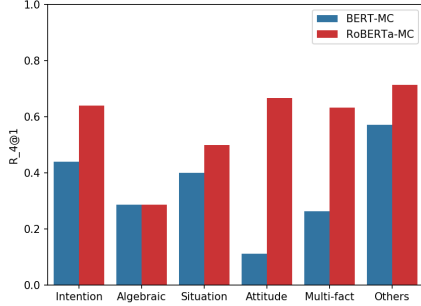
Figure 4: BERT-MC and RoBERTa-MC performance on different reasoning types.

Figure 5: Error analysis. ✗ indicates RoBERTa-MC's prediction.

in the transfer setting, demonstrating the benefits of seeing safe responses during the training process. Moreover, the individual scoring RoBERTa outperforms RoBERTa-MC, showing that the individual scoring method is more robust, when the safe response is not fed during training.

4.3 Discussion

Performance across different reasoning types: To analyze model performance across different reasoning types, we calculate BERT-MC and RoBERTa-MC performance on various question types as we introduce in Section 3.2. As shown in Figure 4, we find that the trends of BERT-MC and RoBERTa-MC are similar across different categories. RoBERTa-MC significantly outperforms BERT-MC in attitude reasoning and multi-fact reasoning. One potential reason is that there are some normal patterns between action and attitude captured by RoBERTa-MC, such as "play football" and "excited". However, instances that involve algebraic and situation show poor performance. These two reasoning types heavily depend on commonsense reasoning. Taking Figure 5 as examples, it takes a simple subtraction step to derive the time difference (5:00 pm - 6h = 11:00 am), but this turns out a significant challenge for RoBERTa-MC. In the second case, RoBERTa-MC fails to infer the dialogue situation, where the goal is to find a flat to rent.

Performance across different context lengths: It is interesting that the performance of RoBERTa does not decrease significantly with the number of turns increasing, which is different from the phenomenon observed on other datasets. As shown in Table 5, the performance drops by only 1.9 points R@1 from 2 turns to long turns (>6), and the performance of 5 turns is higher than those with 4
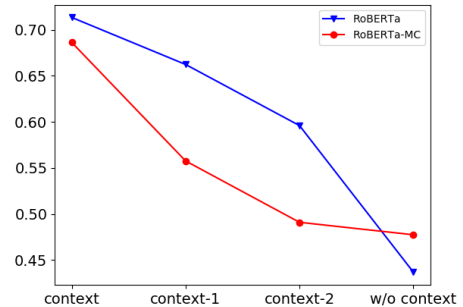

Figure 6: Ablation of context information. w/o context means all contexts are removed, so models just predict correct choice based on four candidates. context-n denotes the earlist n utterances are removed.

	#T=2	#T=3	#T=4	#T=5	#T> 6
#Instances	290	143	115	51	287
RoBERTa	0.731	0.657	0.635	0.804	0.712
RoBERTa-MC	0.681	0.622	0.609	0.725	0.750

Table 5: Performance comparison (R@1) of different number of turns on the test set. #T denotes number of turns. #Instances is the number of instances

turns, indicating the reasoning problems do not become much harder when the context becomes longer. The results also show that the difficulty of MuTual is attributed to reasoning instead of complex conversation history.

Context ablation study: We further verify whether our dataset requires multi-turn understanding rather than degenerating to a single turn reasoning problem. We evaluate Roberta and Roberta-MC performance when some utterances are manually removed. Figure 6 shows the performance when the earliest n utterances are removed in testing. As the ablation utterance increases, the performance of RoBERTa and RoBERTa-MC significantly decreases, which conforms to intuition. RoBERTa and RoBERTa-MC achieve only 43.7% and 47.7%

after ablating all utterances in the context, respectively, indicating the importance of each utterance and the quality of the dataset. Moreover, if we shuffle the sequence of utterance, the performance of RoBERTa-MC drops by 3.8% only, showing that it is insensitive to the utterance sequence information.

5 Conclusion

We introduced MuTual, a high-quality manually annotated multi-turn dialogue reasoning dataset, which contains 8,860 dialogues and aims to test reasoning ability of dialogue models. We describe the process for generating MuTual, and perform a detailed analysis. We find that various state-of-the-art models show poor performance in MuTual. The best model RoBERTa only obtains 71.3% R@1. There is a large gap between the model performance and human performance. We hope that this dataset facilitates future research on multi-turn conversation reasoning problem.

Acknowledgments

References

Joakim Nivre et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 955–964.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Wojciech Kryciski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, pages 552–561.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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 1824–1833, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249266.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ellen Voorhees. 2000. The trec-8 question answering track report.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. [link].

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2019. Domain adaptive training bert for response selection.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163197.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

496–505, Vancouver, Canada. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.