11-747 Assignment 3

Yifan Chen,[1] Haisu Yu,[2] Tianyu Xu,[3]
[1,2,3]Carnegie Mellon University
yifanc2@andrew.cmu.edu, haisuy2@andrew.cmu.edu, txu2@andrew.cmu.edu

Literature Survey

Dataset Analysis

Mutual [Cui *et al.*, 2020] is a novel dataset built for Multi-Turn dialogue reasoning. The dialogues are from Chinese student English listening comprehension exams. Different than traditional conversational dataset, Mutual requires reasoning ability over different types of problems including attitude reasoning, algebraic reasoning, intention prediction, situation reasoning, multi-fact reasoning, etc. The Mutual dataset analysis is published together with the dataset itself, as listed in Table 1. The exact questions and their distributions is shown in the Figure 1.

Apart from the original Mutual dataset, the paper also proposes Mutual-plus, in which one of the four options for each question is randomly replaced with a semantically **safe** answer like "Could you repeat that?" or "I'm really sorry, I didn't catch that.". [Cui *et al.*, 2020] These safe answers will confuse the model and pose great challenge to the reasoning task.

As the dataset has shown [Cui *et al.*, 2020], several category of methods including their baseline performance has been shown here.

	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 1: Data statistics of MuTual.

Related Datasets

In Mutual paper, it listed several dialogue datasets that are frequently used in tasks like next utterance prediction.

DREAM This dataset contains 6.5k dialogues from a variety of English language exams such as National College Entrance Examination.

CoQA CoQA [Reddy *et al.*, 2019](Conversational Question Answering dataset) contains 127k questions with answers obtained from 8k conversations over passages from seven different domains. Each turn of the conversation consists of a question and an answer.

RACE RACE [Lai *et al.*, 2017] is a dataset that contains 28k passages 100k questions generated by human experts collected from the English exams for Chinese high school students. These questions were designed to test the students' ability in understanding and reasoning.

Persona-Chat The Persona-Chat dataset is created by [Zhang *et al.*, 2018]. A persona is a 5 sentences description of a character. In a dialogue, there are two crowdworkers, each of them has a assigned persona. The workers need to get to know each other through a dialogue.

Related Techniques

Option comparison network [Ran *et al.*, 2019] is a multi-choice based method. The model first use a BERT-based [Devlin *et al.*, 2019] pre-trained model to extract basic features from the context, choice and question (If available). Then the network performs several feature extraction methods based on different attention mechanisms in order to produce the enhanced vectors for prediction. Finally, the enhanced features processed by a max pooling layer to generate the probability distribution of the correct answer and trained by a L2-regularized cross entropy loss function. These features [Ran *et al.*, 2019] are used to improve the prediction quality:

- Each candidate choice vectors computed from the BERT model is concatenated with the question vector given by the BERT model.
- The choice vector [Ran *et al.*, 2019] is then enhanced by concatenating itself with co-attention information and the fusing them with a simple linear layer. Then an element-wise gating mechanism [Ran *et al.*, 2019] is fused with the option-wise correlation information produced in the last step
- The option correlation feature is then enhanced by the co-attention and self-attention mechanism with the article vectors generated by the BERT model.

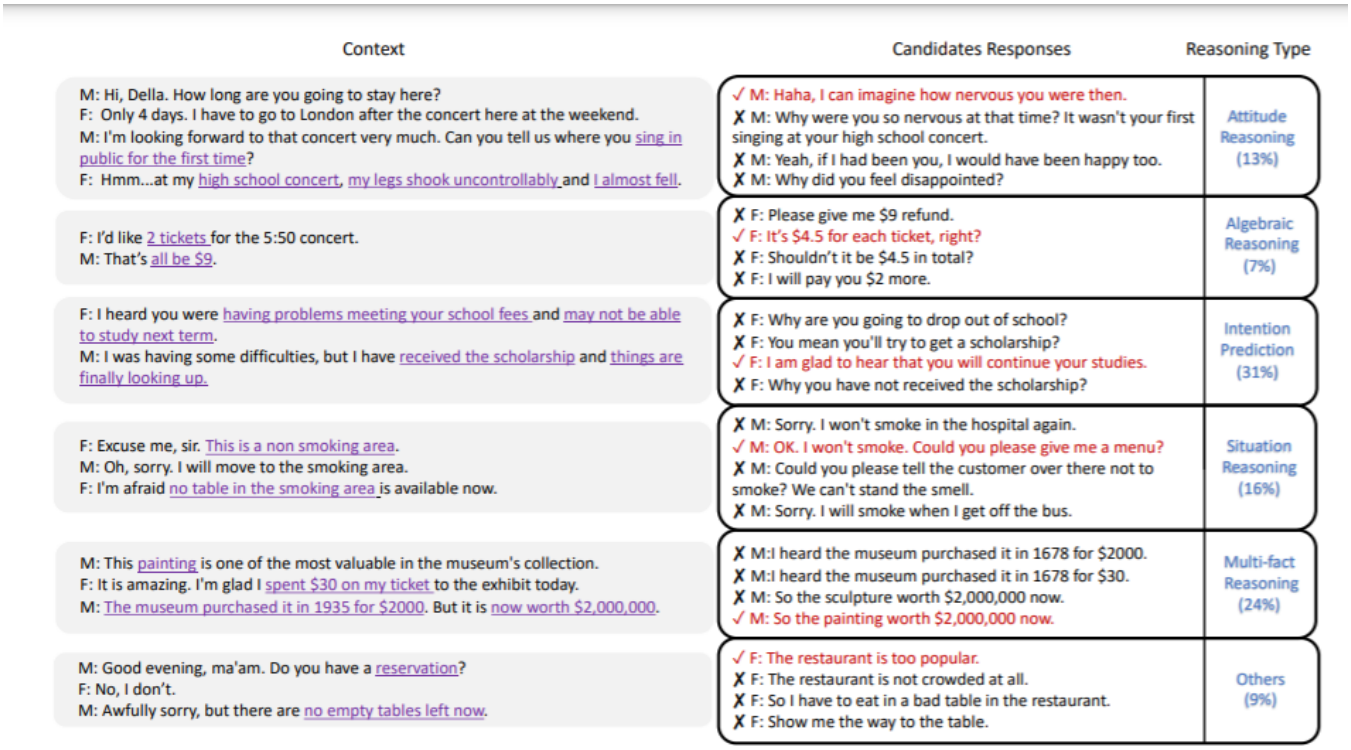

Figure 1: Red options suggests the correct response. The purple and underlined words are clue words to the correct answer

People also make use of the power of Graph Convolution Network in finding the relationship between utterances and each candidate response [Liu *et al.*, 2021b] in the model. The GRN model follows a standard Pretraining-Representation-Reasoning-Aggregation framework. It first creates its own base model, UBERT, which is trained on the target training set with its own unique task and finetuned on the target task. Then the UBERT was used to get the embedding for all contexts and candidate responses. Then, two individual models are used to extract different relationship respectively. The so-called sequence reasoning module (SSR) is actually a multi-head atttention layer that is used to extract suitable responses among candidate answers. The GCN then is used to extract local features among utterance representations. Then the graph matching result from the GCN and the attention result from the SSR are processed by a gate attention layer and a single layer perceptron to find the the final score for each candidate response. The model is trained with negative log likelihood.

People have also tried different pre-trained language models that has been designed to be pre trained with specially designed tasks that is related to the downstream tasks, like this paper [Li *et al.*, 2020]. The author use 49,930 dialogues from data sources that are considered as reliable and mostly correct in grammar, syntax and spelling to build the positive example with a score of 1. Negative examples are generated using utterance ordering, insertion and replacement on the positive examples with a score of 0. These examples are then multiplied with a sentence level coefficient that is based on the ngram Normalized Inverse Document Frequency. Then the pretrained models are trained to predict this score. For the response selection task like mutual dataset, the author suggests using context as textA and response as textB which can be fed into the ELECTRA [Clark *et al.*, 2020]. The best response is 1 and others are 0.

Network to reproduce

Mask-based Decoupling-Fusing Network (MDFN) [Liu *et al.*, 2021a] regards this problem differently by focusing on the dialogue characteristics. It emphasizes that the existing multi-turn dialogue methods rarely captures the speaker role transitions and global or local utterances inherency.

To leverage these information, MDFN specially includes modules to decouple dialogue information and fuse the utterance-aware features and speaker-aware features. The decoupling block consists of four multi-head self-attention blocks with different masks, which respectively attend to current utterances, other utterances, utterances of current speaker, utterances of other speaker. The former two attention outputs are grouped as ”utterance-aware channel” and the latter two are grouped as ”speaker-aware channel”. Different outputs within the same group will be fused and max-pooled to form the final representation of that channel. Note that the attention outputs and fused outputs are still word-level representations, and after the max-pooling over different utterances, the final outputs are utterance-level representation. The utterance-aware and speaker-aware utterance-level representations are further encoded with bi-directional

GRUs and concatenated as the final representation of the whole (dialogue, option) pair.

The authors of MDFN conduct different experiments of removing or stacking different components to find out the optimal model structure. The experiments include using different combination of masking channels, using different number of decoupling layers or GRU layers, using different fusing equations, using max-pooling or CNN to aggregate over utterances, and using different per-trained language models. These experiments prove the superiority of proposed structure.

Experiments Result & Analysis

Experiments result

We choose MDFN as our baseline and use the code from their Github to reproduce the result[1]. By testing on different hyper-parameters, we manage to reach a reasonable performance for Mutual task on the parameters of max_seq_length=256, batch_size=6, learning_rate=4e-6. Since the test set of Mutual has no labels and we have to send our prediction to the author to get the result, we only evaluate the performance on dev set. Apart from the original experiments, we also leverage Mutual plus dataset by mixing up the Mutual and Mutual plus training samples and fine-tune MDFN model on the mixed dataset with learning_rate=1e-6 and get MDFN$^+$. Table 2 shows our reproducing result.

	Mutual			Mutual Plus		
Model	R@1	R@2	MRR	R@1	R@2	MRR
MDFN (Reference)	0.923	0.979	0.958	-	-	-
MDFN (Reproduce)	0.922	0.973	0.956	0.684	0.954	0.834
MDFN$^+$ (Reproduce)	0.924	0.981	0.959	0.848	0.960	0.917

Table 2: Reproduce experiments results.

As a start of the analysis, we perform several experiments mentioned in the original Mutual paper but not covered by MDFN paper. First, we explore whether the performance is affected by the number of turns. Figure 2 shows the prediction R@1 under different context turns. As the number of dialogue turns increase, the R@1 metric decreases slowly and even increases as number of turns keep increasing, because the number of samples are decreasing as well. In the second experiment, we wish to explore whether removing the first several utterances will affect the performance. We prepare several subsets where the i-th subset has the number of turns $T_i > i$ so that we can remove the first i utterances. Figure 3 shows the R@1 of reference subset and removed-utterance subset. We observe an obvious performance loss after removing the first utterances. The experiments above illustrate that MDFN is capable of utilizing the whole dialogue context.

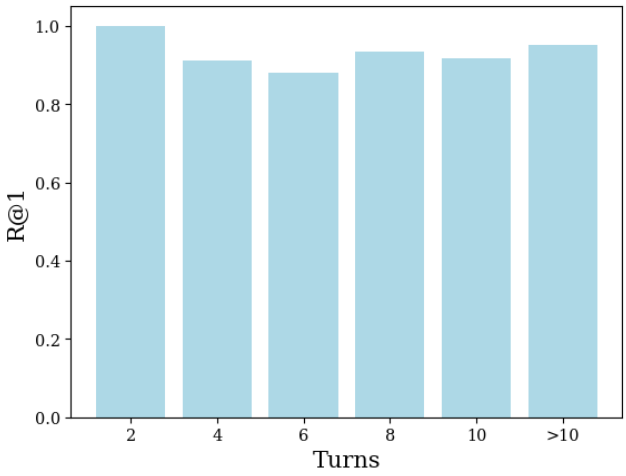

Figure 2: R@1 under different number of turns of the context. We count one turn as one speaker in one utterance.

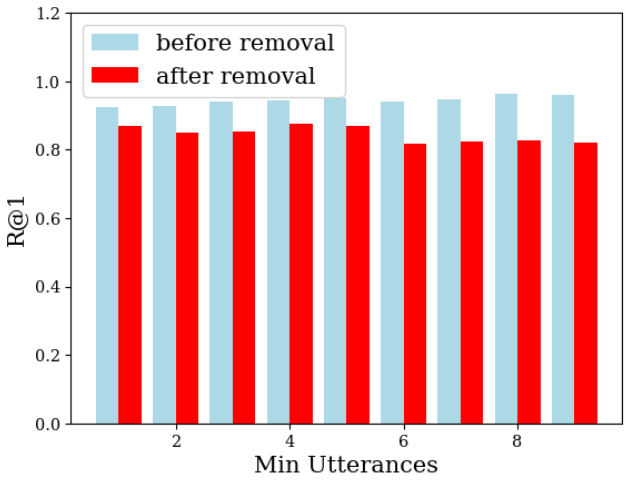

Figure 3: R@1 of different subsets of which the first i utterances are removed.

Inspired by OCN, we also conducted experiments to compare options on different levels. As commonly acknowledged, the model will get confused as the predicted option probabilities have little difference with one another. As figure 4 shows, the incorrect samples mostly have smaller variance. Further more, we evaluate the option similarity by calculating the minimum similarity between the correct answer and the other candidate answers. As figure 5 shows, the incorrect samples mainly place their option similarity between 0.9 to 1, which means similar options lead to a bigger chance of model failure.

[1]https://github.com/comprehensiveMap/MDFN.git

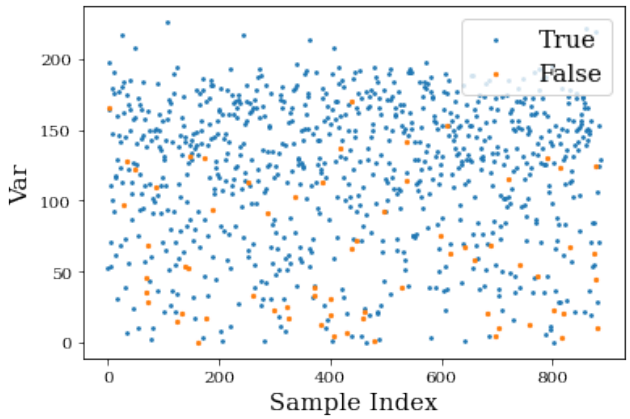

Figure 4: The variance of the predicted probability of all dev samples.

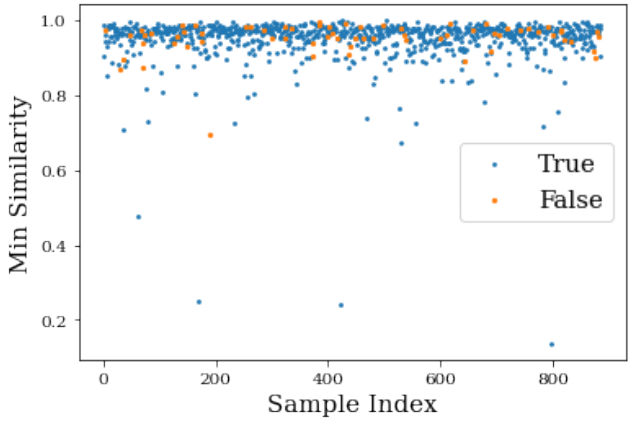

Figure 5: The minimum similarity of the options of all dev samples.

Typical Errors

Arithmetic For the numeric reasoning questions, the difference between options is the number in the sentence, which is related to the numbers mentioned in the dialogue. Like in the following example:
"m : Shelby museum information desk , can I help you ? ", "f : yes , please . could you tell me what the museum 's opening hours are ? ", "m : certainly , opening hours are from 9:00 am to 6:00 pm daily ."
Options:
*"f : I see . the museum opens for **ten hours** every day ."*
*"f : So the park opens for **nine hours** every day , right ?"*
*"f : Okay. Shelby museum is open for **nine hours** daily ."*
*"f : the museum opens for **seven hours** , doesn't it ?"*
To get the correct answer, arithmetic knowledge is required. For example, calculate the opening hours of a museum.

Named Entity For some questions, there are several named entities of the same category in the context. Like in the following example:
"f : i had no idea the countryside was so noisy ! ", "m : it 's usually very quiet in the north carolina mountains , kathy . but this is the year of our 17-year bird calls . ", ... "m : so you don't have these in the city . i thought i heard them when i went to atlanta georgia . ", "f : I've never heard anything like this before . california might not have them . ", "m : maybe they're just in the east"
Options:
*"f : maybe you're right . **north carolina** , where i come from , might have them ."*
*"f : maybe you're right.**georgia** , where i come from , might have them ."*
*"f : maybe you're right. **california** , where i come from , might have them ."*
*"f : maybe you're right.**california** , where you come from , might have them ."*
All of the named entities in the options have at least appeared once in the context. To get the correct answer, we need to analyse the discourse structure.

Sentiment Because of the nature of the dataset, among the semantically correct sentences, we don't want to pick the aggressive ones. Like in the following example:
"f : how about doing some exciting activities this weekend ? there 's a museum outside the village . ", "m : the kids will get bored and start fighting again like they did in that museum we visited last time."
Options:
"f : why don't you agree with my idea of going to a museum this weekend ?"
"f : if you disagree with my suggestion , just give me some suggestions instead."
Both options make sense but in this dataset, the second one is better because it is less aggressive.

Answer with "sorry" This error happens very frequently within the Mutual Plus dataset where random sentences are replaced with safe answers. In all 280 errors made by the model on the train dataset, 144 of them are caused by choosing the answers with the keyword "sorry". More importantly, the predictions for 127 out of 144 error samples gives the correct response the second highest scores, meaning that the model is able to solve most of such errors if it can differentiate safe answer from the correct responses. a typical example would be:
"F: Eat up the eggs and beef. You need your protein if you're going to be strong. ", "M: What I need is a cake. If I eat one more egg, I'll go crazy."
Options:
"F: Cake is actually a better option for you.", "F: I'm sorry, I didn't understand. Could you repeat a little louder, please?"
In this context, the safe answer makes sense but clearly is not the most suitable choice as the context suggests another one would be more related to the article.

Another interesting point to note is that the R@2 for most models are close to 0.99, while the R@1 are under 0.92. This suggests that the models usually have difficulty differentiating the best answer and the second best answer because of error reasons listed above.

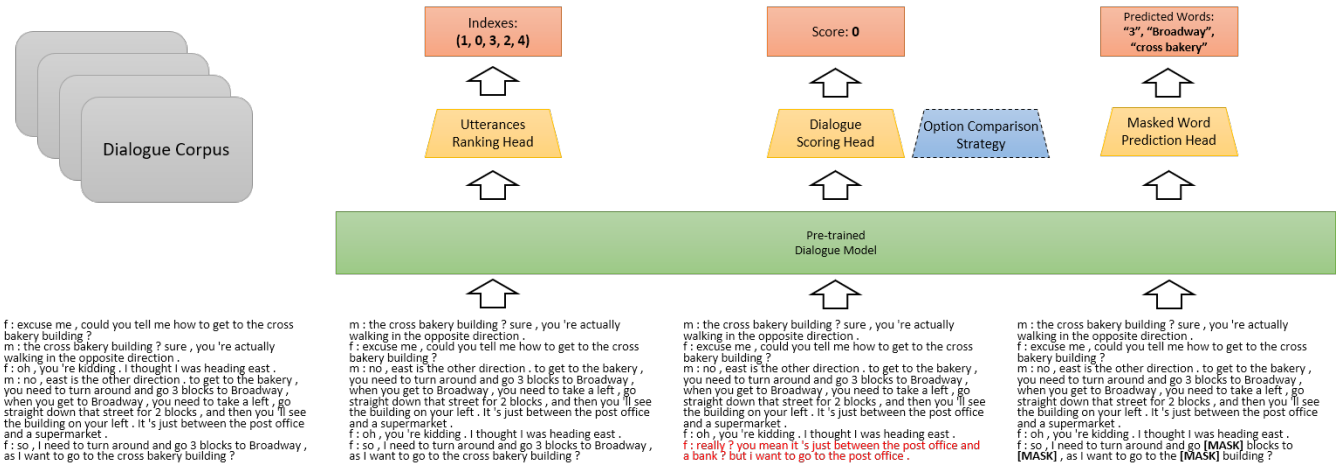

Figure 6: The pre-training tasks overview.

Project Proposal

We plan to improve the performance of Mutual task by increasing three different aspects of the model's capabilities: 1) encoding multi-turn dialogue information; 2) reasoning over certain logic or facts; 3) comparing over the candidate options.

According to the analysis we have made, MDFN has a superior model structure to encode the dialogue information regardless of how many turns are in the dialogue. However, a simple experiment of fine-tuning the model on the mixed dataset of Mutual and Mutual-plus shows that MDFN has not reach its limitation. Mutual-plus is basically an augmented version of MDFN by rule-based option replacement with no human labour at all.

To better encode multi-turn dialogue information, we borrow the idea from pre-training language models on large language corpus. We want to train a **pre-trained dialogue model (PrDM)** on sufficient resources of multi-turn dialogue datasets. These datasets have well-formed dialogues, and some have the same task settings as Mutual. Even with different task settings, they are still helpful if we treat those tasks as a pre-training sub-tasks to train our dialogue-encoding model.

DREAM and RACE are both open-domain conversational dataset composed by reasoning dialogues from English exams, which can be a direct supplement of the Mutual dataset. We also plan to use CoQA, a diverse-domain conversational dataset manually-created rather than migrating from exam papers. PERSONA-CHAT is another dataset focusing more on persona-domain rather than reasoning over facts. We believe those datasets will serve a good complementary material for the model to learn to encode dialogues from different domains and different scenarios.

We will pre-train our model on several tasks on these datasets on different levels inspired by the masked word prediction and next sentence prediction tasks of Bert pre-training. On word level, we will also use masked word prediction, which we believe can help increase the model's reasoning ability as well and we will introduce that later. On utterance level, we will perform utterance perturbations like insertion, deletion, and replacement and train the model to successfully identify the perturbed dialogues. Also, we consider to shuffle the utterances and train the model to rank the utterance correctly with ranking-based losses.

Meanwhile, our typical error analysis also points out the weakness of MDFN. MDFN can choose semantically correct answers hoepfully with its delicate structure and our pre-training setups, while it fails to reason over certain details like arithmetic or facts.

To strengthen the model's reasoning ability, we plan to add GCN structures similar to the one used in the graph reasoning network [Liu *et al.*, 2021c]. The GCN would be responsible to retreive local relationship between utterances which can then be used to enhance the extracted features in the middle of the MDFN in order to make better predictions.

Our designed masked word prediction pre-training techniques will also help increase the model's reasoning ability. Instead of randomly masking our words, we would like to fully leverage the labels of the dataset with differently-formatted tasks (i.e. question answering task). Therefore, we plan to use the correct answers to highlight the **reasoning keywords** in the original dialogue. By masking them out and ask the model to predict them, the model will learn an inductive bias of reasoning over details and facts rather than paying too much attention to semantics.

Now that we are taking our best efforts to help the model understand dialogue and reason better, we hope the model **think twice before making decisions**. According to our analysis as well as our common sense, models make more mistakes when the options are similar to one another. To gain the ability to compare over different options, we plan to add attention-based structures like OCN that is used to extract relationship features between options. Such features can be concatenated with other intermediate features to improve the prediction result.

The overall proposed method is shown in Figure 6.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 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, September 2017. Association for Computational Linguistics.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. Task-specific objectives of pre-trained language models for dialogue adaptation, 2020.

Longxiang Liu, Zhuosheng Zhang, , Hai Zhao, Xi Zhou, and Xiang Zhou. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. A graph reasoning network for multi-turn response selection via customized pre-training, 2021.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. A graph reasoning network for multi-turn response selection via customized pre-training, 2021.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. Option comparison network for multiple-choice reading comprehension, 2019.

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too?, 2018.