IRRGN: An Implicit Relational Reasoning Graph Network for Multi-turn Response Selection

Jingcheng Deng[1,†], **Hengwei Dai**[1,†], **Xuewei Guo**[2,†], **Yuanchen Ju**[1], **Wei Peng**[3,4,*]
[1]College of Computer and Information Science, Southwest University
[2]yz-intelligence Inc
[3]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[4]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
djc123234@163.com, kirobrine2000@163.com, g909336740@gmail.com
juyuanchen0213@163.com, pengwei@iie.ac.cn

Abstract

The task of response selection in multi-turn dialogue is to find the best option from all candidates. In order to improve the reasoning ability of the model, previous studies pay more attention to using explicit algorithms to model the dependencies between utterances, which are deterministic, limited and inflexible. In addition, few studies consider differences between the options before and after reasoning. In this paper, we propose an Implicit Relational Reasoning Graph Network to address these issues, which consists of the Utterance Relational Reasoner (URR) and the Option Dual Comparator (ODC). URR aims to implicitly extract dependencies between utterances, as well as utterances and options, and make reasoning with relational graph convolutional networks. ODC focuses on perceiving the difference between the options through dual comparison, which can eliminate the interference of the noise options. Experimental results on two multi-turn dialogue reasoning benchmark datasets MuTual and MuTual^{plus} show that our method significantly improves the baseline of four pretrained language models and achieves state-of-the-art performance. The model surpasses human performance for the first time on the MuTual dataset. Our code is released in the link.[1]

1 Introduction

The response selection task is one of the most important tasks in neural dialogue systems (Welleck et al., 2019; Demszky et al., 2020; Peng et al., 2022b; Zhang et al., 2020; Chen et al., 2021; Peng et al., 2022c,a; Zhao et al., 2022), which aims to find the most appropriate response from a set of candidate options given a historical dialogue. Most previous studies focus on matching between candidate options and historical dialogues while ignoring the reasoning ability of the model. This causes the model to choose logically incorrect or even counter-common-sense options, resulting in a poor user experience (Shum et al., 2018). On the recently released multi-turn dialogue reasoning benchmark dataset MuTual (Cui et al., 2020), these traditional representative response selection models(Wu et al., 2017; Zhou et al., 2018) perform poorly, which indicates that comparing with matching, reasoning ability is more important in MuTual. Specifically, matching finds semantically related candidates, while reasoning requires obtaining logically consistent responses based on logical and semantic dependencies between sentences.

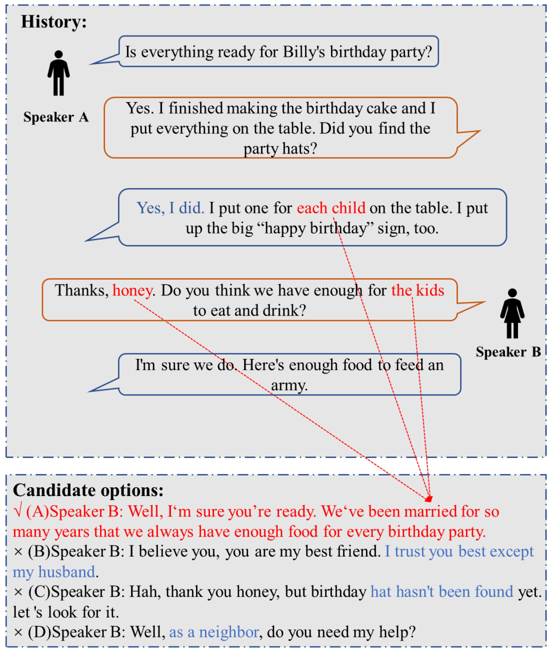


Figure 1: An example from MuTual. All candidate options are semantically related to the historical dialogue. Logical contradictions are marked with sky blue. Ground truth is marked with red. And the reasoning is red dashed lines.

For example, in Figure 1, all the relevant words that appear in option C include "*honey*" and "*birthday*", both of which occur in historical dialogue.

*Corresponding author. †Equal contribution.
[1]The codes are available at: https://github.com/DJC-GO-SOLO/IRRGN

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8529–8541
December 7-11, 2022 ©2022 Association for Computational Linguistics

Since traditional models tend to choose the more semantically relevant option, they consider option C as the best option. However, option C is not logically consistent with the historical dialogue. This conflicts with option C because the hat has already been found in the history dialogue. Similarly, options B and D also have the above problems. For the option A, "*married*" can be inferred from "*each child*", "*honey*" and "*the kids*", and it would be considered by a model with good reasoning ability. There are some work has emerged to improve the reasoning ability of models, but they still have some shortcomings.

Firstly, modeling the dependencies between sentences is an important part of improving the reasoning ability of the model. Experience shows that temporal dependencies (Lu et al., 2019; Yeh and Chen, 2019) between sentences as well as semantic dependencies are critical for multi-turn response selection. Previous studies typically historical dialogues and candidate options as the context (Su et al., 2019), or process each utterance independently (Tao et al., 2019), which lead to ignoring dependencies between sentences. There are also methods to model dependencies through explicit rules, such as using some sentiment dictionaries to obtain features or relying on some community detection algorithm for modeling (Liu et al., 2021b). However, the dependencies established by these explicit rules are deterministic and limited, which affects the performance of the model and can not establish the flexible dependencies between the utterances.

In addition, previous models only focus on the reasoning method, while ignoring the difference between candidate options before and after reasoning. Generally, people first compare the candidate options to understand each of them. After that, people read the article or historical dialogue information to make a deep reasoning and finally make a correct comparison between the candidate options again. Taking Figure 1 as an example, humans first compare the differences among the four options and find that the relationship between speakers differs the most. Then based on historical conversation information, words such as "*child*", "*kids*" and "*honey*" are captured. Comparing the differences between the four options again, it is found that only option A matches the historical dialogue, and finally, conclude that A is the best option. Inspired by human behaviors in reasoning, one can easily come to the correct answer after a two-stage comparison, which is similar to the preview and read methods in the PQ4R learning strategy.

Based on the above ideas, in this paper, we propose Implicit Relational Reasoning Graph Network (IRRGN), which consists of the Utterance Relation Reasoner (URR) and the Option Dual Comparator (ODC). Specifically, the purpose of the URR is to reason and adaptively capture flexible dependencies between utterances, as well as utterances and options, without relying on any explicit algorithm. The ODC is used to perceive the difference between the options before and after reasoning, which can eliminate the interference of the noise options. In summary, our contributions are as follows:

- We propose an URR, which adaptively captures flexible dependencies between utterances, as well as utterances and options, through a relational attention mechanism, and enables reasoning by propagating messages along various utterance paths.
- We propose an ODC, which captures the difference between options before and after reasoning according to the way humans think, which can eliminate the interference of the noise options.
- Empirical results show that our proposed model achieves state-of-the-art performance on MuTual and Mutual^{plus} datasets. This is the first time the model surpass human performance on the MuTual dataset.

2 Related Work

2.1 Response Selection Model

Current research can be roughly divided into three categories (Tao et al., 2021), which are representation-based models, interaction-based models, and pre-trained language model (PLM)-based models (Devlin et al., 2019). Representation-based models usually first encode historical conversations and candidate options by the representation layer, then apply an aggregation function to fuse the historical conversations into a fixed-length vector, and finally use a matching function to calculate a matching score (Yan et al., 2016; Zhou et al., 2016; Xu et al., 2021). Interaction-based models allow historical dialogue and response candidates to interact with each other at the beginning, and they usually follow a representation-match-aggregation paradigm (Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018). The PLM-based mod-

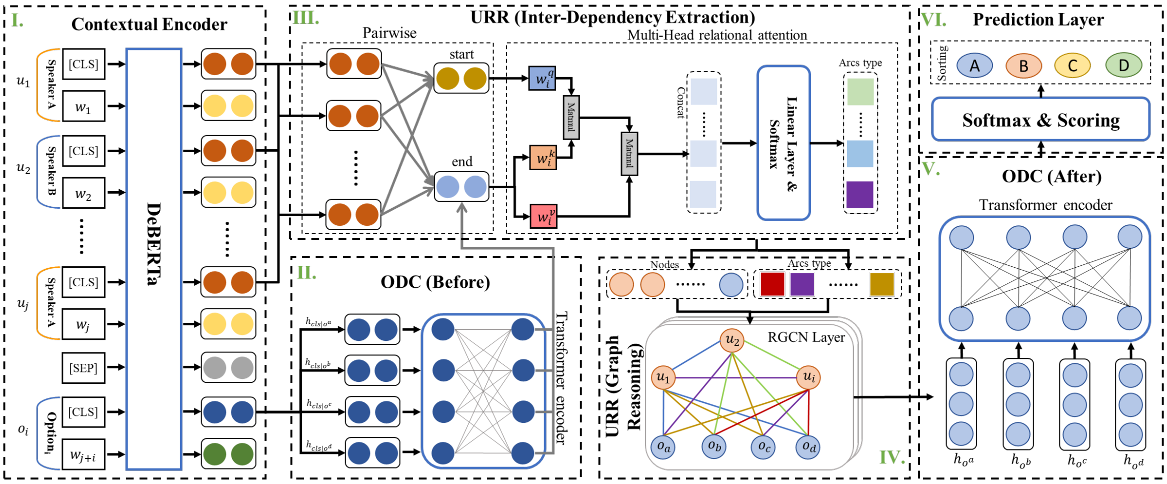


Figure 2: Overview of the proposed IRRGN. It contains six components: I. Contextual Encoder, II. ODC (Before), III. URR (Inter-Sentence Dependency Extraction), IV. URR (Graph Reasoning), V. ODC (After), and VI. Prediction. The gray arrow in II and III indicates that one of the elements is selected for calculation.

els concatenate historical dialogues and candidate options into a pre-trained multilayer self-attention network, and then perform representation, interaction, and aggregation operations in a unified manner through an attention mechanism. Henderson et al. (2019) pre-train a BERT model on a large general-domain conversation corpus, and fine-tune it in the target conversation domain, and finally aggregate each historical conversation-candidate option pair to compute a match score. Gu et al. (2020) incorporate speaker embeddings into BERT to enable the model to perceive speaker change information. Wang et al. (2021) propose a fine-grained comparison model (FCM) that models the logical consistency between dialogue histories and generated responses. Liu et al. (2021b) propose a graph reasoning network (GRN) to solve the problem of insufficient reasoning ability of the model, and the performance of the model can reach close to human level. However, these models ignore the differences between options before and after reasoning and do not have sufficient reasoning ability.

In addition, there are also some models (Zheng et al., 2019; Kang et al., 2021) that introduce visual information into dialogue modeling, but this cannot be applied to plain text modal data.

2.2 Graph Neural Network

Graph neural networks (GNN) achieve excellent performance in improving the reasoning ability of the model (Qiu et al., 2019; Tu et al., 2019). Previous studies also apply graph convolutional networks to models to enhance the reasoning ability of the model (Liu et al., 2021b). Different from previous work, in order to consider the influence of different edge relations on discourse reasoning, we leverage the relational graph structure to model the sequential structure between dialogues and utilize the graph convolutional structure to enable reasoning, which has better generalization than the traditional GCN structure (Schlichtkrull et al., 2018).

3 Model

The architecture is shown in Figure 2, which is divided into six modules in total. Firstly, the **Context Encoder** encodes representations of historical dialogues and candidate options. Secondly, the **Option Dual Comparator (Before)** compares the representation differences of options before reasoning. Then the **Utterance Relation Reasoner (URR)** grasps the different dependencies between utterances, as well as utterances and options, and makes a reasoning between the historical dialogue and candidate options to improve the reasoning ability of the model. Next, the **Option Dual Comparator (After)** module compares the differences of the options after reasoning. Finally, the **Prediction Layer** is used to calculate the score of the options.

3.1 Task Definition

Given a historical dialogue $U = \{u^1, u^2, \ldots, u^N\}$ where an utterance $u^n = \{w_1^n, w_2^n, \ldots, w_M^n\}$ with M words and a set of candidate options $O =$

$\{o^a, o^b, o^c, o^d\}$ where o^i is a candidate option. The goal is to learn the model $f(U, O)$, which can select the most logical candidate option y based on the matching scores of all the candidate O.

3.2 Contextual Encoder

The context encoder mainly obtains fixed-length vector representations of options and historical conversations based on a pre-trained language model.

Given each input example (U, O), the historical dialogue and all options are concatenated and fed into the pre-trained DeBERTa (He et al., 2021). It is worth noting that in order to facilitate sentence-level operations of each historical dialogue and option in the following modules, we insert a [CLS] token before each utterance. Then the fixed-length representation vector for each utterance and option can be obtained, which is denoted as:

$$[H^U; H^O] = \text{DeBERTa}([U; O]) \tag{1}$$

where $H^U \in \mathbb{R}^{|\text{tokenize}(U)| \times d}$ and $H^O \in \mathbb{R}^{|\text{tokenize}(O)| \times d}$ are the token-level vectors of context U and options O, respectively. $\text{tokenize}(\cdot)$ and d are the tokenization function and hidden layer dimension of the DeBERTa model, respectively. $[;]$ represents the concatenation operation, and $\text{DeBERTa}(\cdot)$ returns the output of the last layer of the DeBERTa model. In addition, h_{cls} represents the summary vector of each utterance.

3.3 Option Dual Comparator

This section describes components II and V in Figure 2. The ODC aims to eliminate the interference of noisy options by comparing the differences between the options before and after reasoning based on imitating human reasoning behavior.

Two transformer (Vaswani et al., 2017) encoders are applied to serialization modeling different options before and after the URR (components III and IV), so as to obtain the option representation containing the reasoning difference information, which improves the performance of the model. The multi-head attention mechanism is as follows:

$$\begin{aligned}&\text{Multihead}\\&= \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O\end{aligned} \tag{2}$$

$$\text{head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

where W_i^Q, W_i^K, W_i^V and W^O are all parameter matrices, and h is the number of attention heads. Q, K and V are $h_{cls|o^i}$ and h_{o^i}, where $h_{cls|o^i}$ represents the [CLS] vector of option o^i, and h_{o^i} represents the o^i representation vector after reasoning and $i \in \{a, b, c, d\}$.

3.4 Utterance Relational Reasoner

Inter-Dependency Extraction Temporal and semantic dependencies between different utterances are crucial for the response selection task. Therefore we move away from explicit dependency models to implicit ones. Specifically, we believe that in the process of encoding dialogues by pre-trained language models, the temporal and semantic dependencies between different utterances are hidden in some dimensions in the semantic space, so it only needs to be "mined" and given different types.

We employ a relational attention mechanism to achieve implicit dependency modeling.

$$q_i^s = h_{cls|s} w_i^q \tag{5}$$

$$k_i^e = h_{cls|e} w_i^k \tag{6}$$

$$v_i^e = h_{cls|e} w_i^v \tag{7}$$

$$\begin{aligned}z_e^s = &\frac{[q_1^s; \ldots; q_n^s][k_1^e; \ldots; k_n^e]^T}{\sqrt{d}}\\&* [v_1^e; \ldots; v_n^e]\end{aligned} \tag{8}$$

$$t_e^s \underset{\text{argmax}}{=} \text{softmax}(\text{MLP}(z_e^s))) \tag{9}$$

where q_i^s, k_i^e and v_i^e represent the ith Query of s, the ith Key and Value of e, respectively. $s \in [u_1, u_2, \ldots, u_n]$ and $e \in [u_1, \ldots, u_n, o_a, \ldots, o_d]$ represent start sentence and end sentence, respectively. w_i^q, w_i^k and w_i^v are all parameter vectors. z_e^s and $t_e^s \in \mathrm{T}$ represent the dependency vector and dependency type of s to e, respectively. T represents the set of dependency types, and $|\mathrm{T}|$ represents the number of dependencies. Note that T in Equation 8 represents the matrix transpose.

Graph Reasoning The goal of the reasoning module is to build a graph structure to complete the interaction between historical dialogues and candidate options. The graph structure allows messages to pass through nodes with different contextual information, which can fully consider local information for reasoning purposes. The Graph Convolutional Network (GCN) can achieve better performance in the reasoning task of QA (Ye et al., 2019; Fang et al., 2020; Qiu et al., 2019), and the reason it works is that the GCN can summarize the feature information of the local nodes. However, in traditional GCNs, the influence of different edge relationships on nodes is not considered, which leads to the same way of aggregating neighbor node information. To avoid this, we employ relational graph convolutional networks (Schlichtkrull et al., 2018), which help the model grasp the different dependencies between utterances and between utterances and options. The graph structure is created as follows:

- **Nodes**: The h_{cls} of each utterance and option act as a node in the graph.
- **Edges**: There are two different ways to build edges. An edge is constructed between each historical dialogue node in the graph, and an edge is constructed between each historical dialogue node and a candidate option (Figure 2). The number of edge types is T, which are determined by the dependencies between the two nodes related to this edge.

The graph modeling are now briefly described. In general, the input is a graph $G = (\mathcal{V}, \xi, \mathcal{R})$ with n nodes $v_i \in \mathcal{V}$, edge $e_{ij} = (v_i, v_j, r) \in \xi$, where $r \in \mathcal{R}$ is a relation type. A simple differentiable message-passing framework (Gilmer et al., 2017) is as follows:

$$h_i^{(l+1)} = \sigma(\sum_{m \in \mathcal{M}_i} g_m(h_i^{(l)}, h_j^{(l)})) \tag{10}$$

where $h_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ is the lth layer node representation of v_i and $d^{(l)}$ is the dimensionality of lth layer. The $g_m(\cdot,\cdot)$ function aggregates the incoming messages and passes them through the activation function $\sigma(\cdot)$, such as the $\mathrm{ReLU}(\cdot)$. $\mathcal{M}_i$ is the incoming message set for node v_i, usually chosen as the incoming edge set. Motivated by this architecture, a multi-relational graph message propagation model is defined as:

$$\begin{aligned} h_i^{(l+1)} =& \sigma(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} \\ &+ W_0^{(l)} h_i^{(l)}) \end{aligned} \tag{11}$$

where $\mathcal{N}_i^r$ represents the set of neighbor nodes whose relationship is $r \in \mathcal{R}$ for v_i. $c_{i,r}$ is a problem-specific normalization constant, where $c_{i,r} = |\mathcal{N}_i^r|$. $W_0^{(l)}$ and $W_r^{(l)}$ are the neighbor nodes of v_i and their corresponding parameter matrices respectively, which are used for linear transformation.

3.5 Prediction Layer

Finally, the final score is calculated by two linear layers plus an activation function, which is defined as:

$$s_o = W_2 * \mathrm{ReLU}(W_1 * O + b_1) + b_2 \tag{12}$$

where W_1, W_2, b_1 and b_2 are trainable parameters. O represents the vector of all options after ODC (After). s_o represents the score for all options. The loss function is cross entropy loss, defined as:

$$p_i = \frac{exp(s_{o^i})}{\sum_j exp(s_{o^j})} \tag{13}$$

$$\mathcal{L} = -\sum_i^N y_i log(p_i) \tag{14}$$

where y_i is the true label and N represents the number of samples in a batch.

4 Experiments

In this section, we conduct experiments on the MuTual dataset and MuTual^{plus} dataset to evaluate our proposed IRRGN. In all comparative experiments, in order to ensure the authenticity of the experimental results, all training hyperparameters are kept the same. Only adjust the learning rate when the model does not converge.

4.1 Experimental Settings

4.1.1 Datasets

Our proposed IRRGN is tested on MuTual and MuTual^{plus} datasets[2]. MuTual contains 8860 reasoning questions designed by language experts and professional annotators, which is constructed based on Chinese high school English listening test data. Each candidate is related to the historical dialogue, but only one is logically correct. MuTual^{plus} is more difficult to reasoning, which uses a safe response to replace one of the four candidate options in the original dataset. MuTual^{plus} is used to detect whether the model can choose a safe response when the other three candidate options are not logically correct.

4.1.2 Metrics

The evaluation metrics[3] are the same as those used in previous work. They are recall at position 1 in

[2]The datasets and leaderboard are available at: https://nealcly.github.io/MuTual-leaderboard/

[3]The evaluation code is available at: https://github.com/Nealcly/MuTual/blob/master/eval_sample/eval.py

Source	Method	MuTual			MuTual^{plus}		
		R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR
From paper (Cui et al., 2020)	Random	0.250	0.500	0.604	0.250	0.500	0.604
	TF-IDF (Paik, 2013)	0.279	0.536	0.542	0.278	0.529	0.764
	Dual LSTM (Lowe et al., 2015)	0.260	0.491	0.743	0.251	0.479	0.515
	SMN (Wu et al., 2017)	0.299	0.585	0.595	0.265	0.516	0.627
	DAM (Zhou et al., 2018)	0.241	0.465	0.518	0.272	0.523	0.695
	BERT (Devlin et al., 2019)	0.648	0.847	0.795	0.514	0.787	0.715
	RoBERTa (Liu et al., 2019)	0.713	0.892	0.836	0.626	0.866	0.787
	GPT2 (Radford et al., 2019)	0.332	0.602	0.584	0.316	0.574	0.568
	GPT2-FT (Radford et al., 2019)	0.392	0.670	0.629	0.226	0.611	0.535
	BERT-MC (Devlin et al., 2019)	0.667	0.878	0.810	0.580	0.792	0.749
	RoBERTa-MC (Liu et al., 2019)	0.686	0.887	0.822	0.643	0.845	0.792
From Mutual leaderboard	MUSN	0.912	0.983	0.953		-	
	CFDR	0.913	0.986	0.954	0.735	0.904	0.849
	GRN (Liu et al., 2021b)	0.915	0.983	0.954	0.841	0.957	0.913
	MDFN (Liu et al., 2021a)	0.916	**0.988**	0.956		-	
	BIDeN	0.930	0.983	0.962		-	
	Human	0.938	0.971	0.964	0.930	0.972	0.961
Ours	**IRRGN**	**0.939**	0.979	**0.965**	**0.845**	**0.962**	**0.916**

Table 1: Results on the test set of the two benchmark datasets. The top half includes eleven baseline models, and the bottom half includes recent studies on these two datasets.

4 candidate options (R_4@1), recall at position 2 in 4 candidate options (R_4@2) and Mean Reciprocal Rank (MRR) (Baeza-Yates and Ribeiro-Neto, 1999).

4.1.3 Baselines

Eleven baseline models were used for comparison. Besides traditional TF-IDF (Paik, 2013) and Dual LSTM (Lowe et al., 2015), it also includes Sequential Matching Network (SMN) (Wu et al., 2017), Deep Attention Matching Network(DAM) (Zhou et al., 2018), BERT and BERT-MC (Devlin et al., 2019), RoBERTa and RoBERTa-MC (Liu et al., 2019), GPT2 and GPT2-FT (Radford et al., 2019).

4.1.4 Parameter Settings

We utilize the open-source pre-trained model DeBERTa-V2$_{xxlarge}$ as the context encoder, which has 48 hidden layers, 1536 hidden-size and 24 attention heads. The $L2$ weight decays λ is set to 0.01. The maximum sequence length is 512. We use the AdamW optimizer to optimize the model parameters with a learning rate of 2e-6. The learning rate was changed with a cosine annealing strategy in ten epochs with batch size of 2. The total number of types T was set to 8. The model with the best performance on the validation set is set as the final model. We run the experiments on an A100 SXM4 GPU with 80G of memory. For more details on experimental parameter settings, please refer to our open-source code.

4.2 Experimental Results

4.2.1 Comparison with baselines

Table 1 reports the test results of IRRGN and the results of all models available for comparison. It can be observed that the R_4@1 metric of IRRGN significantly outperforms all compared models on both datasets, and more importantly, our proposed model outperforms human performance on all three metrics on the MuTual dataset, which proves that IRRGN has excellent reasoning ability. It is worth noting that the performance of traditional models (TF-IDF, DuLSTM, SMN and DMN) is relatively low, which indicates their insufficient reasoning ability. Pre-trained models (BERT and RoBERTa) improve in performance, but are still far behind human performance. Generative pre-trained models (GPT2) are not suitable for multi-turn dialogue reasoning problems. Our method achieves state-of-the-art performance compared to other studies (from MuTual leaderboard) on improving the reasoning ability of the model, which again validates that our method is effective. See the appendix A for the results of all baselines on the validation set.

Method	R_4@1	R_4@2	MRR
IRRGN	0.931	0.972	0.959
w/o ODC (After)	0.929	0.972	0.955
w/o ODC (Before)	0.925	0.975	0.954
w/o ODC	0.917	0.970	0.951
w/o URR	0.913	0.967	0.952
w/o ALL	0.904	0.964	0.946

Table 2: Ablation experimental results of GRN on MuTual validation set. -RAO Comparison: Remove the reasoning-after option comparison module (V). -RBO Comparison: Remove the reasoning-before option comparison module (II). -ODC: Remove Option Dual Comparison module (II and IV). -URR: Remove Utterance Relational Reasoner (III and IV). -ALL: Remove all modules.

To better verify the effectiveness of our method, we conduct ablation experiments and apply our method to other pre-trained language models for comparison.

4.2.2 Ablation Study

To get better insight into our IRRGN, we perform the ablation study. Specifically, five variants of IRRGN are designed: 1) **w/o ODC (After)**, the transformer encoder after the URR is removed; 2) **w/o ODC (Before)**, the transformer encoder before URR was removed; 3) **w/o ODC**, the transformer encoder before and after the URR is removed; 4) **w/o URR**, the relational attention and RGCN layers are removed. 5) **w/o ALL**, all components except pre-trained DeBERTa and Prediction are removed. The results are shown in Table 2. When ODC (After) or ODC (Before) is removed, the performance of the model decreases, which verifies the effectiveness of the dual comparison. It is worth noting that ODC (Before) appears to have a larger role than ODC (After), which is exactly what other studies overlook. When the ODC is removed, the performance of the model begins to drop significantly, which verifies that it is essential to capture the differences between the options before and after reasoning. When the URR is removed and the model shows a significant performance drop, which means that it is important for the reasoning ability of the model. Compared to using only the DeBERTa model (w/o ALL), our proposed IRRGN significantly enhances the performance on three metrics. It is worth noting that IRRGN only increases the amount of parameters by 2%, which can be observed in the code.

Number of RGCN layers l	R_4@1	R_4@2	MRR
4	0.900	0.975	0.945
3	0.877	0.980	0.935
2	0.931	0.971	0.959
1	0.882	0.972	0.935

Table 3: Results of different number of RGCN layers on MuTual validation set.

4.2.3 Generality of IRRGN

To test the generality of the proposed IRRGN, we apply it to a widely used pre-trained language model, which includes BERT_{base}, BERT_{large}, $\mathrm{RoBERTa}_{base}$, $\mathrm{RoBERTa}_{large}$, $\mathrm{ALBERTV2}_{base}$, and $\mathrm{ALBERTV2}_{large}$ (Lan et al., 2020). As shown in Figure 3, the performance of different pre-trained language models plus our IRRGN improves, which proves that IRRGN is generally effective.

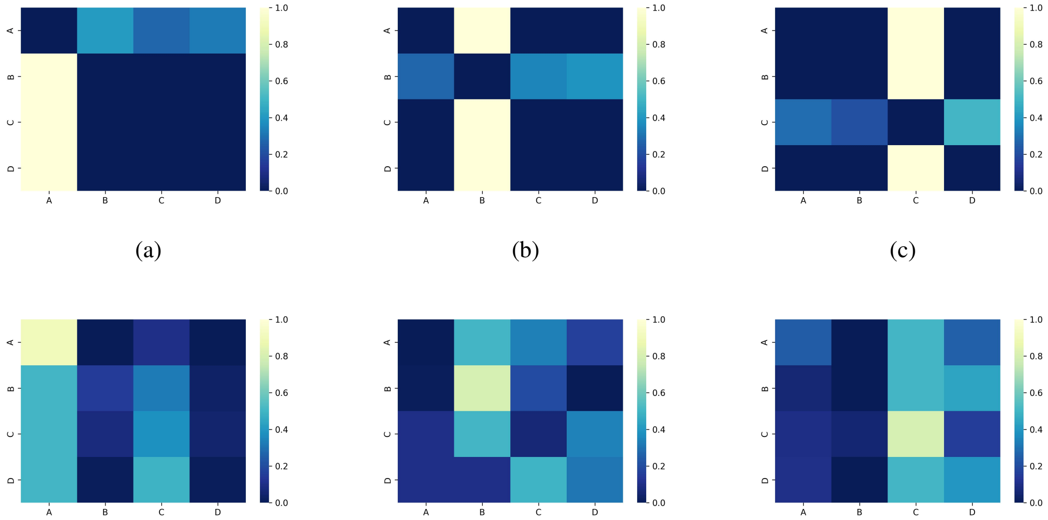
5 Analysis

5.1 Number of RGCN layers

Table 3 shows the effect of the number of RGCN layers on the performance of the model. It can be seen that when $l = 2$, the comprehensive performance of the model is the highest. This corresponds to what was analyzed in previous work (Klicpera et al., 2019), where the number of GCN layers is related to the depth of the graph and the sparsity of the adjacency matrix. The historical dialogue turns in the MuTual and MuTual^{plus} datasets are mostly within 5, which makes $l = 2$ more suitable for our model.

5.2 Number of Arc Types

To see the impact of the number of implicit arc types on the performance of the model, we experiment with it. As shown in Table 4, when $\mathrm{T} = 8$, the effect is the best. When $\mathrm{T} < 8$ or $\mathrm{T} > 8$, the performance of the model is weakened. We guess that on the MuTual and MuTual^{plus} datasets, 8 different arc types can model the dependencies between utterances and between utterances and options well. When $\mathrm{T} < 8$, the number of arc relations is not enough to express the number of dependencies, and when $\mathrm{T} > 8$, too much noise are introduced, which lead to the degradation of performance of the model. When $\mathrm{T} = 1$, the RGCN layer degenerates into the ordinary GCN layer.



(a) (b) (c)

(d) (e) (f)

Figure 3: Visualization of attention weights between different options. The upper row and lower row represent the attention weight maps in ODC (Before) and ODC (After), respectively. The correct answers from left to right are A, B, and C, from the example in Figure 1, dev_1 and dev_4, respectively.

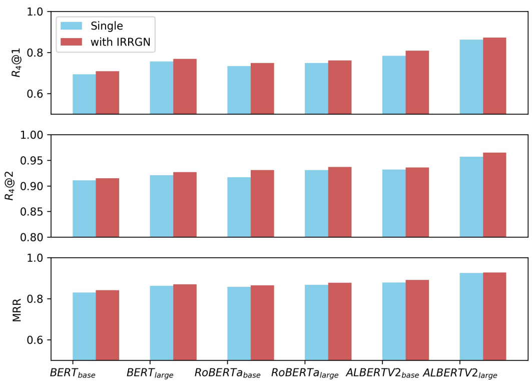


Figure 4: The $R_4@1$, $R_4@2$ and MRR performance of different pre-trained language models with IRRGN on MuTual validation set.

Number of arc types T	R_4@1	R_4@2	MRR
8	0.931	0.971	0.959
+1	-0.001	+0.006	-0.007
+2	-0.009	+0.006	-0.004
+3	-0.002	-0.007	+0.002
+4	-0.011	+0.001	-0.005
-1	0	+0.009	-0.002
-2	-0.004	-0.002	+0.003
-3	-0.002	-0.003	-0.005
-4	-0.006	+0.002	+0.001

Table 4: Results of different number of arc types on MuTual validation set. + and - represent addition and subtraction on the basis of $\mathrm{T} = 8$ and its three metrics, respectively.

5.3 Effect of Option Dual Comparator

In order to see the effect of ODC, we extract the attention weights between the options in ODC (Before) and ODC (After) for visualization, as shown in Figure 4. In the first column, correct option A in ODC (Before) focuses on options B, C, and D, which means that it is affected by the wrong options. However, in ODC (After), the correct option focuses on itself close to 1. For wrong options, they cannot focus on themselves, both in ODC (Before) and ODC (After), which shows that they can still be influenced by other options. The display of other columns is similar to the first column. It can be seen that the purpose of the ODC is to focus the correct option on itself.

6 Conclusion

In this paper, we propose a novel Implicit Relational Reasoning Graph Network (IRRGN). It can implicitly define dependencies between utterances, as well as utterances and options for more efficient and flexible graph reasoning. Among other things, it captures the differences between options before and after reasoning. State-of-the-art performance is achieved on the MuTual and MuTual^{plus} datasets that focus on the multi-turn dialogue reasoning task. In future work, we will further implement more fine-grained reasoning, explore model interpretability through bad cases, and let the model consider security responses.

7 Limitations

Although IRRGN outperforms all other models on these two datasets, there are still some points that can be improved.

- Fine-grained reasoning. Although IRRGN has excellent reasoning ability, it may not perceive more fine-grained reasoning. The nodes on the reasoning graph are at the utterance-level rather than the word-level, and we will use more Fine-grained reasoning clues to assist the dialogue selection task in the future.
- Security response. Like all other models, the performance of IRRGN on the MuTual^{plus} dataset is lower than that of the MuTual dataset. This suggests when none of the other candidate options are logical, how to choose a security response is worth researching.

Acknowledgements

We thank the anoymous reviewers for valuable and inspiring comments and suggestions.

References

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021. GoG: Relation-aware graph-over-graph network for visual dialog. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 230–243.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1406–1416. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4040–4054. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8823–8838. Association for Computational Linguistics.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Matthew Henderson, Ivan Vulic, Daniela Gerz, Iñigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5392–5404. Association for Computational Linguistics.

Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. 2021. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 327–339. Association for Computational Linguistics.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Predict then propagate: Graph neural networks meet personalized pagerank. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021a. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13406–13414. AAAI Press.

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. *CoRR*, abs/1907.11692.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. 2021b. A graph reasoning network for multi-turn response selection via customized pre-training. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13433–13442. AAAI Press.

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 SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.

Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 44–50. Association for Computational Linguistics.

Jiaul H. Paik. 2013. A novel TF-IDF weighting scheme for effective ranking. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 343–352. ACM.

Wei Peng, Yue Hu, Yuqiang Xie, Luxi Xing, and Yajing Sun. 2022a. Cogintac: Modeling the relationships between intention, emotion and action in interactive process from cognitive perspective. In *IEEE Congress on Evolutionary Computation, CEC 2022, Padua, Italy, July 18-23, 2022*, pages 1–8. IEEE.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022b. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *CoRR*, abs/2204.12749.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022c. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6140–6150. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers Inf. Technol. Electron. Eng.*, 19(1):10–26.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 22–31. Association for Computational Linguistics.

Chongyang Tao, Jiazhan Feng, Rui Yan, Wei Wu, and Daxin Jiang. 2021. A survey on response selection for retrieval-based dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4619–4626. ijcai.org.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1–11. Association for Computational Linguistics.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2704–2713. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xu Wang, Hainan Zhang, Shuai Zhao, Yanyan Zou, Hongshen Chen, Zhuoye Ding, Bo Cheng, and Yanyan Lan. 2021. FCM: A fine-grained comparison model for multi-turn dialogue reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4284–4293. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 496–505. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14176–14184. AAAI Press.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 55–64. ACM.

Deming Ye, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2019. Multi-paragraph reasoning with knowledge-enhanced graph neural network. *CoRR*, abs/1911.02170.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 86–90. Association for Computational Linguistics.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4429–4440. International Committee on Computational Linguistics.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3740–3752. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4524–4530. ijcai.org.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6669–6678. Computer Vision Foundation / IEEE.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan.

2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 372–381. The 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1118–1127. Association for Computational Linguistics.

Source	Method	MuTual			MuTual^{plus}		
		$R_4@1$	$R_4@2$	MRR	$R_4@1$	$R_4@2$	MRR
From paper (Cui et al., 2020)	Random	0.250	0.500	0.604	0.250	0.500	0.604
	TF-IDF (Paik, 2013)	0.276	0.541	0.541	0.283	0.530	0.763
	Dual LSTM (Lowe et al., 2015)	0.266	0.528	0.538		-	
	SMN (Wu et al., 2017)	0.274	0.524	0.575	0.264	0.524	0.578
	DAM (Zhou et al., 2018)	0.239	0.463	0.575	0.261	0.520	0.645
	BERT (Devlin et al., 2019)	0.657	0.867	0.803	0.514	0.787	0.715
	RoBERTa (Liu et al., 2019)	0.695	0.878	0.824	0.622	0.853	0.782
	GPT2 (Radford et al., 2019)	0.335	0.595	0.586	0.305	0.565	0.562
	GPT2-FT (Radford et al., 2019)	0.398	0.646	0.628	0.226	0.577	0.528
	BERT-MC (Devlin et al., 2019)	0.661	0.871	0.806	0.586	0.791	0.751
	RoBERTa-MC (Liu et al., 2019)	0.693	0.887	0.825	0.621	0.830	0.778
From leaderboard	GRN (Liu et al., 2021b)	0.935	0.985	0.971		-	
	MDFN (Liu et al., 2021a)	0.923	0.979	0.958		-	
Ours	IRRGN	0.930	0.971	0.959	0.863	0.958	0.924

Table 5: Results on the validation set of the two benchmark datasets. The top half includes eleven baseline models, and the bottom half includes recent studies on these two datasets.

A Result of the Baselines on the Validation Set

The performance of all baselines on the validation set is shown in Table 5. The performance of the traditional model is still not high. Some models on the leaderboard achieve relatively high performance.