Dialogues Are Not Just Text: Modeling Cognition for Dialogue Coherence Evaluation

Xue Li[1]*, Jia Su[2], Yang Yang[1]†, Zipeng Gao[3], Xinyu Duan[2], Yi Guan[1]

[1]Faculty of Computing, Harbin Institute of Technology
[2]Huawei Cloud
[3]School of Computer Science and Technology, University of Science and Technology of China
20s103245@stu.hit.edu.cn, {sujia3,duanxinyu}@huawei.com, yangyang_hit_wi@163.com,
gaozp619@mail.ustc.edu.cn, guanyi@hit.edu.cn,

Abstract

The generation of logically coherent dialogues by humans relies on underlying cognitive abilities. Based on this, we redefine the dialogue coherence evaluation process, combining cognitive judgment with the basic text to achieve a more human-like evaluation. We propose a novel dialogue evaluation framework based on Dialogue Cognition Graph (DCGEval) to implement the fusion by in-depth interaction between cognition modeling and text modeling. The proposed Abstract Meaning Representation (AMR) based graph structure called DCG aims to uniformly model four dialogue cognitive abilities. Specifically, core-semantic cognition is modeled by converting the utterance into an AMR graph, which can extract essential semantic information without redundancy. The temporal and role cognition are modeled by establishing logical relationships among the different AMR graphs. Finally, the commonsense knowledge from ConceptNet is fused to express commonsense cognition. Experiments demonstrate the necessity of modeling human cognition for dialogue evaluation, and our DCGEval presents stronger correlations with human judgments compared to other state-of-the-art evaluation metrics.

Introduction

Dialogue coherence evaluation is essential for research on open dialogue systems, which refers to the coherence and consistency of the content and structure of the dialogue (See et al. 2019; Ye et al. 2021). Dialogues exhibit higher coherence when the responses are fluent in the language, clear in meaning, context-sensitive, and logically tight. Human evaluation is widely used in modern dialogue systems, but it is expensive and time-consuming (Huang et al. 2020). When humans understand dialogues, cognition plays a crucial role (Sperber and Wilson 1986). The auto metrics used for automatic dialogue evaluation can be mainly categorized into three types. 1) Traditional evaluation metrics calculate lexical word-overlap between generated responses and reference responses, such as BLEU (Papineni et al. 2002) and ROUGE (Lin 2004). Such approaches overlook semantic information, leading to challenges in addressing response diversity.(Zhang et al. 2021; Deriu et al. 2021). 2) To address the above issue, recent works propose learnable evaluation metrics. They use the large-scale pre-trained language models to consider the semantic information of the dialogue (Sai et al. 2020; Zhao et al. 2022). 3) Besides, a few methods take some additional information into account when modeling the dialogue, such as the dialogue topic graph (Huang et al. 2020). However, there still exists a gap between these metrics and human evaluation (Zhao et al. 2022).

This is because dialogue is a cognitive activity that requires multiple cognitive abilities (Van Dijk 1984; Jang et al. 2013; Branigan et al. 2007), such as adherence to the human commonsense, seamless integration of context, temporal coherence, and non-confusing participant roles. Dialogue evaluation should not solely focus on the text itself but should also consider these cognitive abilities. However, none of the existing methods consider these cognitive abilities. As a result, these metrics fail to capture the corresponding cognitive errors in the generated responses of dialogue systems, including commonsense error, core-semantic error, temporal error, and role error.

Commonsense error: The generated response does not conform to commonsense, as shown in Figure 1 (a). **Core-semantic error**: The generated response deviates from the core semantic of the utterance due to neglecting crucial entities or misinterpreting fine-grained intent. As shown in Figure 1 (b), the second participant emphasized the word 'else' but the response ignored it leading to a wrong understanding of intent. **Temporal error**: The generated response ignores the previous dialogue content and does not continue the dialogue according to the temporal order. As shown in Figure 1 (c), the response ignores the second participant's utterance. **Role Error**: The generated response confuses its current role, responding from the perspective of another role. As shown in Figure 1 (d), the response is generated from the perspective of the second participant rather than the first participant. Usually, an unreasonable response may exhibit multiple of the four error types simultaneously, for example, temporal error is accompanied by role error.

To effectively capture four types of cognitive errors, we define four cognitive abilities that a qualified dialogue evaluation model should possess: commonsense cognition, core-semantic cognition, temporal cognition, and role cognition.

*This work was done during an internship at Huawei Cloud
†Corresponding author

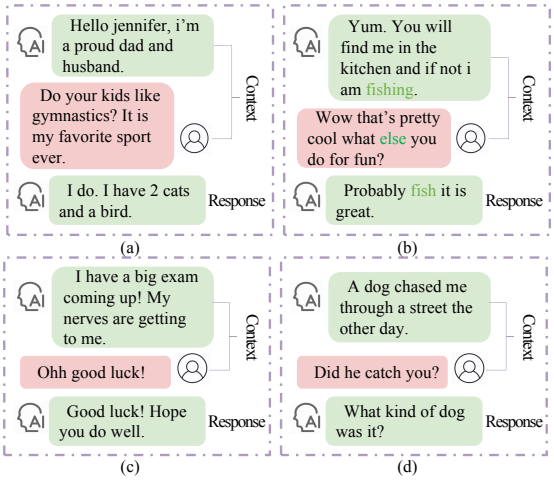

Figure 1: Examples from real datasets illustrating four cognitive error types, with 'Response' by DialogGPT or Transformer Ranker. Errors in (a)-(d) are commonsense, core-semantic, temporal, and role errors, respectively.

Modeling these cognitive abilities in the model enables dialogue evaluation to simultaneously consider core semantics, temporal aspects, role consistency, and implicit commonsense information in the dialogue. To this end, we need to address two critical issues: 1) How to model these cognitive abilities in a unified way for dialogue evaluation? 2) How to better integrate cognition information and text information to achieve mutual enhancement?

We propose a novel dialogue coherence evaluation framework based on Dialogue Cognition Graph (DCGEval), modeling four cognitive abilities in a unified way, in which role cognition is first considered. In order to effectively fuse cognition modeling and text modeling, we design a deep interaction module based on dual-process theory, which is widely used in describing the human thinking process (Daniel 2017). Specifically, core-semantic cognition is modeled by converting the utterance into an abstract meaning representation (AMR) graph, which can extract core semantic information without redundancy (Bai et al. 2021; Banarescu et al. 2013). The temporal and role cognition are modeled by establishing logical relationships among the different AMR graphs. And the commonsense knowledge from ConceptNet is fused to express commonsense cognition.

Our contributions are as follows:

- We redefine the process of dialogue coherence evaluation by combining cognitive judgment with the basic text to evaluate dialogue in a more human-like way. We design a novel neural framework based on the dual-process theory to implement the in-depth interaction between cognition modeling and text modeling.
- We define four cognitive abilities for dialogue coherence evaluation: commonsense cognition, core-semantic cognition, temporal cognition, and role cognition. Based on this, we design a novel AMR-based graph, DCG, to model these dialogue cognitive abilities.
- We demonstrate the effectiveness of the four cognitive abilities. Extensive experiments show that DCGEval has significantly stronger correlations with human judgments than other state-of-the-art (SOTA) metrics.

Related Work

Human Cognition and Discourse Modeling

(Van Dijk 1984) expounded some indispensable abilities in the dialogue process, which includes the following parts: 1) the necessity of temporal sequential modeling for dialogue understanding, 2) common world knowledge plays a decisive role, 3) sequence management in dialogue should be different from monological discourse sequences, which need to be considered role, 4) at any point of a dialogue a hearer needs to understand the actual semantic of the current speaker. (Pickering and Garrod 2004; Clark and Marshall 1981) argue that communication requires common world knowledge, (Schegloff and Sacks 1973; Sacks, Schegloff, and Jefferson 1978) indicate the importance of the role of participants in a dialogue, (Pickering and Garrod 2004) emphasize the temporal sequential nature of dialogue sequences, and (Evers-Vermeul, Hoek, and Scholman 2017) emphasize the temporal information is one of the prominent features that determine the coherence in discourse. In addition, Semantic information has consistently remained pivotal in discourse modeling, regardless of whether it pertains to monologue discourse or dialogue (Bai et al. 2021; Yeh, Wu, and Yang 2006).

Automatic Dialogue Evaluation

As traditional evaluation metrics are proven ineffective in dialog evaluation (Deriu et al. 2021), some learnable metrics are gradually proposed. Most metrics evaluate dialogue by directly considering the text representations obtained by pre-trained language models. BERTScore (Zhang et al. 2019), BERT-RUBER (Ghazarian et al. 2019), DEB (Sai et al. 2020), these methods use BERT (Kenton and Toutanova 2019) to encode the dialogue text. USR (Mehri and Eskenazi 2020b), FBD (Xiang et al. 2021) use RoBERTa (Liu et al. 2019) to encode the dialogue. FED (Mehri and Eskenazi 2020a) uses DialogGPT (Zhang et al. 2020) to measure 18 fine-grained qualities of dialog. They implicitly model semantics based on large-scale pre-trained models, which can easily overlook important entities (Bai et al. 2021).

Several methods consider additional information, which coincidentally implies certain cognitive abilities, leading to more reasonable dialogue evaluation results. GRADE (Huang et al. 2020) models the topic transition dynamics in dialogue by constructing a dialogue-level topic graph and incorporating commonsense information into the graph. DynaEval (Zhang et al. 2021) constructs an utterance-level dialogue graph for each dialogue, capturing dependencies between utterances, implicitly considering some temporal information. FlowEval (Li et al. 2021) models dialogue as segment act flow and evaluates the dialogue by encoding the segment act sequences, which may contain some temporal information. Therefore, the above methods are insufficient in four cognitive abilities modeling for dialogue evaluation.

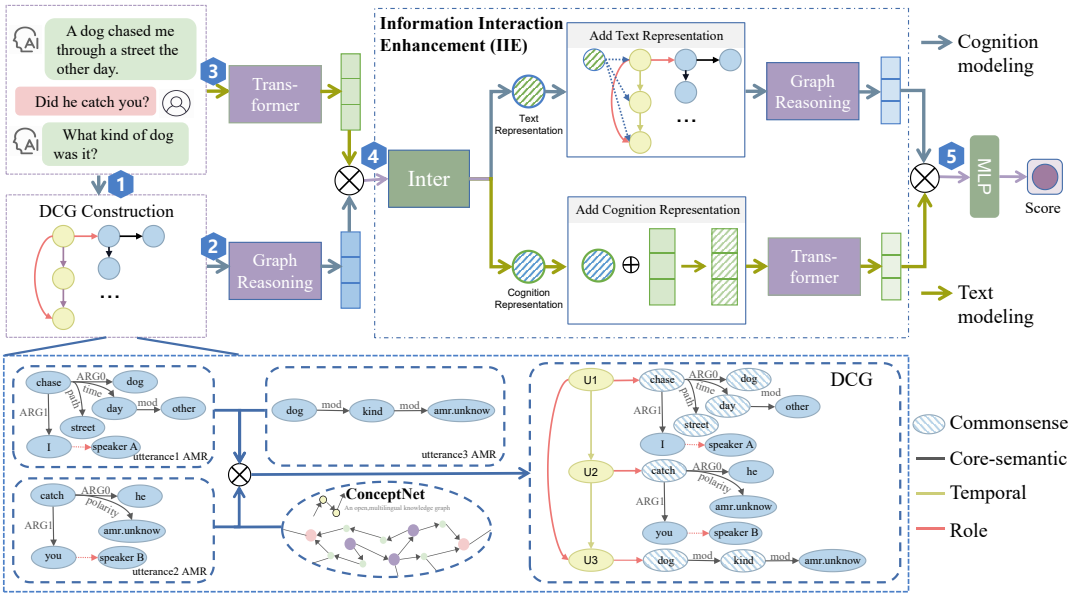

Figure 2: The figure illustrates the framework's key steps as numbers 1 to 5, with the lower part detailing the DCG construction process, further explained in Method. Specifically, the framework of DCGEval consists of two branches: text modeling and cognition modeling. The cognitive modeling branch constructs DCG based on the dialogue and ConceptNet, which models four cognitive abilities and fuses them through graph-based reasoning. The text modeling branch encodes the dialogue using a Transformer. Text and cognition encoding are fed into an information interaction enhancement (IIE) module to interact and enhance each other deeply. The enhanced text and cognition encoding are then concatenated and fed into an MLP to compute the final coherence score.

AMR

AMR (Banarescu et al. 2013) is a kind of semantic representation based on a directed acyclic graph to provide important concepts and explicit structure of sentences (Bai et al. 2021). AMR Parsing automatically transforms a sentence into an AMR graph applying for downstream tasks (Flanigan et al. 2014; Cai and Lam 2020; Cai et al. 2021). Recently, several works have employed AMR to model core-semantic representations (Xu et al. 2021a,b; Bai et al. 2021; Bonial et al. 2020), addressing concerns about large models inadvertently neglecting crucial details in processing lengthy texts. In this work, we aim to investigate the impact of AMR-based modeling of core-semantic ability on dialog evaluation.

Method

Problem Definition

We propose an automatic dialogue coherence evaluation framework that automatically evaluates responses generated by dialogue systems. Formally, let A and B denote the two speakers participating in the dialogue. Given a context, it can be regarded as a sequence containing n utterances, expressed as $c = \{U_1^A, U_2^B, ..., U_{n-1}^A, U_n^B\}$ and a response r, generated by the dialogue system. We add the response r as a new utterance to the utterance sequences of the context c and obtain the dialogue D as $D = \{U_1^A, U_2^B, ..., U_{n-1}^A, U_n^B, U_{n+1}^A\}$, where $U_{n+1}^A = r$. Our goal is to learn a function $f : (D) \rightarrow s$ to predict the coherence score s of the dialogue D.

Model

The proposed framework aims to simulate the dual-process theory of human decision-making (Daniel 2017) proposed by Daniel Kahneman. This theory posits two decision-making systems: the intuition system and the analytic system. The intuition system operates rapidly based on intuition under unconscious circumstances, while the analytic system operates under conscious control and requires cognitive reasoning before responding, making it slower. These systems interact and collectively influence human behavior. As shown in Figure 2, our framework consists of five components. The intuition system corresponds to Component 3, Text Modeling. It utilizes language models to make quick decisions based on empirical intuition. The analytic system corresponds to Components 1 and 2, Cognitive Modeling. It explicitly constructs the four cognitive abilities required for dialogue evaluation and unifies them in the form of a DCG due to their strong interdependence. Finally, these four cognitive abilities are integrated to perform graph-based cognitive reasoning for decision-making. The interaction between intuition and analytic systems corresponds to Component 4, Information Interaction Enhancement. Text modeling considers all words in a dialogue, while cognitive modeling aims to reason and extract key information. The interaction between these results in mutually enhancing their effects. The final decision-making corresponds to component 5, which provides the coherence score for the dialogue.

DCG Construction The bottom half of Figure 2 describes our process of constructing DCG. Since an unreasonable response may often contain multiples of the four error types, we use a unified graph DCG to simultaneously model the four cognitive abilities. First, we convert each utterance into an AMR graph to model core-semantic cognition. For example, in (b) in Figure 1, AMR can clearly model the 'else' in utterance 'U2', and accurately identify the fine-grained intent of 'U2' as 'what else' rather than 'what'. We implement it by using the AMR parser[1], which is pre-trained on BART(Lewis et al. 2019).

Then we model temporal and role cognition by establishing logical relationships among multiple utterance AMRs. For temporal cognition, we add a new node for each utterance to represent the index within the utterance sequence, ranging from 1 to $n+1$. The utterance index nodes are connected according to the temporal relationship between utterance pairs. In addition, We align role nodes in AMR to actual roles. For Speaker A, 'I' becomes 'Speaker A,' and 'you' becomes 'Speaker B.' This is applied to Speaker B's utterance AMRs as well.

Finally, we ground dialogue AMR graph nodes to ConceptNet. Specifically, we utilize the node embedding from ConceptNet as the initial embedding for the corresponding nodes in the AMR graph. When obtaining the ConceptNet node embedding, we consider the neighbor nodes of the current node to incorporate more commonsense information.

Graph Reasoning We obtain dialogue cognition representations by reasoning on the constructed DCG. For a given DCG, expressed as $G = (V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ represents denotes a set of nodes and E denotes a set of labeled edges, where $N > n$. We apply a graph convolution operation (Kipf and Welling 2016) on the DCG to aggregate four types of cognition information of neighbor nodes in the graph, the aggregated representation h_i^{l+1} at the layer $l+1$ for the node v_i is formulated as follows:

$$\{h_1^{(l+1)}, h_2^{(l+1)}, ..., h_N^{(l+1)}\} = GCN(\{h_1^{(l)}, h_2^{(l)}, ..., h_N^{(l)}\}) \tag{1}$$

$$h_i^{(l+1)} = \sigma(\hat{A} h_i^{(l)} W^{(l)}) \tag{2}$$

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \tag{3}$$

where $\tilde{A} = A + I$, A is the adjacency matrix of the DCG, I is the identity matrix, and $\tilde{D}$ is the degree matrix of $\tilde{A}$. h_i^l is the representation of the L-th layer of the node v_i, and σ is the nonlinear activation function. h_0^i is the node representation obtained by ConceptNet. We follow(Huang et al. 2020) uses k-hop neighbor nodes to obtain the node embedding of ConceptNet. We average the representation of each node in the graph to obtain a cognition representation H of the dialogue:

$$H = \frac{1}{N} \sum_{i=1}^{N} h_i \tag{4}$$

[1]https://github.com/bjascob/amrlib.

Text Representation For a given context c and response r, we obtain dialogue text representations by encoding c and r with transformer (Kenton and Toutanova 2019):

$$x_1, x_2, ..., x_L = Transformer(c, r) \tag{5}$$

$$X = \frac{1}{L} \sum_{i=1}^{L} x_i \tag{6}$$

where x_i is $1 \times d$, where L is the sum of the number of tokens in c and the number of tokens in r, and d is the dimension of embedding. X is the text representation of the entire dialogue, and X is $1 \times d$.

Information Interaction Enhancement Inspired by (Zhang et al. 2022), we design an Information Interaction Enhancement module (IIE) to facilitate deep interaction and mutual enhancement between text encoding and cognition encoding, thereby achieving the integration of cognitive judgment and text modeling. The information interaction enhancement module propagates information between cognition modeling and text modeling, enabling their deep interaction. Firstly, an information interaction layer (Inter) is designed to fuse the dialogue's text and cognition representations. We concatenate the text representation X and cognition representation H and obtain a joint representation through the information interaction layer. Subsequently, we separate the fused representations into h_{inter} and x_{inter}:

$$[h_{inter}; x_{inter}] = Inter([H; X]) \tag{7}$$

We implement $Inter$ using a multi-layer perceptron (MLP). Then we apply the fused representation h_{inter}, x_{inter} to cognition encoding and text encoding, respectively. Specifically, for cognition encoding, we add a text node v_{N+1} to the DCG, and its embedding is set to h_{inter}. This node establishes directed edges with all nodes in the graph. The direction of the edges is from the v_{N+1} to other nodes. The new DCG is expressed as $D\bar{C}G = (\bar{V}, \bar{E})$:

$$\bar{V} = \{v_1, v_2, ..., v_N, v_{N+1}\} \tag{8}$$

We perform graph reasoning on the new DCG to further facilitate a deeper interaction between text and cognition information, resulting in a text-enhanced dialogue cognition representation denoted as $\bar{H}$:

$$\{\bar{h}_1, \bar{h}_2, ..., \bar{h}_N, \bar{h}_{N+1}\} = GCN(\{h_1, h_2, ..., h_N, h_{N+1}\}) \tag{9}$$

$$\bar{H} = \frac{1}{N+1} \sum_{i=1}^{N+1} \bar{h}_i \tag{10}$$

On the other hand, for text encoding, we incorporate x_{inter} into each token representation and utilize a transformer to update them, resulting in a cognition-enhanced dialogue text representation denoted as $\bar{X}$.

$$\bar{x}_1, ..., \bar{x}_L = Transformer(x_1 + x_{L+1}, ..., x_L + x_{L+1}) \tag{11}$$

$$\bar{X} = \frac{1}{L} \sum_{i=1}^{L} \bar{x}_i \tag{12}$$

	ConvAI2				EmpatheticDialogues			
Metric	Pearson	Spearman	Kendall	Average	Pearson	Spearman	Kendall	Average
BLEU	0.003*	0.128	0.088	0.073	-0.051*	0.002*	0.005*	-0.015
ROUGE	0.136	0.140	0.097	0.124	0.029*	-0.013*	-0.010*	0.002
METEOR	0.145	0.181	0.123	0.15	0.118	0.055*	0.04*	0.071
BERTScore	0.225	0.225	0.154	0.201	0.046*	0.033*	0.021*	0.033
ADEM	0.026*	0.037*	0.049*	0.037	0.007*	0.009*	0.040*	0.019
BERT-RUBER	0.266	0.266	0.185	0.239	-0.022*	-0.040*	-0.029*	-0.030
BLEURT	0.152	0.149	0.103	0.135	0.203	0.192	0.13	0.175
QuantiDCE	0.554	0.554	0.395	0.501	0.412	0.393	0.274	0.360
DynaEval	0.066	0.070	0.047	0.061	0.071	0.066	0.045	0.061
GRADE	0.496	0.503	0.356	0.452	0.350	0.344	0.243	0.312
ChatGPT	0.498	0.515	0.374	0.462	0.407	0.358	0.283	0.349
DCGEval	**0.562**	**0.572**	**0.406**	**0.513**	**0.436**	**0.436**	**0.306**	**0.393**

Table 1: Correlations between automatic evaluation metrics and human judgments on two datasets (ConvAI2 and EmpatheticDialogues). The results are the average of five experiments, each with a different random seed. The star * indicates results with p-value > 0.05, which are not statistically significant.

Score Prediction In the score prediction, we predict a score based on text representation $\bar{X}$ and cognition representation $\bar{H}$ obtained by the information interaction enhancement module. The text representation $\bar{X}$ and cognition representation $\bar{H}$ are concatenated and fed into an MLP to transform the cognitively enhanced dialogue representation into a score s.

$$s = MLP(\bar{X}; \bar{H}) \tag{13}$$

Training

We employ the relative ranking MLR loss in the pretraining phase and the absolute scoring KD-MSE loss in the finetuning phase as training objectives. The design aims to balance the inconsistency between using ranking as a training objective due to limited labeled data and the expected behavior (absolute score) of the model. A more detailed explanation of these two losses is provided in the Appendix. Here, we present the overall formulas:

$$L_{pre\_train} = \frac{1}{N_1} \sum_{i=1}^{N_1} L_{MLR}(s_i^{pre-train}) \tag{14}$$

$$L_{fine\_tune} = \frac{1}{N_2} \sum_{i=1}^{N_2} L_{KD-MSE}(s_i^{fine-tune}, \bar{s}_i) \tag{15}$$

where N_1 and N_2 are the total number of samples in the dataset for pre-training and fine-tuning, respectively, $s_i^{pre-train}$, $s_i^{fine-tune}$ and $\bar{s}_i$ are the score of the $i-th$ dialogue sample from the pre-training stage, fine-tuning stage and human-annotated, respectively.

Experiment

Experimental Setup

Baseline We compare our evaluation metrics with eleven popular automatic dialogue evaluation metrics, including three lexical word-overlap metrics: BLEU, ROUGE, and METEOR (Banerjee and Lavie 2005), five metrics that consider semantic representation: BERTScore, ADEM (Lowe et al. 2017), BERT-RUBER, BLEURT, QuantiDCE (Ye et al. 2021), two metrics that take into account additional information about the dialogue: DynaEval, GRADE, and ChatGPT. **Evaluation** The common practice to show the effectiveness of a dialogue evaluation metric is to calculate the correlation between the model-predicted and the human-rated scores (Zhang et al. 2021; Huang et al. 2020). Specifically, we adopt Pearson, Spearman (Zar 2005) and Kendall (Kendall 1938) as the correlation measures. **Datasets** We use two daily dialogue datasets, DailyDialog++ (Sai et al. 2020) and DailyDialogEVAL (Huang et al. 2020), as training data. To evaluate model performance, we use ConvAI2 (Huang et al. 2020) and EmpatheticDialogues (Huang et al. 2020) as unseen datasets, including substantial human scoring. For more data details, please refer to the Appendix.

Experiment Results

The results in Table 1 show that DCGEval achieves an improvement on both ConvAI2 and EmpatheticDialogues datasets. This suggests modeling cognitive abilities in DCGEval enables a more human-like approach to dialogue evaluation. On ConvAI2, DCGEval improves the average correlation by at least 1.2%. Surprisingly, for the EmpatheticDialogues dataset, DCGEval achieves an absolutely significant improvement over the previous SOTA QuantiDCE. The correlations of Pearson, Spearman, and Kendall have increased by 2.4%, 4.3%, and 3.2%, respectively, with an average increase of 3.3%.

DCGEval also performs better than DynaEval and GRADE, the models that imply temporal and commonsense information to model dialogue. As a result, DCGEval achieved an absolute averaged correlation improvement of 6.1% and 8.1% over GRADE on the ConvAI2 and EmpatheticDialogues datasets, respectively. DynaEval achieved an extremely low correlation, there are two possible causes. One reason is that DynaEval treats each utterance as a sin-

	ConvAI2				EmpatheticDialogues			
Metric	Pearson	Spearman	Kendall	Average	Pearson	Spearman	Kendall	Average
DCGEval	**0.562**	**0.572**	**0.406**	**0.513**	**0.436**	**0.436**	**0.306**	**0.393**
Without commonsense	0.549	0.550	0.392	0.497	0.402	0.383	0.267	0.350
Without core-semantic	0.549	0.555	0.396	0.500	0.421	0.425	0.291	0.379
Without temporal	0.549	0.550	0.392	0.497	0.404	0.401	0.281	0.362
Without role	0.534	0.538	0.380	0.484	0.426	0.428	0.299	0.384
Without temporal,role	0.531	0.531	0.378	0.480	0.409	0.395	0.276	0.360
Without DCG	0.507	0.514	0.363	0.461	0.338	0.321	0.220	0.293
Without IIE (concat)	0.512	0.517	0.369	0.466	0.374	0.343	0.237	0.318
Without IIE (MLP)	0.514	0.530	0.377	0.473	0.395	0.385	0.264	0.348
Without IIE (h_{inter})	0.543	0.550	0.393	0.495	0.418	0.390	0.268	0.358
Without IIE (x_{inter})	0.556	0.564	0.404	0.508	0.404	0.376	0.26	0.346
Without IIE (h_{inter};x_{inter})	0.545	0.565	0.402	0.504	0.383	0.362	0.251	0.332

Table 2: The extensive ablation experiments on the ConvAI2 and Empathetic Dialogues datasets yielded results, including six ablations on the DCG graph and five ablations on the IIE module. The results are the average of five experiments, each with a different random seed.

gle node in the graph, which can lead to neglecting core-semantic information. Additionally, DynaEval is more focused on evaluating long dialogues, and for short dialogues, it constructs a fully connected graph where temporal and role information becomes confused.

Besides, we also utilize ChatGPT for this task. We meticulously craft a series of prompts. Compared with other baselines, ChatGPT evaluation results have shown good performance and are easy to operate. Nevertheless, there is still a certain gap compared to our model. We conduct a detailed analysis of the results in the Appendix. These results demonstrate that explicit modeling of four cognitive abilities and utilizing the interaction enhancement module can lead to a more accurate evaluation of dialogue coherence.

Ablation Studies

We conduct ablation studies on DCGEval to better analyze the impact of four cognitive abilities and deep interactions on dialogue evaluation. **Does the DCG work?** To demonstrate the necessity of modeling cognitive abilities for dialogue evaluation, we perform six ablations on the DCG graph: 1) Only remove commonsense cognition, expressed as **Without commonsense**. We employ word2vec for node initialization instead of commonsense knowledge. 2) Only remove core-semantic cognition, expressed as **Without core-semantic**. We replace the edges of each utterance AMR with a set of randomly generated edges and randomly remove some nodes, which destroys the ability of AMR to model core semantics. 3) Only remove the temporal cognition, expressed as **Without temporal**. 4) Only remove the role cognition, expressed as **Without role**. 5) Remove the temporal and role cognition, expressed as **Without temporal,role**. 6) Remove the entire DCG graph, expressed as **Without DCG**.

As shown in Table 2, all six ablations result in varying degrees of correlation decline, and the performance decrease is positively correlated with the amount of ablated information. The more information was ablated, the greater the performance degradation. Specifically, the most significant correlation decrease was observed in Without DCG, with an average reduction of 7.6% on both datasets. Furthermore, the decrease in Without temporal,role is also more significant compared to when only one of them was ablated. Therefore, these results provide strong evidence that the modeling of any cognitive ability is indispensable.

Does the information in-depth interaction work? We conduct detailed ablation experiments on the IIE module, which aims to obtain dialogue representations after interacting with cognition and text modeling. As shown in Table 2, we design five other methods to obtain dialogue representations: 1) **Without IIE (concat)**, the initial cognitive representation H and the text representation X are directly concatenated as the dialogue representation, without any interaction. 2) **Without IIE (MLP)**, a two-layer MLP is employed as a replacement. 3) **Without IIE (h_{inter})**, h_{inter} obtained after the interaction layer Inter as the dialogue representation. 4) **Without IIE (x_{inter})**, x_{inter} obtained after the interaction layer Inter as the dialogue representation. 5) **Without IIE ($h_{inter}; x_{inter}$)**, the concatenation of h_{inter} and x_{inter} as the dialogue representation.

From the results, it is evident that all five ablations led to a decrease in model performance. The most significant drop is observed in the method without any interaction (Without IIE (concat)), indicating that using an MLP to replace IIE retains some ability for information exchange. Moreover, using representations obtained after the interaction layer for score prediction led to decreased performance as well. This demonstrates the necessity of the IIE module for facilitating deep interaction.

Granular Analyses of the Four Cognitive Abilities

We conduct more in-depth experiments and granular analyses of the four cognitive abilities on the EmpatheticDialogues dataset. Specifically, we divide the EmpatheticDialogues dataset into four sub-datasets based on human annotation for four dialogue error types: Commonsense Error,

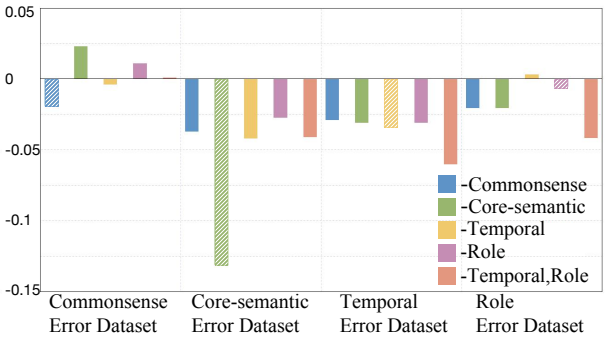

Figure 3: Experiment results of fine-grained analysis of four cognitive abilities. The horizontal axis represents four different data subsets, and the bars represent the results of five ablation experiments. Above and below the horizontal axis represent the performance increase and decrease, respectively. The height of the column represents the degree of change. The shaded bars represent experiments where the corresponding ability was ablated on the current dataset. All comparisons are made against the model without any ablated modules on the same sub-dataset.

Core-semantic Error, Temporal Error, and Dialogue Role Error Dataset. Across these subsets, we perform five ablation experiments, involving the individual removal of each ability and a combined ablation of temporal and role abilities. As shown in Figure 3, the horizontal axis is divided into four sections corresponding to the four sub-datasets mentioned above. Each section contains five bars representing the results of each ablation. All comparisons are made against the model without any ablated modules on the same sub-dataset.

From the results in Figure 3, we can deduce that ablating the corresponding ability for each sub-dataset leads to the greatest drop in model performance, indicating the effectiveness of the four cognitive abilities considered in the model. This is demonstrated in the Commonsense errors Dataset, Core-semantic errors Dataset, and Temporal errors Dataset. In the Role Error Dataset, ablating commonsense and core-semantic abilities results in substantial performance degradation, indicating that these abilities are crucial not only for the entire dataset but also for each sub-dataset. In the Role Error Dataset, ablating role ability leads to a performance drop, though not the most significant one. The joint ablation of role and temporal abilities causes the largest performance degradation, suggesting that when commonsense, core-semantic, and role cognitive ability coexist, some role information can be considered. It is also verified that role errors often occur together with temporal error types.

Case Study

Table 3 presents the scores of different methods on cases of four typical cognitive error responses. The closer the scores are to human scores, the more accurate the method. This suggests that the model can closely align with human judgment. Regarding the four types of cognitive errors, existing models show a significant gap between their scores and human scores. Conversely, the scores of our model are closest to human scores. Specifically, the differences between our model scores and human scores in the four examples are 0.29, 0.42, 0.21, and 0.67. The previous SOTA model had differences of 0.64, 1.44, 0.64, and 1.11 with human scores in the same examples. Notably, our approach has reduced the gap by 3 times compared to the previous best method in terms of core-semantic error. Results show that our model can better capture cognitive errors in dialogue, leading to more accurate and human-like evaluations. This demonstrates the necessity of incorporating these four cognitive abilities into dialogue evaluation.

Error type	Human	**DCGEval** / ChatGPT QuantiDCE / GRADE Score
Commonsense error	1.9	**2.19** / (1.5, 1.8) / 2.54 / 3.32
Core-semantic error	2.5	**2.92** / (1.8, 2.8) / 3.94 / 4.27
Temporal error	2.89	**3.10** / (2.8, 4.5) / 3.53 / 3.97
Role error	2.4	**3.07** / (4.7, 4.8) / 3.51 / 4.48

Table 3: Case study of four types of cognitive errors in Figure 1, including human scoring, our model scoring, and the scoring of existing metrics. A score closer to the human scores indicates a more accurate model. The score of ChatGPT is the result of two runs under the same configuration.

The results show that the output of ChatGPT with the same configuration is often unstable, showing significant score variations. Specifically, ChatGPT can effectively capture commonsense and core semantic errors, but role and temporal errors are more challenging for it. For a detailed analysis, please refer to the Appendix.

Conclusion and Discussion

In this paper, we present DCGEval, a novel framework for evaluating dialogue coherence that combines cognitive judgment with text information. Overall, we implement the in-depth interaction between cognition modeling and text modeling. Specifically, we design a new graph structure called DCG, which uniformly and explicitly models four types of dialog cognitive abilities, including commonsense cognition, core-semantic cognition, temporal cognition, and role cognition. Experiments demonstrate the effectiveness of the cognitive abilities to capture four kinds of cognitive errors in the process of dialogue evaluation, which is lacking in existing dialogue evaluation metrics. Finally, empirical results show that DCGEval has stronger correlations with human judgments. A limitation of DCGEval is that it relies on a pre-trained AMR parser to generate AMR graphs. AMR parsers have shown good performance, but there is still a risk of using an incorrect AMR graph. This error is usually on relation types rather than the entities. Therefore, it has little impact on DCGEval which does not require relation types.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 72293584), and the National Key Research and Development Program of China (Grant Nos. ZDYF20220008 and ZDYF20220003-01).

References

Bai, X.; Chen, Y.; Song, L.; and Zhang, Y. 2021. Semantic Representation for Dialogue Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4430–4445.

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 178–186.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bonial, C.; Donatelli, L.; Abrams, M.; Lukin, S.; Tratz, S.; Marge, M.; Artstein, R.; Traum, D.; and Voss, C. 2020. Dialogue-AMR: abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 684–695.

Branigan, H. P.; Pickering, M. J.; McLean, J. F.; and Cleland, A. A. 2007. Syntactic alignment and participant role in dialogue. *Cognition*, 104(2): 163–197.

Cai, D.; and Lam, W. 2020. AMR parsing via graph-sequence iterative inference. *arXiv preprint arXiv:2004.05572*.

Cai, D.; Li, X.; Ho, J. C.-S.; Bing, L.; and Lam, W. 2021. Multilingual amr parsing with noisy knowledge distillation. *arXiv preprint arXiv:2109.15196*.

Clark, H. H.; and Marshall, C. R. 1981. Definite knowledge and mutual knowledge.

Daniel, K. 2017. *Thinking, fast and slow*.

Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; and Cieliebak, M. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54: 755–810.

Evers-Vermeul, J.; Hoek, J.; and Scholman, M. C. 2017. On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse*, 8(2): 1–20.

Flanigan, J.; Thomson, S.; Carbonell, J. G.; Dyer, C.; and Smith, N. A. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1426–1436.

Ghazarian, S.; Wei, J.; Galstyan, A.; and Peng, N. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 82–89.

Huang, L.; Ye, Z.; Qin, J.; Lin, L.; and Liang, X. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. *arXiv preprint arXiv:2010.03994*.

Jang, G.; Yoon, S.-a.; Lee, S.-E.; Park, H.; Kim, J.; Ko, J. H.; and Park, H.-J. 2013. Everyday conversation requires cognitive inference: neural bases of comprehending implicated meanings in conversations. *NeuroImage*, 81: 61–72.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, Z.; Zhang, J.; Fei, Z.; Feng, Y.; and Zhou, J. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 128–138.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Mehri, S.; and Eskenazi, M. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Mehri, S.; and Eskenazi, M. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pickering, M. J.; and Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2): 169–190.

Sacks, H.; Schegloff, E. A.; and Jefferson, G. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, 7–55. Elsevier.

Sai, A. B.; Mohankumar, A. K.; Arora, S.; and Khapra, M. M. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8: 810–827.

Schegloff, E. A.; and Sacks, H. 1973. Opening up closings.

See, A.; Roller, S.; Kiela, D.; Weston, J.; and Stanford, N. 2019. What makes a good conversation. *How controllable attributes affect human judgments. CoRR abs/1902.08654*.

Sperber, D.; and Wilson, D. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Van Dijk, T. A. 1984. Dialogue and cognition. In *Cognitive Constraints on Communication: Representations and Processes*, 1–17. Springer.

Xiang, J.; Liu, Y.; Cai, D.; Li, H.; Lian, D.; and Liu, L. 2021. Assessing Dialogue Systems with Distribution Distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2192–2198.

Xu, W.; Deng, Y.; Zhang, H.; Cai, D.; and Lam, W. 2021a. Exploiting reasoning chains for multi-hop science question answering. *arXiv preprint arXiv:2109.02905*.

Xu, W.; Zhang, H.; Cai, D.; and Lam, W. 2021b. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering. *arXiv preprint arXiv:2105.11776*.

Ye, Z.; Lu, L.; Huang, L.; Lin, L.; and Liang, X. 2021. Towards quantifiable dialogue coherence evaluation. *arXiv preprint arXiv:2106.00507*.

Yeh, J.-F.; Wu, C.-H.; and Yang, M.-Z. 2006. Stochastic discourse modeling in spoken dialogue systems using semantic dependency graphs. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 937–944.

Zar, J. H. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Zhang, C.; Chen, Y.; D'Haro, L. F.; Zhang, Y.; Friedrichs, T.; Lee, G.; and Li, H. 2021. DynaEval: Unifying turn and dialogue level evaluation. *arXiv preprint arXiv:2106.01112*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.

Zhao, J.; Li, Y.; Du, W.; Ji, Y.; Yu, D.; Lyu, M. R.; and Wang, L. 2022. FlowEval: A Consensus-Based Dialogue Evaluation Framework Using Segment Act Flows. *arXiv preprint arXiv:2202.06633*.