

GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems

Lishan Huang^{1*}, Zheng Ye^{1*}, Jinghui Qin¹, Liang Lin^{1,2}, Xiaodan Liang^{1,2†}

¹ Sun Yat-Sen University, ² Dark Matter AI Inc.

{huanglsh6,yezh7,qinjinh}@mail2.sysu.edu.cn,
linliang@ieee.org, xdliang328@gmail.com

Abstract

Automatically evaluating dialogue coherence is a challenging but high-demand ability for developing high-quality open-domain dialogue systems. However, current evaluation metrics consider only surface features or utterance-level semantics, without explicitly considering the fine-grained topic transition dynamics of dialogue flows. Here, we first consider that the graph structure constituted with topics in a dialogue can accurately depict the underlying communication logic, which is a more natural way to produce persuasive metrics. Capitalized on the topic-level dialogue graph, we propose a new evaluation metric **GRADE**, which stands for **Graph-enhanced Representations for Automatic Dialogue Evaluation**. Specifically, GRADE incorporates both coarse-grained utterance-level contextualized representations and fine-grained topic-level graph representations to evaluate dialogue coherence. The graph representations are obtained by reasoning over topic-level dialogue graphs enhanced with the evidence from a commonsense graph, including k-hop neighboring representations and hop-attention weights. Experimental results show that our GRADE significantly outperforms other state-of-the-art metrics on measuring diverse dialogue models in terms of the Pearson and Spearman correlations with human judgements. Besides, we release a new large-scale human evaluation benchmark to facilitate future research on automatic metrics.

1 Introduction

Coherence, what makes dialogue utterances unified rather than a random group of sentences, is an essential property to pursue an open-domain

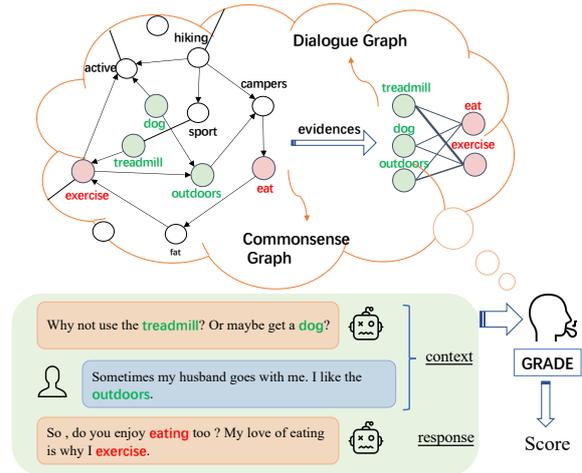


Figure 1: An illustrative example of how our GRADE evaluates dialogue coherence by incorporating graph information on topic transitions from a commonsense graph. Topic keywords of the context and the response are highlighted in green and red respectively, which can be aligned to the corresponding nodes in the commonsense graph. The white nodes and all the edges in the commonsense graph are pieces of evidence that assist in constructing the dialogue graph. Taking advantage of such evidence, GRADE can better capture the topic transition dynamics between the context and the response, as shown in the thickness of edges in the dialogue graph.

dialogue system aiming at conversing with humans. Although open-domain dialogue systems have achieved significant progress and performed much more human-like skills in recent years (Zhou et al., 2020; Adiwardana et al., 2020; Roller et al., 2020), automatically measuring dialogue coherence for state-of-the-art open-domain dialogue models is still an open and under-explored research problem attributing to the open-ended nature of dialogue (See et al., 2019).

Statistic-based automatic metrics, such as BLEU (Papineni et al., 2002), mostly rely on the degree of word overlap between a dialogue response and its corresponding gold response. However,

*Equal Contribution.

†Corresponding Author.

due to the ignorance of the underlying semantic of a response, they are biased and correlate poorly with human judgements in terms of response coherence (Liu et al., 2016). To overcome this issue, some learning-based metrics were proposed to train a coherence scoring model by considering the utterance-level semantics, such as ADEM (Lowe et al., 2017), RUBER (Tao et al., 2018), and BERT-RUBER (Ghazarian et al., 2019). However, a coherent real-world dialogue should be not only coherent among utterances but also smooth at topic transition. As shown in Figure 1, the topics inside a coherent dialogue are close to each other in the commonsense graph, which embodies a smooth topic transition. Although the above metrics have demonstrated higher correlations with human judgements than statistic-based metrics, they only model dialogue coherence at utterance level without explicitly considering the fine-grained topic transition dynamics of dialogue flows.

To address the above problems, we propose a new automatic metric for open-domain dialogue systems, named as **Graph-enhanced Representation for Automatic Dialogue Evaluation (GRADE)**, which explicitly models topic transition dynamics by reasoning over dialogue graphs and incorporates them into utterance-level contextualized representations. As a result, our method can capture more accurate semantic transition information, thus measuring dialogue coherence in a more human-like manner.

Specifically, our GRADE consists of two semantic extraction branches. One branch deploys BERT (Devlin et al., 2019) to learn the coarse-grained utterance-level contextualized representations, while another learns the fine-grained topic-level graph representations by constructing topic-level dialogue graphs and applying a graph neural network on the graphs to model the topic transition dynamics. As to the dialogue graph construction, we determine nodes and edges by utilizing the evidence from the commonsense knowledge graph, ConceptNet (Speer et al., 2017), including k-hop neighboring representations and hop-attention weights. GRADE is trained in an unsupervised manner with data automatically generated by a negative sampling strategy considering both lexical and semantic aspects rather than random sampling adopted by previous works (Tao et al., 2018; Ghazarian et al., 2019). Experimental results show that GRADE significantly outperforms other state-of-

the-art metrics in terms of the Pearson and Spearman correlations with human judgements and can generalize to unseen chat datasets well.

Our contributions are summarized as follows:

- We propose GRADE, a novel automatic coherence metric for evaluating open-domain dialogue systems, which is the first attempt to introduce graph reasoning into dialogue evaluation.
- We demonstrate the effectiveness of incorporating graph information into dialogue evaluation. Extensive experiments show that GRADE has significantly stronger correlations with human judgements than other state-of-the-art metrics.
- We construct and release a new large-scale human evaluation benchmark with 11910 human annotations to the research community for encouraging future study on automatic metrics.

The code and data are available at <https://github.com/li3cmz/GRADE>.

2 Related Work

Automatic evaluation for open-domain dialogue systems is difficult since there are many appropriate responses for a dialogue context under the open-domain setting, known as the one-to-many problem (Zhao et al., 2017).

Initially, the statistic-based metrics in language generation tasks are adopted for dialogue evaluation, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004). These metrics use statistical rules to measure the surface similarity between generated responses and reference responses. For example, BLEU computes the geometric average of the n-gram precisions. However, they can not cope with the one-to-many problem and have weak correlations with human judgements (Liu et al., 2016).

In recent years, learning-based metrics have increasingly attracted interest from researchers. ADEM proposed by Lowe et al. (2017) achieves higher correlations with human judgements than the statistic-based metrics, which is trained with human-annotated data in a supervised manner. However, it is time-consuming and expensive to obtain large amounts of annotated data. To reduce the cost of obtaining annotated data, Tao et al. (2018) trained their metric RUBER with auto-constructed negative samples in an unsupervised manner.

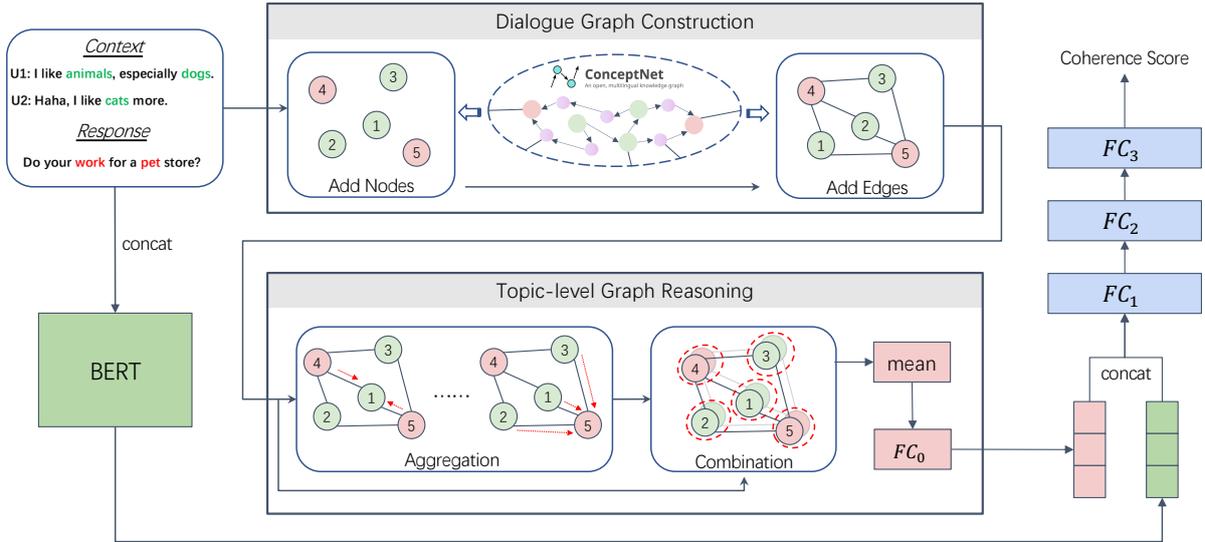


Figure 2: The architecture of GRADE consists of two semantic extraction branches. One branch encodes the context-response pair with BERT, while the other constructs a topic-level dialogue graph for the pair by utilizing the evidence from ConceptNet and performs reasoning over the constructed graph. The representations from the two branches are concatenated and fed into a MLP to compute the final coherence score. Note that the **green** and **red** nodes are corresponding to the keywords in the context and the response respectively.

With the advances of the pre-trained language model, BERT (Devlin et al., 2019) has been adopted for dialogue or NLG evaluation. Ghazarian et al. (2019) proposed BERT-RUBER, which outperforms RUBER significantly by incorporating BERT embeddings. BERTScore (Zhang et al., 2020) performs soft-overlap between candidate and reference sentences by using BERT embeddings directly without fine-tuning, and has been shown to correlate with human judgment robustly. Besides, Sellam et al. (2020) introduced BLEURT by further training regular pre-trained BERT with an elaborate pre-training scheme and fine-tuning on small amounts of rating data, which yields superior results.

Note that our model differs from the above learning-based metrics in two folds. First, our metric is trained with high-quality negative samples that are similar to the ground truths in both lexical and semantic aspects instead of randomly sampling. Second, different levels of representations are considered in our GRADE, especially the fine-grained topic-level graph representation.

3 GRADE Metric

In this paper, we focus on designing an evaluation metric that can automatically assess the coherence of responses produced by dialogue models. Formally, given a dialogue context $c = \{c_1, \dots, c_m\}$ and a response $r = \{r_1, \dots, r_n\}$, where each c_k

is a token in the context and each r_k is a token in the response, our goal is to learn a function $f : (c, r) \rightarrow s$ that predicts the coherence score s .

As illustrated in Figure 2, our GRADE predicts a coherence score s between a context c and a response r in three steps: (1) producing the utterance-level contextualized representation v_c (Section 3.1); (2) generating the topic-level graph representation v_g (Section 3.2 and Section 3.3); (3) predicting the coherence score s based on v_c and v_g (Section 3.4). The training details of our GRADE is elaborated in Section 3.5.

3.1 Utterance-level Contextualized Encoding

We use BERT (Devlin et al., 2019) to encode the context c and the response r . The pooled output feature of BERT is then taken as the utterance-level contextualized representation v_c :

$$v_c = BERT(c, r). \quad (1)$$

3.2 Dialogue Graph Construction

We construct a topic-level dialogue graph based on c and r , denoted as $G = (V, E)$, where V is a set of topic nodes and E is a set of edges between topics. The details are described as follows.

Nodes. To determine the nodes in G , we first apply a rule-based keyword extractor that combines both TF-IDF and Part-Of-Speech features (Tang et al., 2019), to extract the keywords of c and r . Then the keywords in c is the context-topic nodes

of G , denoted as $V_c = \{t_1, t_2, \dots, t_p\}$, while the keywords in r is the response-topic nodes of G , denoted as $V_r = \{t_{p+1}, t_{p+2}, \dots, t_{p+q}\}$, where p and q are the numbers of keywords in the context c and the response r respectively. Therefore, $V = V_c \cup V_r$. After determining the nodes, we utilize ConceptNet to obtain node representations. Specifically, each topic node t_i is aligned to the corresponding node in ConceptNet and first initialized as $\mathbf{h}_i = CN(t_i) \in \mathbb{R}^d$, $i \in [1, p+q]$, where \mathbf{h}_i is the initial representation of the node t_i , CN means the ConceptNet Numberbatch embeddings¹, d is the dimension of each node representation. Furthermore, in order to preferably capture the topic relations in reality, \mathbf{h}_i is updated with the representations of its k -hop neighbors in ConceptNet, named as k -hop neighboring representations:

$$\mathbf{h}_{\bar{\mathcal{N}}_i^k} = \frac{1}{|\bar{\mathcal{N}}_i^k|} \sum_{t_j \in \bar{\mathcal{N}}_i^k} CN(t_j), \quad (2)$$

$$\bar{\mathbf{h}}_i = \mathbf{h}_i + \sum_{k=1}^K (\mathbf{W}_k \mathbf{h}_{\bar{\mathcal{N}}_i^k} + \mathbf{b}), \quad (3)$$

where K is the maximum number of hops taken into account and is set as 2, $\bar{\mathcal{N}}_i^k$ is the k^{th} hop neighboring nodes of t_i in the ConceptNet graph, \mathbf{W}_k and \mathbf{b} are the weight matrix and bias vector respectively.

Edges. Since our goal is to predict a coherence score of a response based on a context, we only consider the edges between the context nodes V_c and the response nodes V_r . In other words, the edges only exist between each context-topic node V_c^i and each response-topic node V_r^j . Moreover, we consider G as a weighted undirected graph and assign a weight to each edge of G by heuristically using the hop information in the ConceptNet commonsense graph, named as hop-attention weights. Specifically, let the weighted adjacency matrix of G as \mathbf{A} , then the hop-attention weight of the edge between the nodes t_i and t_j (i.e., $\mathbf{A}[i][j]$) is determined by:

$$\mathbf{A}[i][j] = \frac{1}{\#hops(V_c^i, V_r^j)}, \quad (4)$$

where $\#hops(\cdot)$ indicates the shortest path between V_c^i and V_r^j over the ConceptNet graph. As a result, the distances between topic nodes are re-defined and the nodes that are far away from each

¹<https://github.com/commonsense/conceptnet-numberbatch>

other will have low weight values. After determining the edges, we randomly deactivate a certain number of edges from G at each training step to prevent over-smoothing, and normalize the adjacency matrix \mathbf{A} (Rong et al., 2020):

$$\bar{\mathbf{A}} = (\mathbf{D} + \mathbf{I})^{-1/2} (\mathbf{A} + \mathbf{I}) (\mathbf{D} + \mathbf{I})^{-1/2}, \quad (5)$$

where $\bar{\mathbf{A}}$ is the augmented normalized adjacency matrix, \mathbf{D} is the corresponding degree matrix of \mathbf{A} and \mathbf{I} is the identity matrix.

3.3 Topic-level Graph Reasoning

We explicitly model the topic transition dynamics by reasoning over the constructed topic-level graph G via two steps: aggregation and combination (Hamilton et al., 2017).

In the first step, we apply the graph attention network (GAT) (Veličković et al., 2018) to aggregate neighboring information of each node t_i . The aggregated representation $\mathbf{z}_i^{(l)}$ at the layer l for the node t_i is formulated as follows:

$$\mathbf{z}_i^{(l)} = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_l \mathbf{h}_j^{(l)}, \quad (6)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{n \in \mathcal{N}_i} \exp(e_{in})}, \quad (7)$$

$$e_{ij} = \bar{\mathbf{A}}[i][j] * \rho \left(\mathbf{a}_l^T \left[\mathbf{W}_l \mathbf{h}_i^{(l)} \parallel \mathbf{W}_l \mathbf{h}_j^{(l)} \right] \right), \quad (8)$$

where $\mathbf{h}_i^{(0)} = \bar{\mathbf{h}}_i$, \mathcal{N}_i is the neighboring nodes of t_i in the dialogue graph G , $\mathbf{W}_l \in \mathbb{R}^{d \times d}$ and $\mathbf{a}_l \in \mathbb{R}^{2d}$ are learnable parameters, α_{ij} is the attention coefficient, ρ is LeakyReLU, and \cdot^T represents transposition. Note that we scale the attention coefficients with the above augmented normalized adjacency matrix $\bar{\mathbf{A}}$, as shown in equation 8, so that the network will pay more attention to the nodes that are closer to t_i in the ConceptNet graph during the aggregation.

In the second step, the aggregated representation $\mathbf{z}_i^{(l)}$ is combined with the i^{th} node representation $\mathbf{h}_i^{(l)}$ to get the updated node representation $\mathbf{h}_i^{(l+1)}$:

$$\mathbf{h}_i^{(l+1)} = ELU \left(\mathbf{V}_l \mathbf{h}_i^{(l)} + \mathbf{z}_i^{(l)} \right), \quad (9)$$

where $\mathbf{V}_l \in \mathbb{R}^{d \times d}$ is the weight matrix to transform $\mathbf{h}_i^{(l)}$, and ELU represents an exponential linear unit (Clevert et al., 2016).

Finally, the topic-level graph representation \mathbf{v}_g is obtained by:

$$\mathbf{v}_g = FC_0(\text{mean}(\{\mathbf{h}_i^{(L)} | i \in [1, p+q]\})), \quad (10)$$

where $\mathbf{h}_i^{(L)}$ is the i^{th} node representation at the last layer, *mean* represents mean pooling and FC_0 is a fully-connected layer with a ELU activation.

3.4 Coherence Scoring

To compute the coherence score s , the contextualized representation \mathbf{v}_c and the graph representation \mathbf{v}_g are concatenated together and fed into a multi-layer perceptron (MLP) to transform the high-dimensional representation into a real number:

$$s = FC_3(FC_2(FC_1([\mathbf{v}_c; \mathbf{v}_g])), \quad (11)$$

where FC_1 , FC_2 and FC_3 are three different fully-connected layers whose activation functions are ELU, ELU and sigmoid, respectively.

3.5 Training

Training Objective. Inspired by Tao et al. (2018), we train our GRADE in an unsupervised manner. Given a dataset $D = \{(\mathbf{c}_i, \mathbf{r}_i, \bar{\mathbf{r}}_i) | i \in [1, N]\}$, where \mathbf{c}_i and \mathbf{r}_i are a ground-truth context-response pair and $\bar{\mathbf{r}}_i$ is a false response for the context \mathbf{c}_i selected by using negative sampling described in the next paragraph, then GRADE is trained to predict a higher score for each ground-truth response \mathbf{r}_i than its corresponding false response $\bar{\mathbf{r}}_i$ by minimizing the following margin ranking loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \max(0, \bar{s}_i - s_i + m), \quad (12)$$

where N is the size of the dataset, m is a margin value set as 0.1, s_i and \bar{s}_i are the coherence scores of \mathbf{r}_i and $\bar{\mathbf{r}}_i$ respectively in the i^{th} example.

Negative Sampling. Following Sato et al. (2020), we select the false response $\bar{\mathbf{r}}$ that is similar to the ground-truth response \mathbf{r} , instead of random sampling adopted in previous works (Tao et al., 2018; Ghazarian et al., 2019). Overall, we generate negative samples by two sampling methods: lexical sampling and embedding-based sampling. For lexical sampling, we use Lucene² to retrieve utterances that are related to the ground-truth response \mathbf{r} from the training set, and select the middle one in the retrieved utterances as the false response $\bar{\mathbf{r}}$. For embedding-based sampling, we first randomly sample 1000 utterances and take the utterances with the top-5 cosine similarity against the ground-truth response \mathbf{r} .³ The false response $\bar{\mathbf{r}}$ is then randomly selected from the top-5 utterances.

²<https://lucene.apache.org>

³All the utterances are encoded with BERT.

4 Experiments

4.1 Experimental Setup

Dialogue Models. We consider both retrieval-based and generation-based dialogue models to obtain diverse responses for metric evaluation so that the performance of the metrics can be assessed comprehensively. Specifically, we first deploy Transformer-Ranker and Transformer-Generator from the ParlAI platform (Miller et al., 2017), where the former is retrieval-based and the latter is generation-based. Besides, we also deploy two state-of-the-art dialogue models, BERT-Ranker (Urbanek et al., 2019) and DialogPT (Zhang et al., 2019) that can output more human-like responses than Transformer-Ranker and Transformer-Generator.

Baseline Metrics. We compare our GRADE with seven dialogue metrics, consisting of three statistic-based metrics: BLEU (Papineni et al., 2002) ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), four learning-based metrics: ADEM (Lowe et al., 2017), BERT-RUBER (Ghazarian et al., 2019), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). Note that, for comparison, we only present the BLEU-4 results for BLEU metric, and ROUGE-L for ROUGE, BERTScore-F1 for BERTScore.

Datasets. We use the DailyDialog⁴ (Li et al., 2017) dataset which contains high-quality open-domain conversations about daily life including diverse topics, to learn our GRADE. In addition, another two chit-chat datasets, ConvAI2⁵ (Dinan et al., 2019) and EmpatheticDialogues⁶ (Rashkin et al., 2019), are considered as unseen datasets to verify the transferability of the metrics. The details of the datasets are provided in Appendix A.

Implementation Details. We use $BERT_{BASE}$ for the utterance-level contextualized encoding. For the graph reasoning module, the GAT layer is set as 3 and the number of heads is 4, where both the input and output dimensions are 300. To train GRADE, we use Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and set batch size as 16, learning rate as $2e-5$. Our GRADE is implemented with a natural language processing toolkit, Texar-Pytorch (Hu et al., 2019).

Human Judgements. We collected human judge-

⁴<http://yanran.li/dailydialog>

⁵<http://convai.io>

⁶<https://github.com/facebookresearch/EmpatheticDialogues>

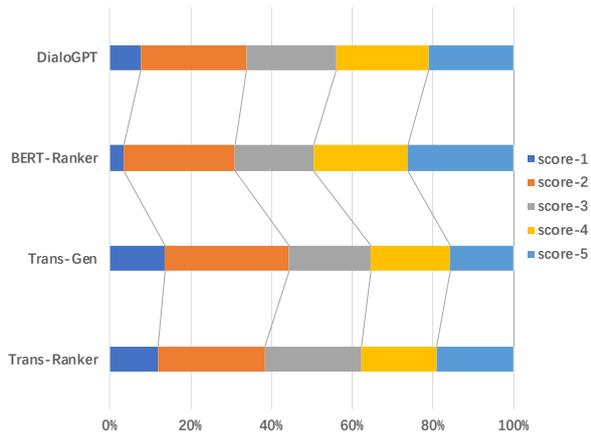


Figure 3: Score distributions of human judgements on the ConvAI2 dataset. Trans-Gen and Trans-Ranker refer to the Transformer-Generator and Transformer-Ranker dialogue models respectively.

ments from Amazon Mechanical Turk (AMT). Each survey contained six questions, including five coherence questions and one attention check question. The submissions failed in the attention check are directly discarded. For each coherence question, workers were provided with a context-response pair and asked to assess the coherence between the context and the response on a scale of 1-5 (not coherent at all to very coherent). Each pair was assessed by 8 to 10 individual workers. In total, there are 1200 different pair and 11910 human annotations from 217 unique workers, as the final human judgements. As shown in Figure 3, the distributions of human judgements are balanced from score 1 to 5. Moreover, It also demonstrates that the dialogue models we selected are diverse in performance, which helps comprehensively assess the abilities of the metrics.

4.2 Experimental Results

DailyDialog Dataset. The test set results of the DailyDialog dataset are presented in Table 1. Overall, our GRADE obtains the highest correlations with human judgements in average. Although the Spearman value of GRADE on the Transformer-Ranker is lower than BLEURT which is trained on a very large-scale dataset, the averaged correlation result of GRADE is 1% higher than BLEURT. Besides, all the correlation results of GRADE are statistically significant with p -value < 0.05 , which is more reliable than the baselines.

Other Unseen Datasets. To verify the transferability of our GRADE, we further evaluate the human correlations of GRADE compared with other baselines on two unseen chit-chat datasets, ConvAI2 and EmpatheticDialogues. Results in Table

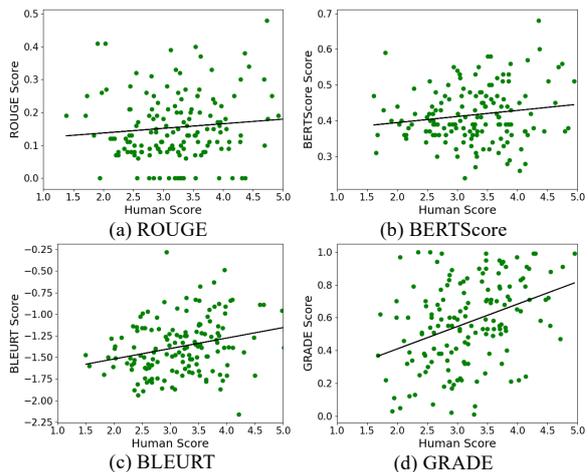


Figure 4: Score correlations between auto-metrics and human judgements, presented in a scatter plot form. Each point is associated with a context-response pair where the context is from the ConvAI2 dataset, and the response is generated by the DialogGPT model.

1 show that GRADE can easily adapt to other unseen datasets without any re-training and obtain more stable and higher correlations with human judgements than the baseline metrics. It is noteworthy that all Pearson and Spearman correlations of GRADE are statistically significant with p -value < 0.05 , and most of them are with p -value < 0.01 . Particularly, GRADE achieves a significant Pearson correlation of 0.606 and Spearman correlation of 0.617 for evaluating Transformer-Generator on the ConvAI2 dataset, bringing an improvement of 0.411 (Pearson) and 0.417 (Spearman) compared with BLEURT. Furthermore, Table 2 presents the correlation results of GRADE and other baselines for evaluating two state-of-the-art dialogue models, BERT-Ranker and DialogGPT. Our GRADE significantly outperforms the baseline metrics on human correlations, which shows that GRADE is better at evaluating the coherence of high-quality responses. Besides, Figure 4 illustrates the scatter plots against human judgements for DialogGPT on the ConvAI2 dataset. We can see that the scores predicted by GRADE are closer to the human scores than the baseline metrics, which intuitively shows the superiority of our GRADE.

4.3 Ablation Studies

We perform ablation studies⁷ for the main components of GRADE to better analyze their relative contributions. The results are shown in Table 3.

Does the negative sampling strategy work? We

⁷For each ablation experiment, We run five times and take the averaged result since the results fluctuate over different runs (more details in Section 5).

Metric		Transformer-Ranker		Transformer-Generator		Average
		Pearson	Spearman	Pearson	Spearman	
<i>DailyDialog</i>						
Statistic-based	BLEU	0.065 *	0.114 *	0.084 *	0.246	0.127
	ROUGE	0.163	0.169	0.138 *	0.126 *	0.149
	METEOR	0.079 *	0.036 *	0.115 *	0.016 *	0.062
Learning-based	BERTScore	0.163	0.138 *	0.214	0.156	0.168
	ADEM	0.162	0.179	0.077 *	0.092 *	0.128
	BERT-RUBER	0.185	0.225	0.142 *	0.182	0.184
	BLEURT	0.230	0.258	0.347	0.299	0.284
	GRADE	0.261	0.187	0.358	0.368	0.294
<i>ConvAI2</i>						
Statistic-based	BLEU	0.161	0.240	0.130 *	0.013 *	0.136
	ROUGE	0.177	0.240	0.130 *	0.126 *	0.168
	METEOR	0.215	0.274	0.101 *	0.131 *	0.180
Learning-based	BERTScore	0.310	0.344	0.266	0.241	0.290
	ADEM	-0.015 *	-0.040 *	0.063 *	0.057 *	0.016
	BERT-RUBER	0.204	0.274	0.160	0.173	0.203
	BLEURT	0.259	0.229	0.195	0.200	0.221
	GRADE	0.535	0.558	0.606	0.617	0.579
<i>EmpatheticDialogues</i>						
Statistic-based	BLEU	-0.073 *	0.081 *	-0.056 *	-0.089 *	-0.034
	ROUGE	0.170	0.143 *	-0.200	-0.202	-0.022
	METEOR	0.275	0.269	-0.126 *	-0.130 *	0.072
Learning-based	BERTScore	0.184	0.181	-0.087 *	-0.115 *	0.041
	ADEM	0.001 *	-0.004 *	0.087 *	0.086 *	0.042
	BERT-RUBER	0.021 *	-0.034 *	-0.128 *	-0.177	-0.080
	BLEURT	0.187	0.181	0.017 *	-0.031 *	0.090
	GRADE	0.375	0.338	0.257	0.223	0.298

Table 1: Correlations between automatic evaluation metrics and human judgements on three different datasets (DailyDialog, ConvAI2 and EmpatheticDialogues) and two dialogue models (Transformer-Ranker and Transformer-Generator). The star * indicates results with p-value > 0.05 , which are not statistically significant.

	Bert-Ranker		DialoGPT	
	Pearson	Spearman	Pearson	Spearman
ROUGE	0.157	0.121 *	0.084 *	0.098 *
METEOR	0.070 *	0.088 *	0.020 *	0.029 *
BERTScore	0.165	0.135 *	0.208	0.177
BERT-RUBER	0.141 *	0.111 *	0.113 *	0.085 *
BLEURT	0.133 *	0.071 *	0.273	0.275
GRADE	0.502	0.425	0.487	0.485

Table 2: Correlations between auto-metrics and human judgements on the ConvAI2 dataset and two dialogue models, Bert-Ranker and DialoGPT, respectively.

first verify the effectiveness of our negative sampling strategy by replacing it with random sampling. As shown in Table 3, adopting the random sampling strategy hurts performance significantly with a 6.6% drop in average, which indicates the importance of our negative sampling strategy.

Does the graph work? To prove the contribution of our graph components, we perform three ablations respectively: 1) remove the entire graph branch of GRADE; 2) remove the k-hop neighboring representations used for initializing the node representations in the dialogue graph; 3) remove

the hop-attention weights used for computing a weight for each edge in the dialogue graph. Consequently, the performance of GRADE decreased after removing the graph branch or one of the components in the graph branch.

How much graph information we need? Finally, we explore the number of k-hop neighboring representations needed for initializing the dialogue graph’s nodes in two aspects: the maximum number of hops (refer to the K in Equation 3), and the number of neighboring nodes in the k^{th} hop (denoted as N_k , i.e., the number of nodes in \bar{N}_i^k in Equation 3). By comparing the results among the first row and the last three rows in Table 3, we confirm that incorporating both the 1st hop and the 2nd hop neighboring nodes brings the best performance. Furthermore, we also observe that considering too much graph information may result in relatively poor performance, as shown in the last row. Therefore, the final version of GRADE adopts the 2-hop neighboring representations where $N_1 = 10, N_2 = 10$.

Metric	Transformer-Ranker		Transformer-Generator		Average
	Pearson	Spearman	Pearson	Spearman	
Our GRADE ($N_1 = 10, N_2 = 10$)	0.227 ± 0.018	0.162 ± 0.015	0.364 ± 0.017	0.372 ± 0.018	0.281 ± 0.008
random sampling	0.225 ± 0.022	0.153 * ± 0.016	0.237 ± 0.034	0.245 ± 0.028	0.215 ± 0.023
no graph branch	0.211 ± 0.028	0.146 * ± 0.020	0.324 ± 0.034	0.336 ± 0.029	0.254 ± 0.024
no k-hop neighboring representations	0.219 ± 0.011	0.153 * ± 0.008	0.347 ± 0.032	0.356 ± 0.034	0.269 ± 0.019
no hop-attention weights	0.227 ± 0.013	0.162 ± 0.012	0.349 ± 0.019	0.352 ± 0.015	0.273 ± 0.007
1-hop neighboring representations ($N_1 = 10$)	0.211 ± 0.022	0.150 * ± 0.019	0.347 ± 0.014	0.352 ± 0.017	0.265 ± 0.018
1-hop neighboring representations ($N_1 = 20$)	0.206 ± 0.025	0.148 * ± 0.015	0.356 ± 0.030	0.358 ± 0.032	0.267 ± 0.025
2-hop neighboring representations ($N_1 = 20, N_2 = 20$)	0.216 ± 0.016	0.150 * ± 0.014	0.360 ± 0.019	0.364 ± 0.017	0.273 ± 0.015

Table 3: Ablation results on the DailyDialog dataset, averaged across five random seeds, with standard deviations presented in gray color. N_1 and N_2 refer to the numbers of the 1st and 2nd hop neighboring nodes in ConceptNet, respectively. The symbol * indicates that three or more than three correlation results over the five random seeds are not statistically significant, namely, p-value > 0.05.

Context	Response	Graph	Coherence Score								
<p>U1: I enjoy a great meal, but usually just eat when there is nothing else to do. haha!</p> <p>U2: yeah that is cool, what is your favorite color ?</p>	<p>R_{model}: I like green and blue.</p> <p>R_{ref}: Green , and it shows with my bright green crew cut ! what is yours ?</p>		<table border="1"> <tr><td>Human</td><td>4.10</td></tr> <tr><td>GRADE</td><td><u>4.92</u></td></tr> <tr><td>BERT-RUBER</td><td>2.84</td></tr> <tr><td>ROUGE</td><td>1.70</td></tr> </table>	Human	4.10	GRADE	<u>4.92</u>	BERT-RUBER	2.84	ROUGE	1.70
Human	4.10										
GRADE	<u>4.92</u>										
BERT-RUBER	2.84										
ROUGE	1.70										
<p>U1: I see, you must have felt hopeful.</p> <p>U2: I gave thanks in spite of the difficulties.</p>	<p>R_{model}: That's what you should do.</p> <p>R_{ref}: Good, one should always be positive.</p>		<table border="1"> <tr><td>Human</td><td>4.10</td></tr> <tr><td>GRADE</td><td>2.56</td></tr> <tr><td>BERT-RUBER</td><td><u>3.77</u></td></tr> <tr><td>ROUGE</td><td>1.72</td></tr> </table>	Human	4.10	GRADE	2.56	BERT-RUBER	<u>3.77</u>	ROUGE	1.72
Human	4.10										
GRADE	2.56										
BERT-RUBER	<u>3.77</u>										
ROUGE	1.72										
<p>U1: I have a big exam coming up! My nerves are getting to me.</p> <p>U2: Ohh good luck!</p>	<p>R_{model}: Good luck! Hope you do well.</p> <p>R_{ref}: Thanks! It is my final exam for a summer class studying Database Design. I actually should be taking it right now.</p>		<table border="1"> <tr><td>Human</td><td>2.89</td></tr> <tr><td>GRADE</td><td>4.28</td></tr> <tr><td>BERT-RUBER</td><td>5.00</td></tr> <tr><td>ROUGE</td><td>1.00</td></tr> </table>	Human	2.89	GRADE	4.28	BERT-RUBER	5.00	ROUGE	1.00
Human	2.89										
GRADE	4.28										
BERT-RUBER	5.00										
ROUGE	1.00										

Figure 5: Visualization results of our GRADE, compared with two baseline metrics, ROUGE and BERT-RUBER. Keywords of the contexts and the model responses R_{model} are highlighted in green and red respectively. R_{ref} is the reference response. For comparison, the auto-metric scores are normalized to the range of human scores, i.e., [1,5].

4.4 Case Study

To more intuitively analyze the performance of our GRADE, three representative examples are shown in Figure 5. From the example in the first row, we can see that the score given by our metric is closer to the human score than the other two baseline metrics. However, in the second-row example, our metric performs poorly. The potential reason may be the lack of topics (i.e., keywords) in the model response, as illustrated in the graph that only contains context-topic nodes. As a result, the graph reasoning module in our GRADE fails to induce an appropriate graph representation, which harms the coherence scoring. Finally, the example in the last row shows a hard case that both our GRADE and the baseline metrics are failed to cope with. In this hard case, the topics of the model response

are relevant to the dialogue context so that both our GRADE and BERT-RUBER, as learning-based metrics, deem that the response greatly matches the context. However, the truth is that the model response is more likely a response for the previous utterance U1 rather than U2, which is hard for metrics to recognize.

5 Conclusion and Discussion

In this paper, we proposed GRADE (Graph-enhanced Representations for Automatic Dialogue Evaluation), a novel metric for dialogue coherence evaluation of open-domain dialogue systems. Empirical results show that GRADE has stronger correlations with human judgements and can generalize to other unseen chat-chat datasets. Besides, we also release a new large-scale human evaluation bench-

mark to facilitate future research on automatic metrics.

A limitation of GRADE is the inconsistency between the training objective (relative ranking) and the expected behavior (absolute scoring). Specifically, the ranking loss we adopted only requires good responses to be ranked higher than bad responses, which is a relatively loose constraint compared with the absolute scoring that humans do. Therefore, GRADE may deviate from the human scoring criterion and fail to quantify the dialogue responses accurately, and that the human correlation results fluctuate over different runs. Overall, to develop a dialogue metric that can quantify in a more human-like manner, it is critical to reducing the gap between the training objective and the model behavior we truly care about.

Acknowledgments

We thank all anonymous reviewers for their constructive comments. This work was supported in part by National Key RD Program of China under Grant No. 2018AAA0100300, National Natural Science Foundation of China (NSFC) under Grant No.U19A2073 and No.61976233, Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) Grant No.2019B1515120039, Nature Science Foundation of Shenzhen Under Grant No. 2019191361, Zhijiang Lab’s Open Fund (No. 2020AA3AB14).

References

- Daniel De Freitas Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.
- Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhunoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *ArXiv*, abs/1902.00098.
- Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 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*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034.
- Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, et al. 2019. Texar: A modularized, versatile, and extensible toolkit for text generation. In *ACL 2019, System Demonstrations*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. **DailyDialog: A manually labelled multi-turn dialogue dataset**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. **How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. **Towards an automatic Turing test: Learning to evaluate dialogue responses**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

- A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bor-des, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. **Towards empathetic open-domain conversation models: A new benchmark and dataset**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Young Ju, Mary F. Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *ArXiv*, abs/2004.13637.
- Yu Rong, Wen bing Huang, Tingyang Xu, and Junzhou Huang. 2020. Droppedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*.
- Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. *ArXiv*, abs/2004.14302.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL-HLT*.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *ArXiv*, abs/2004.04696.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. **Target-guided open-domain conversation**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. **Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems**. In *AAAI*.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. **Learning to speak and act in a fantasy text adventure game**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. **Graph attention networks**. In *International Conference on Learning Representations*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. **Dialogpt: Large-scale generative pre-training for conversational response generation**.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. **Learning discourse-level diversity for neural dialog models using conditional variational autoencoders**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.
- Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. **The design and implementation of xiaoice, an empathetic social chatbot**. *Computational Linguistics*, pages 1–62.

A Details of the Datasets

The detailed processing procedure of DailyDialog and the introduction of the other two unseen datasets are presented.

DailyDialog is a chit-chat dataset with strong annotations for topic, emotion and utterance act. It contains total 13,118 open-domain multi-turn dialogues. We use the initial split of DailyDialog where training/validation/test sets have 11,118/1,000/1,000 dialogues respectively. Next, we subdivide these dialogues into context-response pairs each of which is composed of a context c with length = 2 and a ground-truth response r . Therefore, the processed training/validation/test sets now have 59264/6015/5705 pairs respectively. Then, for each context-response pair, we obtain two false responses \bar{r}_l and \bar{r}_e based on the lexical sampling

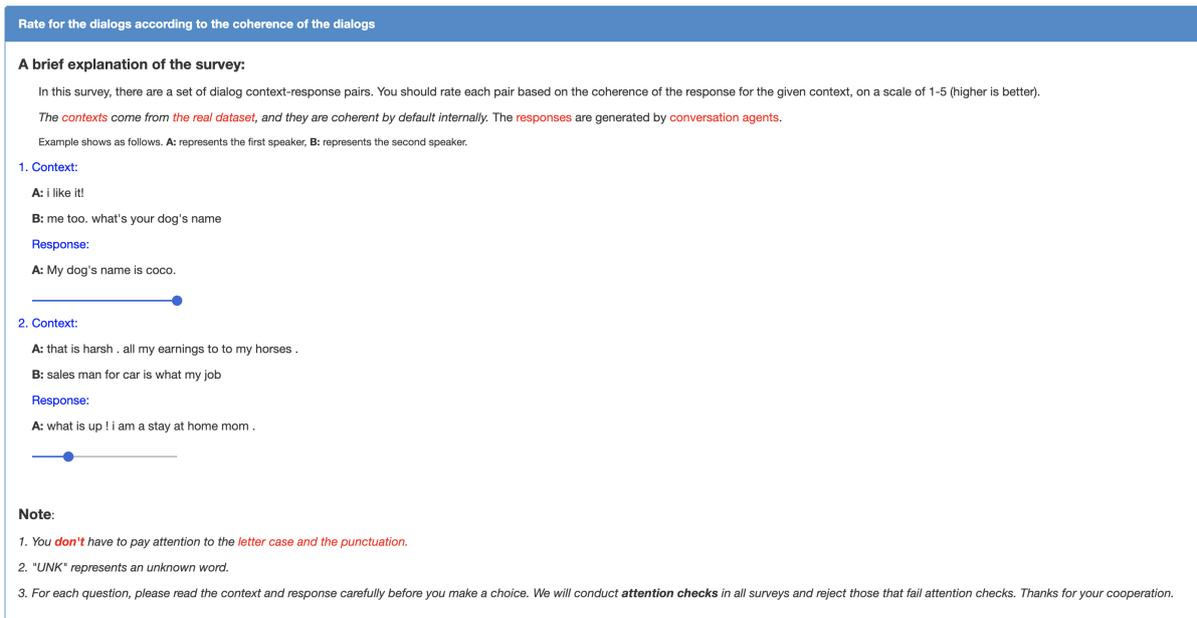


Figure 6: Screenshot of the survey’s introduction on AMT for collecting the human judgements.

and embedding-based sampling methods respectively, and get two tuples (c, r, \bar{r}_l) , (c, r, \bar{r}_e) . In total, there are 118528/12030/11410 tuples as our final data for training GRADE.

ConvAI2 is a chit-chat dataset based on the PersonaChat dataset (Dinan et al., 2019) for a NIPS 2018 competition. The dataset was collected by asking workers to chat with each other naturally with a given persona. The conversations cover a broad range of topics and frequently change during the conversations since both the speakers want to say out their persona information.

EmpatheticDialogues is a novel dataset of 25k conversations grounded in a wide range of emotions to facilitate training and evaluating dialogue systems. It has been verified that dialogue models trained on this dataset are perceived to be more empathetic by human evaluators.

B Screenshot of the Survey’s Introduction on AMT

See Figure 6.