DynaEval: Unifying Turn and Dialogue Level Evaluation

Chen Zhang[†,⋆]   **Yiming Chen**[†]   **Luis Fernando D'Haro**[‡]   **Yan Zhang**[†]
Thomas Friedrichs[⋆]   **Grandee Lee**[†]   **Haizhou Li**[†,⋆⋆]

[†]National University of Singapore   [⋆]Robert Bosch (SEA), Singapore
[‡]Universidad Politécnica de Madrid, Spain   [⋆⋆]Kriston AI Lab, China
{chen_zhang,yiming.chen,grandee.lee}@u.nus.edu,
{haizhou.li,eleyanz}@nus.edu.sg,
luisfernando.dharo@upm.es, thomas.friedrichs@sg.bosch.com

Abstract

A dialogue is essentially a multi-turn interaction among interlocutors. Effective evaluation metrics should reflect the dynamics of such interaction. Existing automatic metrics are focused very much on the turn-level quality, while ignoring such dynamics. To this end, we propose DynaEval[1], a unified automatic evaluation framework which is not only capable of performing turn-level evaluation, but also holistically considers the quality of the entire dialogue. In DynaEval, the graph convolutional network (GCN) is adopted to model a dialogue in totality, where the graph nodes denote each individual utterance and the edges represent the dependency between pairs of utterances. A contrastive loss is then applied to distinguish well-formed dialogues from carefully constructed negative samples. Experiments show that DynaEval significantly outperforms the state-of-the-art dialogue coherence model, and correlates strongly with human judgements across multiple dialogue evaluation aspects at both turn and dialogue level.

1 Introduction

Modern dialogue systems (Smith et al., 2020; Zhang et al., 2020; Adiwardana et al., 2020) leveraging large-scale language model pre-training (Devlin et al., 2019; Radford et al., 2019) are capable of generating fluent and contextually relevant utterances. Yet, they still face difficulties in mimicking human conversations in the sense that they lack certain conversation-level attributes, such as coherence (Cervone et al., 2018), consistency (Welleck et al., 2019; Nie et al., 2020), diversity (Li et al., 2016; Wu et al., 2020) and engagement (Ghandeharioun et al., 2019; Ghazarian et al., 2020). One of the main reasons is the dearth of effective dialogue-level evaluation mechanisms to guide the studies and to monitor progress.

Commonly used static metrics, such as BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and ROUGE (Lin, 2004), correlate poorly with human judgements (Liu et al., 2016) rendering them unsuitable for dialogue evaluation. While some recent automatic dialogue evaluation metrics (Ghazarian et al., 2019; Mehri and Eskenazi, 2020b; Huang et al., 2020; Zhang et al., 2021b) demonstrate strong correlations with human judgement at the turn-level, they only focus on context-response pairs without explicitly modeling the interaction over an entire dialogue. To perform dialogue-level evaluation, we need to rely on the aggregation of turn-level scores over the dialogue as a proxy for a dialogue-level score.

Furthermore, a recent study by Mehri and Eskenazi (2020a) found out that even though state-of-the-art chatbots outperform humans across multiple turn-level evaluation criteria, such as interestingness, engagement and specificity, their dialogue-level ratings like coherence, Likability and diversity are still far below human level. This further reinforces the idea that turn-level quality evaluation may be insufficient to assess the performance of open-domain dialogue systems.

In this work, we address the problem of automatic open-domain dialogue evaluation by focusing on the quality of an entire dialogue. This is a departure from the way we frame the problem as a weakly supervised next sentence prediction (Mehri and Eskenazi, 2020b; Sato et al., 2020) or language modeling tasks (Nedelchev et al., 2020; Pang et al., 2020) for context-response pairs. To this end, we need to answer two important questions: (1) How to effectively represent the entire dialogue? (2) How to incorporate this dialogue-level knowledge into our evaluation framework? We propose DynaEval to provide meaningful dialogue-level representation with explicit modeling of the interactive

---

[1]https://github.com/e0397123/DynaEval

5676

dynamics among interlocutors, for a unified turn and dialogue level quality assessment.

The main contributions of this work include: (1) The unified turn and dialogue level evaluation represents a departure from turn-level evaluation scheme; (2) DynaEval is one of the first few metrics where dialogue level dynamics is considered with structured graph representation. (3) Empirical results show that DynaEval outperforms the state-of-the-art dialogue coherence model and strongly correlates with human judgements at both turn and dialogue level.

2 Related Work

2.1 Open-ended Dialogue Evaluation

Turn-Level Evaluation The current trend for automatic dialogue evaluation is shifting towards the reference-free paradigm. Lately, the research community has witnessed a surge in the automatic metrics along these lines. Many of them focus on evaluating naturalness of generated responses. Typical examples include perplexity (Adiwardana et al., 2020), USR-MLM (Mehri and Eskenazi, 2020b) and GPT-2 (Radford et al., 2019) based fluency metrics (Nedelchev et al., 2020; Pang et al., 2020).

Another group of metrics evaluates contextual relevance of the responses. For example, RU-BER (Tao et al., 2018), BERT-RUBER(Ghazarian et al., 2019) and USR-DR (Mehri and Eskenazi, 2020b) predict the relatedness between generated responses w.r.t the corresponding context by training a discriminative network to distinguish the original response from negative samples bootstrapped from the training set. Sato et al. (2020) and Lan et al. (2020) provide a better sampling strategy for bootstrapping negative samples.

Besides these two major aspects, there are many metrics for other qualities, such as adequacy (D'Haro et al., 2019; Zhang et al., 2021a), consistency (Welleck et al., 2019; Dziri et al., 2019), engagement (Ghazarian et al., 2020).

Even though all these automatic metrics demonstrate strong correlation with human judgements, they are laser-focused on one aspect of the evaluation. In addition, they do not explicitly model the speaker-level and utterance-level interactions, which we believe is essential for the dialogue-level representation, and eventually benefits the dialogue evaluation task.

Interactive Evaluation A popular human evaluation method is the interactive evaluation whereby

human judges converse with dialogue systems and make the assessment at the end of the conversations (See et al., 2019; Finch and Choi, 2020; Li et al., 2019; Deriu et al., 2020). It has been shown to be more reliable than turn-level static evaluation (Mehri and Eskenazi, 2020a).

There are few studies on fully automating this process. Ghandeharioun et al. (2019) propose a self-play scenario where the dialog system chats with itself and a combination of three metrics measuring sentiment, semantic coherence and engagement respectively along the conversation trajectory is computed to approximate dialogue-level quality estimation. Mehri and Eskenazi (2020a) propose the FED metric, which evaluates the quality of a system utterance in an interactive setting by computing the likelihood of a particular follow-up utterance responded by dialoGPT (Zhang et al., 2020). Moreover, Sinha et al. (2020) come up with MaUde, a reference-free metric tailored for online dialogue evaluation, which leverages a pre-trained DistilBERT (Sanh et al., 2019) model to extract the semantic representation of dialogue turns and uses bidirectional LSTM to explicitly model the discourse structure.

While the interactive evaluation is more reliable than the turn-level static evaluation, it still relies on the aggregation of turn-level scores. An ideal approximation of the human evaluation process is a top-down approach whereby we examine the quality of the entire dialogue at macro level before zooming into the dialogue turns. Hence, a unified framework, which holistically models the entire dialogue, is highly sought after.

2.2 Dialogue Coherence

Examining a dialogue at macro level is related to discourse coherence (Halliday and Hasan, 2014; Grosz et al., 1995; Barzilay and Lapata, 2008), which considers whether a piece of text is in a consistent and logical manner, as opposed to a random collection of sentences. Dialogue is a special kind of discourse structure, of which coherence assessment is an essential part of quality evaluation.

Many studies have followed the standard discourse coherence evaluation protocol (Cervone and Riccardi, 2020; Zhou et al., 2019; Mesgar et al., 2020). Very few have considered customizing their dialogue coherence models for evaluating the performance of dialogue systems. It is common to leverage supervised approaches (Higashinaka et al.,

2014; Gandhe and Traum, 2016; Cervone et al., 2018; Yi et al., 2019), that is closely linked to modeling with entities and dialogue acts (Cervone and Riccardi, 2020; Zhou et al., 2019; Mesgar et al., 2020).

Hence, we are motivated to study the application of dialogue coherence modeling for automatic dialogue evaluation by designing a self-supervised framework, without dependence on any human annotations for coherence features.

2.3 Graph Modeling of Dialogue

Recently, the graph neural network (GNN) (Scarselli et al., 2008; Kipf and Welling, 2017; Schlichtkrull et al., 2018) has been successfully applied in various dialogue applications. For example, Ghosal et al. (2019) adopts GCN for utterance-level emotion recognition. Chen et al. (2018) modeled structured dialogue policy with GNN and (Qin et al., 2020) proposes a joint framework leveraging graph attention network (Veličković et al., 2018) for both dialogue act recognition and sentiment classification.

GNN is useful for dialogue modeling, because the relative position of target and context utterances decides how past utterances influence future utterances and vice versa (Ghosal et al., 2019). The interaction of utterances can be effectively captured with a graph structure as long as they are connected by relation-aware edges. However, GNN has not been well studied for dialogue evaluation. Huang et al. (2020) recently proposes the GRADE metric, leveraging graph modeling for turn-level coherence evaluation. The way we use GNN is different from Huang et al. (2020) because GRADE is focused on turn-level coherence evaluation while we are interested in a turn-dialogue joint evaluation. Furthermore, GRADE considers the keywords in context-response pairs, and we explicitly use graph structure to model the speaker and utterance level interaction within a dialogue.

3 DynaEval Framework

DyanEval represents an integration of several ideas. It takes advantage of the structured graph representation of dialogues, useful information on the utterance and speaker level interaction. It is motivated by dialogue coherence modeling.

In this paper, we only consider dyadic dialogues, but the formulation can be easily generalized to multi-party conversations. Formally, let A and B denote the two speakers participating in the dialogue. A dialogue, D, consists of a sequence of n utterances, $[u_1^A, u_2^B, \ldots, u_{n-1}^A, u_n^B]$[2]. Let $\bar{D}$ represent the negative dialogue sample obtained via various sampling strategies described in Section 3.5.

Figure 1 illustrates the learning process of DynaEval in four steps[3]: (1) Deriving contextualized representation, $\mathbf{e}_i$, for utterances within D. (Section 3.1). (2) Constructing the directed dialogue graph. The nodes are initialized with $\mathbf{e}_i$ and the edges between node pairs represent the speaker and temporal dependencies (Section 3.2). (3) Generating utterance-level graph representation, $\mathbf{h}_i$, via feature transformation to aggregate useful contextual information from all connected neighbours to the current node (Section 3.3). (4) producing a dialogue-level score, which indicates whether D is preferred over $\bar{D}$ (Section 3.4).

3.1 Dialogue Utterance Representation

A sentence-encoder is needed to map the individual utterances within D onto the vector space. Firstly, we fine-tune a RoBERTa-base pre-trained language model (Liu et al., 2019) with training data of the target dialogue domain, because task-adaptive fine-tuning of the pre-trained language model on the target domain data benefits the final performance (Gururangan et al., 2020; Lee and Li, 2020). Next, the mean pooling operation is performed on the token embeddings within each utterance of D to derive their respective utterance-level representations. Formally, let SRoBERTa denotes the sentence encoder and u_i^* in D is mapped into vector representations, $\mathbf{u}_i \in \mathcal{R}^d$, whereby

$$\mathbf{u}_i = \text{SRoBERTa}(u_i^*) \qquad (1)$$

Note that $*$ can be either speaker A or speaker B. Then, to capture a more fine-grained temporal dependency among the utterances, a bidirectional LSTM is adopted to model the sequential flow of information within D. The context-aware utterance representation, $\mathbf{e}_i$ is then obtained via:

$$\mathbf{e}_i = \overleftrightarrow{\text{LSTM}}(\mathbf{e}_{i(+,-)1}, \mathbf{u}_i) \qquad (2)$$

3.2 Dialogue Graph Construction

D is represented with a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is the sets of graph nodes and $\mathcal{E}$ is the set of

---

[2]n is assumed to be even to simplify the mathematical expressions.

[3]Note that all the operations from Section 3.1 through Section 3.4 are illustrated with D. They are applied in the same way on $\bar{D}$.
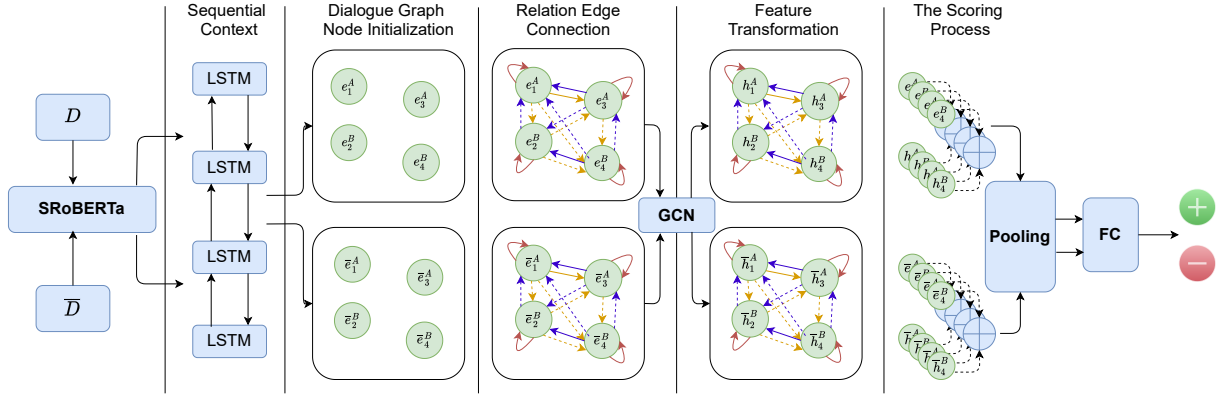
Figure 1: The architecture of DynaEval. The input is a pair of contrasting dialogues, D and $\bar{D}$. The output is a unified score indicating whether D is preferred than $\bar{D}$. Utterance-level representation derived from SRoBERTa model is used for dialogue graph node initialization. Different types of arrows in relation edge connection represent different types of relations: (1) Solid line denotes intra-speaker dependency. (2) Dotted line denotes inter-speaker dependency. (3) Red color means self-connection. (4) Purple color means connection from future utterances to previous utterances. (5) Yellow color means connection from previous utterances to future utterances. Since there are two speakers, A and B. Hence, there will be a total of $2 \times 2 \times 2 + 1 = 9$ distinct relation types.

edges, which reflects the contextual dependencies among utterance pairs.

Graph Nodes Each graph node corresponds to an utterance within D. Hence, for a dialogue with n utterances, $\mathcal{V} = \{v_1, v_2, \ldots, v_{n-1}, v_n\}$. All the graph nodes are initialized with utterance-level contextualized embeddings: $v_i = \mathbf{e}_i$.

Edges For short conversations, $\mathcal{G}$ will be a fully-connected graph whereby all graph nodes are connected to each other, including self-connection. The intuition is that short conversations tend to focus on a single topic and thus, each utterance is contextually dependent on all the other utterances in the dialogue. For long conversations, there may be frequent topic shifts. Distant utterances within the same dialogue may not be contextually relevant to the current utterance. Sometimes, adding more context leads to diminishing performance gain or even negative impact (Zhong et al., 2019). Therefore, a context window length, M, is set, which means that v_i is only connected to $v_j \in \{v_{i-M}, v_{i-M+1}, \ldots, v_i, v_{i+1}, \ldots, v_{i+M}\}^4$. Let $v_{ij} \in \mathcal{E}$ denote the edge from v_j to v_i. Each edge is associated with an edge weight, a_{ij}, and a relation type, θ_{ij}. They are illustrated as follows:

Edge Weights The edge weight determines the relative importance of the neighbour nodes w.r.t the current node. A similarity based attention module is applied to determine the edge weights. For a graph node, v_i, the set of weights, a_i, w.r.t all its incoming edges, should sum up to 1. The attention weight is formulated in the following way:

$$a_i = \text{softmax}(\mathbf{e}_i^T W_e[\mathbf{e}_{i-M}, \ldots, \mathbf{e}_{i+M}]),$$

$$\text{where} \sum_{j=i-M}^{i+M} a_{ij} = 1, W_e \in \mathcal{R}^{d \times d} \tag{3}$$

More importance is placed upon neighbouring utterances on the same topic. Little attention is paid to the irrelevant utterances.

Edge Relations Following (Ghosal et al., 2019), there are two aspects to take into account when defining the relation types. One aspect is to capture speaker dependencies. This is because we want to model the interaction between the interlocutors in a dialogue. The other aspect is to consider the temporal dependencies. This pertains to the relative position of an utterance w.r.t another. The explicit modeling of such dependency is important since the ordering of utterances within a dialogue is an essential feature for learning dialogue coherence. With these considerations, the total number of distinct types of relations[5] will be 2 (u_i^* occurs before or after u_j^*) $\times$ 2 (either u_i^A or u_i^B) $\times$ 2 (either u_j^A or u_j^B) plus the self-connection ($i = j$). This is depicted with different arrows connecting the graph nodes in Figure 1. We define this set of 9 relation types as Θ and $\theta_{ij} \in \Theta$.

---

[4]For simplicity purpose, we do not explicitly include the cases when $i <= M$ or $i + M$ is greater than the total number of utterances in a dialogue in the formula.

[5]Since we are considering dyadic dialogues, there are only two speakers involved. The formulation can be generalized to multi-party dialogue.

3.3 Feature Transformation

This section describes the process of transforming the initial node representation, $\mathbf{e}_i$, into both a speaker and context aware vector representation, $\mathbf{h}_i$, which captures the dynamics of interaction w.r.t u_i^*. Basically, the whole process is a two-stage graph convolution.

The first stage aggregates information from neighbourhood nodes to the current node v_i based on the relation-aware transformation motivated by (Schlichtkrull et al., 2018) whereby edges of different relation types are associated with different transformation matrix, W_θ':

$$\mathbf{h}_i' = \sigma(\sum_{\theta \in \Theta} \sum_{j \in S_i^\theta} \frac{a_{ij}}{c_{i,\theta}} W_\theta' \mathbf{e}_j + a_{ii} W_0' \mathbf{e}_i)$$
$$\text{for } i = 1, 2, \ldots, n \tag{4}$$

In Equation 4, $\mathbf{h}_i'$ is the intermediate node representation and σ denotes the activation function, such as ReLU. S_i^θ represents the set of indices of nodes connected to v_i with their edges v_{ij} having the relation type $\theta \in \Theta$. a_{ij} and a_{ii} are the edge weights of v_{ij} and v_{ii} respectively. $W_\theta' \in \mathcal{R}^{d' \times d}$ and $W_0' \in \mathcal{R}^{d' \times d}$ are learnable parameters of the feature transformation. $c_{i,\theta}$ is a problem specific normalization constant, which can be set as a learnable parameter or fixed in advance.

The second stage applies another graph convolution operation on the intermediate node representation, $\mathbf{h}_i'$ and the final node representation, $\mathbf{h}_i$ is obtained via:

$$\mathbf{h}_i = \sigma(\sum_{j \in S_i^\theta} W'' \mathbf{h}_j' + W_0'' \mathbf{h}_i')$$
$$\text{for } i = 1, 2, \ldots, n \tag{5}$$

where $W'' \in \mathcal{R}^{d'' \times d'}$ and $W_0'' \in \mathcal{R}^{d'' \times d'}$ are two learnable parameters in the second stage of feature transformation.

Through Equation 4 and Equation 5, relevant contextual information from neighbouring nodes is effectively accumulated to the current node while irrelevant information is filtered out.

3.4 The Scoring Process

In the scoring step, $\mathbf{h}_i$ is first concatenated with $\mathbf{e}_i$ to obtain the final utterance representation, $\mathbf{g}_i$. Next, a mean pooling layer is applied on all the utterance representations in a conversation to derive the dialogue-level representation, $\mathbf{o}$:

$$\mathbf{o} = \frac{\sum_{i=1}^n \mathbf{g}_i}{|\sum_{j=1}^n \mathbf{g}_j|} \tag{6}$$

$\bar{\mathbf{o}}$, which corresponds to $\bar{D}$, is obtained in the same way. A unified score, s_{dial} or $s_{d\bar{i}al}$, is derived by passing $\mathbf{o}$ or $\bar{\mathbf{o}}$ through a fully-connected layer.

3.5 Training Setup

Learning Objective Inspired by the preference learning approaches, the label, y for the D and $\bar{D}$ pair is defined as:

$$y = \begin{cases} 1 & \text{if } D \text{ is preferred over } \bar{D} \\ -1 & \text{if } \bar{D} \text{ is preferred over } D \end{cases} \tag{7}$$

The margin ranking loss function is adopted to train DynaEval.

$$\mathcal{L} = \max(0, -y * (s_{dial} - s_{d\bar{i}al}) + 1) \tag{8}$$

Sampling Strategy Two negative sampling strategies are explored in this paper to construct $\bar{D}$: Utterance Replacement (UR) and Speaker Level Utterance Shuffling (SS).

Utterance Replacement (UR) An utterance randomly selected from a dialogue is replaced with another utterance randomly chosen from a different dialogue. This sampling strategy perturbs a dialogue at the semantic level. An utterance from a different dialogue is considered topically in-congruent w.r.t the current dialogue context. It breaks down the current dialogue by suddenly injecting irrelevant information.

Speaker Level Utterance Shuffling (SS) With this strategy, the order of utterances from one speaker in a dialogue is kept the same while that from another speaker is shuffled. SS changes the coherence structure of a dialogue w.r.t specific speaker. This strategy is motivated by (Healey et al., 2014), which adopts a "Chance Other" method to measure how much syntactic and lexical repetition of a speaker happen by chance. The reason why we do not randomly permute the order of all utterances in the dialogue is because random permutation of all utterances is a very simple discrimination task.

4 Experiments

In this work, we consider two experiment settings to assess the effectiveness of DynaEval. The first setting (Section 4.2) is similar to the studies on

dialogue coherence (Cervone et al., 2018; Mesgar et al., 2020) where accuracy score is applied to evaluate its discrimination capability in distinguishing original dialogues from negative samples. The second setting (Section 4.3) is to evaluate its dialogue-level and turn-level judgement capability via correlation analysis on the human-chatbot conversational datasets. The domain of the evaluation set is different from that of human-human conversation datasets that DyanEval is trained on.

4.1 Dialogue Datasets

Three bench-marking open-domain dialogue datasets are included in our experiments, Empathetic Dialogue (Rashkin et al., 2019), ConvAI2 PERSONACHAT (Zhang et al., 2018b; Dinan et al., 2020) and DialyDialog (Li et al., 2017). For training, we remove dialogues containing less than 4 utterances or more than 30 utterances. Statistics of the three human-human dialogue corpora after filtering is presented in Table 1.

Empathetic Dialogue is designed for mimicking the real-life human conversation scenario whereby the interlocutors need to recognize and acknowledge the others' feelings in the conversation. This dataset pertains to the short conversation scenario where interlocutors stick to a single topic.

ConvAI2 PERSONACHAT is a crowd-sourced dataset where each pair of interlocutors try to get to know each other by conditioning their conversations on their respective persona profile provided in prior. The dataset contains more number of turns per dialogue as compared to Empathetic Dialogue. Hence, topic shift is more likely to occur within a dialogue and this simulates the long conversation scenario mentioned in Section 3.2.

DailyDialog is a high-quality human-human conversation dataset, which reflects our day-to-day communications and covers different topics about our daily life, such as relationship and health. The average dialogue length of DailyDialog lies in the middle of that of Empathetic Dialogue and ConvAI2. Topic shift in the conversations of DailyDialog occurs less frequently as compared to those in ConvAI2.

4.2 The Dialogue-level Discrimination Task

Similar to the previous works (Cervone and Riccardi, 2020; Mesgar et al., 2020), 20 perturbations are created for each dialogue w.r.t both UR and SS. For each perturbation, two pairs are formed, $\{D, \bar{D}\}$ with label $y = 1$ and $\{\bar{D}, D\}$ with label

Empathetic Dialogue	training	validation	test
#dialog	19,531	2,768	2,547
#turn	84,160	12,075	10,973
#word	1,306,060	201,816	194,772
#avg turn per dialogue	4.31	4.36	4.31
#avg words per dialogue	66.87	72.91	76.47
ConvAI2	**training**	**validation**	**test**
#dialog	17,878	1,000	-
#utterance	262,626	15,566	-
#word	3,068,672	189,374	-
#avg turn per dialogue	14.69	15.57	-
#avg words per dialogue	171.64	189.37	-
DailyDialog	**training**	**validation**	**test**
#dialog	10,245	933	918
#utterance	84,916	7,908	7,536
#word	1,189,527	109,172	106,627
#avg turn per dialogue	8.29	8.48	8.21
#avg words per dialogue	116.11	117.01	116.15

Table 1: Human-Human Dialogue Corpora Statistics

$y = -1$. Then, we train, fine-tune, and evaluate DynaEval on the training, validation, and test sets for each sampling strategy. Note that all these sets are constructed with the same perturbation method.

Baselines we compare DynaEval against three baselines: RANDOM, CoSim (Xu et al., 2018) and S-DiCoh (Mesgar et al., 2020). RANDOM baseline arbitrarily assigns a label to the input dialogue pairs. It suggests the peformance lower bound. CoSim is a common method for dialogue coherence assessment (Xu et al., 2018; Zhang et al., 2018a). It obtains a dialogue-level score by averaging the cosine similarities between sentence embeddings of all adjacent utterance pairs within the dialogue. For fair comparison, we apply the same procedure described in Section 3.1 to derive the sentence embedding of an utterance in CoSim. S-DiCoh (Mesgar et al., 2020) is a recent state-of-the-art dialogue coherence model. It models a dialogue with a neural network framework consisting of two bidirectional LSTM layers with attention mechanism at both the token and utterance level.

Results and Analysis It can be observed in Table 2 that on all bench-marking dialogue datasets, DynaEval outperforms the baselines in both UR and SS category. Even though the dialogue datasets possess different characteristics as indicated in Section 4.1, DynaEval exhbits robust performance across all the datasets. This confirms our hypothesis that DynaEval provides useful dialogue-level representation for distinguishing the original dialogues from the corresponding negative samples. Especially when compared to S-Dicoh, which mod-

Model	Empathetic		ConvAI2		DailyDialog	
	UR	SS	UR	SS	UR	SS
RANDOM	50.07	50.07	50.25	50.25	50.17	49.62
CoSim	63.54	63.33	68.79	92.93	69.59	63.80
S-DiCoh	80.33 ± 2.83	86.04 ± 0.31	66.80 ± 1.93	90.35 ± 0.08	83.67 ± 0.41	84.92 ± 0.70
DynaEval	**94.30 ± 0.07**	**90.37 ± 0.37**	**85.23 ± 0.96**	**98.65 ± 0.29**	**91.89 ± 0.58**	**91.65 ± 0.62**

Table 2: The accuracy (%) of DynaEval vs baselines on the test sets of Empathetic Dialogue and DailyDialog as well as the validation set of ConvAI2. UR & SS are the sampling strategies defined in Section 3.5. Experiments involving training are repeated five times with different random seeds for model weights initialization. The average and standard deviation are reported in the table.

els a dialogue sequentially with bidirectional LSTM and does not explicitly incoporate the speaker level interaction, the structured graph modeling of a dialogue in DynaEval is more effective for capturing both the interaction between the interlocutors and the contextual information within a dialogue.

Based on the experimental results, it can be deduced that the discrimination task with UR strategy is more challenging compared to that with SS strategy. The accuracy scores achieved by S-DiCoh in the SS category is much higher than that in the UR category on both datasets. Similar observation can be made w.r.t CoSim and DynaEval on the ConvAI2 dataset. DynaEval performs remarkably in this task as it outperforms S-DiCoh by a significant margin of 13.97, 18.43 and 8.22 on Empathetic Dialogue, ConvAI2 and DailyDialog respectively. Given these observations, we further hypothesize that DynaEval model trained with UR strategy offers more useful dialogue representation to the dialogue evaluation task.

4.3 Dialogue Evaluation Task

To validate the above hypothesis, we assess the usefulness of DynaEval in both the dialogue-level and turn-level evaluation tasks. In both settings, Spearman correlations between the scores generated by DynaEval and the corresponding human evaluation scores are computed. The performance of DynaEval is compared against several recently proposed dialogue evaluators.

Evaluation Dataset FED (Mehri and Eskenazi, 2020a) is a bench-marking dataset useful for both dialogue-level and turn-level evaluation. It contains both human-human conversations and human-chatbot conversations, which are collected by the authors of the Meena chatbot (Adiwardana et al., 2020) in an interactive setup. In total, 124 conversations are collected, out of which 40 come from

interacting with the Meena Chatbot, 44 come from interacting with the Mitsuku Chatbot and 40 are drawn from human-human conversations. The average number of utterances per conversation is 13.72 and the average number of words per utterance is 9.23. Human quality annotations of these conversations are performed at both the dialogue and turn level. There are 9 quality aspects for turn-level annotations and 11 for dialog-level annotations outlined in the first column of Table 3. FED includes 3348 turn-level and 1364 dialog-level annotations, for a total of 4712. The inter-annotator agreements for all the quality aspects, which indicate the metric performance upper bound, is shown in the last column of Table 3.

Metrics to Compare The recently proposed reference-free state-of-the-art dialogue metrics, including USR (Mehri and Eskenazi, 2020b), BERT-RUBER (Ghazarian et al., 2019) (BERT-R), GPT-2 based coherence metric (Pang et al., 2020) (GPT-2) and FED (Mehri and Eskenazi, 2020a)[6], serve as the baseline dialogue evaluators. Since USR, BERT-R and GPT-2 are turn-level metrics, aggregation of all the turn-level scores in a dialogue is required for dialogue-level evaluation. The best correlation scores at dialogue level are reported in Table 3 among all the aggregation strategies for these three metrics. For completeness, we report their correlation scores w.r.t difference aggregation strategies in Appendix A.2. Similar to DynaEval, S-Dicoh provides a unified score for each dialogue. Based on insights from Section 4.2, the best performing model in the UR category is chosen to score the dialogues for both S-Dicoh and DynaEval.

Dialogue-level Evaluation DynaEval achieves

---

[6]The correlation scores of FED is obtained from the original paper. For each evaluation category, the highest score is reported among the scores provided by all its variants.

	Dialogue-level Spearman Correlation						
Dialogue Aspects	**BERT-R**	**GPT-2**	**USR**	**S-DiCoh**	**FED**	**DynaEval**	**Human**
Coherence	0.229	*0.123*	0.194	*0.038*	0.251	**0.423**	0.809
Error Recovery	0.242	*0.096*	*0.170*	-0.054	*0.165*	**0.311**	0.840
Consistency	*0.163*	*0.091*	0.169	*0.017*	*0.116*	**0.352**	0.562
Diversity	0.196	*0.147*	0.242	*0.059*	**0.449**	0.332	0.789
Topic Depth	0.192	*0.097*	0.341	*0.046*	**0.522**	0.439	0.833
Likability	0.281	0.179	0.221	-0.070	0.262	**0.398**	0.838
Understanding	0.198	*0.070*	0.172	-0.100	0.306	**0.361**	0.809
Flexibility	0.253	*0.134*	0.209	*0.044*	**0.408**	0.389	0.816
Informativeness	0.211	*0.116*	0.288	*0.028*	0.337	**0.396**	0.806
Inquisitiveness	0.337	*0.071*	0.188	-0.054	0.298	**0.388**	0.769
Overall	0.248	*0.123*	0.288	-0.073	0.443	**0.482**	0.830
	Turn-level Spearman Correlation						
Interestingness	0.235	-0.107	*0.085*	*0.031*	**0.431**	0.289	0.819
Engagement	0.206	-0.086	0.107	*0.040*	**0.318**	0.255	0.798
Specificity	0.327	-0.112	*0.095*	*0.062*	**0.326**	0.272	0.790
Relevance	0.151	-0.105	0.183	-0.051	0.152	**0.265**	0.753
Correctness	*0.081*	*0.041*	0.098	-0.040	0.133	**0.216**	0.780
Semantically Appropriateness	*0.044*	-0.084	0.201	-0.069	0.177	**0.233**	0.682
Understandable	*0.051*	-0.071	0.110	-0.075	0.111	**0.185**	0.522
Fluency	*0.079*	-0.151	0.220	-0.007	**0.224**	*0.096*	0.714
Overall	0.195	*-0.095*	0.137	-0.022	0.209	**0.264**	0.820

Table 3: Comparison of both dialogue and turn level Spearman correlations among state-of-the-art automatic metrics on the FED evaluation dataset. The results are reported for the 11 and 9 unique quality categories at turn and dialogue level respectively. Scores with p-values larger than 0.05 are italicized (indicating statistical insignificance). The best score for each category is highlighted in bold.

the highest correlation scores in 8 out of 11 dialogue aspects, including the overall category. For the other three categories, DynaEval attains second highest correlation scores. We can see that DynaEval significantly outperforms S-DiCoh. These results showcase that structured graph modeling of a dialogue with explicit incorporation of speaker and utterance level dependencies provides meaningful dialogue-level representations. Such representations capture information of various dialogue attributes that are beneficial for the dialogue-level evaluation task.

Moreover, BERT-R, GPT-2 and USR are state-of-the-art turn-level evaluation metrics. They evaluate a dialogue based on aggregation of scores of all the context-response pairs within the dialogue. It can be observed that their correlation scores across individual dialogue aspects are not as high as those of DynaEval. This supports our hypothesis in Section 1 that turn-level quality evaluation may be insufficient to assess the performance of open-domain dialogue systems.

In addition, dialogue aspects, including coherence, likability, informativeness and Inquisitiveness, are highly dependent on the interaction of the interlocutors. Amongst all the dialogue aspects,

DynaEval achieves significantly higher scores in these four categories. This attributes to its incorporation of the speaker level dependency.

Turn-level Evaluation Furthermore, it can be observed that DynaEval achieves the highest correlation in 5 out of 9 categories including the overall category. This demonstrates that DynaEval is not only useful for holistic evaluation of a dialogue, but also useful for turn level evaluation. In this sense, DynaEval serves as a better proxy to the human evaluation process (Li et al., 2019) whereby humans mainly evaluate the conversations in a holistic manner and laser-focus on the problematic turns.

Specifically, DynaEval performs well in turn-level aspects, such as relevance, semantic appropriateness and correctness. These aspects highly correlate to the dialogue-level attributes, such as coherence and understanding, suggesting that the evaluation of these turn-level attributes also benefit from the explicit modeling of the speaker and utterance level interaction in a unified framework.

Error Analysis An interesting finding is that DynaEval and FED actually complement each other at both dialogue and turn level. For example, at the dialogue level, FED performs well in diversity and topic depth, but struggles with coher-

ence and consistency. DynaEval performs well in coherence and consistency, but its performance in diversity is much lower in comparison to FED. This may be because dialoGPT, the backbone of FED, was trained on a large amount of Reddit data, which contain diverse amount of topics and variation of expressions while DynaEval is trained on a single dialogue domian. Moreover, dialoGPT does not explicitly model such speaker-level interaction, but DynaEval does. Hence, DynaEval is more useful for evaluating coherence and consistency aspects of a dialogue. One way to improve DynaEval for evaluating topic depth and diversity is to pre-train on a large amount of dialogue data with a variety of topics and then fine-tune it on the target domain.

Another observation is that DynaEval performs significantly poorer for the fluency aspect at turn-level than for other turn-level aspects. Additionally, GPT-2, USR and FED, which leverage pretrained language model, perform significantly better than DynaEval in this category. This may be because DynaEval directly models a dialogue at the utterance level instead of at the token level, while the other metrics consider the language modeling objective, which focuses more on the token-level dependencies rendering them effective for evaluating the naturalness of a response. A remedy to this problematic aspect of DynaEval is to introduce perturbation strategies targeting the token level, such as word drop, word shuffling and word replacement (Sinha et al., 2020; Park et al., 2021). Such strategies provide negative samples mimicking the non-sensical or non-grammatical responses produced by certain seq2seq generative models. Another simple solution is to combine DynaEval with turn-level metrics specifically designed for evaluating naturalness of dialogue responses.

Besides the fluency aspect, DynaEval's performance in interestingness, engagement and specificity at the turn level is not as pronounced as that of FED. This may be because purely modeling the dialogue itself is not enough for all the aspects. The model may need to incorporate external knowledge concerning a diverse range of topics to be able to reflect these attributes. The same conclusion can also be drawn from DynaEval's relatively weaker performance in the diversity category at the dialogue level.

Lastly, DynaEval primarily targets open-domain dialogues where there is no clear or predefined task to perform. When evaluating task-oriented dialogues, task completion will take a more central role. Meta-information such as intents and request types are important to determine task completion and therefore, the evaluation framework will require further adaptation accounting for these information when evaluating task-oriented dialogues.

5   Conclusion & Future Work

DynaEval serves as a unified framework for both turn and dialogue level evaluation in open-domain dialogue. It provides meaningful representations that incorporate information reflecting various important dialogue attributes. Its explicit modeling of speaker and utterance level interaction leveraging GCN has been proven beneficial for the evaluation task. Lastly, the error analysis in Section 4.3 sheds light on how DynaEval can be further improved. DynaEval can also be combined with the specialized turn-level metrics, such as those targeting fluency and engagement, to fully approximate the interactive human evaluation process.

Acknowledgement

Ethical Considerations & Broader Impact

This study conforms to the prevailing ethical guidelines. All datasets used are in the public domain. In addition, we have identified a way that DynaEval can help address the ethical concerns. By explicitly training the framework to discriminate safe dialogues from unsafe ones, it can help detect dialogues containing inappropriate sentences, such as those regarding injustice and discrimination. Such application may be useful in many real-life scenarios where the behaviors of chatbots need to be properly monitored to avoid insensitive and irresponsible comments from the chatbots.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Alessandra Cervone and Giuseppe Riccardi. 2020. Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, 1st virtual meeting. Association for Computational Linguistics.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. *Proc. Interspeech 2018*, pages 1011–1015.

Lu Chen, Bowen Tan, Sishan Long, and Kai Yu. 2018. Structured dialogue policy with graph neural networks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1257–1268, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Fernando D'Haro, Rafael E Banchs, Chiori Hori, and Haizhou Li. 2019. Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech & Language*, 55:200–215.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Sudeep Gandhe and David Traum. 2016. A semi-automated evaluation metric for dialogue model coherence. In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 217–225. Springer.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *Advances in Neural Information Processing Systems*, 32:13658–13669.

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.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, Hong Kong, China. Association for Computational Linguistics.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. 9. Routledge.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2014. Evaluating coherence in open domain conversational systems. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–37.

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 860–870, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

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.

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. *arXiv preprint arXiv:1907.11692*.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450, Online. Association for Computational Linguistics.

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020. Language model transformers as evaluators for open-domain dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6797–6808, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020. I like fish, especially dolphins: Addressing contradictions in dialogue modelling. *arXiv preprint arXiv:2012.13391*.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue

generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

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.

ChaeHun Park, Eugene Jang, Wonsuk Yang, and Jong Park. 2021. Generating negative samples by manipulating golden responses for unsupervised learning of a response evaluation model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1525–1534, Online. Association for Computational Linguistics.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2020. Co-GAT: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *arXiv preprint arXiv:2012.13260*.

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

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 593–599, Online. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. 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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

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

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken

dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75, Tokyo, Japan. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Rafael E Banchs, Thomas Friedrichs, and Haizhou Li. 2021a. Deep AM-FM: Toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade*, pages 53–69. Springer.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021b. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 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*, pages 270–278, Online. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. 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 165–176, Hong Kong, China. Association for Computational Linguistics.

Yunxiao Zhou, Man Lan, and Wenting Wang. 2019. Hierarchical intention enhanced network for automatic dialogue coherence assessment. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

A Additional Experimental Results

A.1 Utterance-level Pooling Techniques

To derive the dialogue-level representation, we have adopted the mean pooling method in DynaEval. In this section, we examine the effects of different pooling methods in the dialogue-level discrimination task. Specifically, we compare the performance of mean pooling against max pooling and the concatenation of sentence vectors derived with both mean and max pooling. The performance comparison is presented in Table 4. It can be observed that the performance difference across various pooling strategies is not statistically significant.

Strategy	UR	SS
Mean	94.30 ± 0.07	90.37 ± 0.37
Max	94.17 ± 0.16	90.75 ± 0.24
Mean+Max	94.19 ± 0.04	90.64 ± 0.06

Table 4: The accuracy scores (%) of DynaEval on the test set of Empathetic Dialogue with different utterance-level pooling techniques. The average and standard deviation are reported in the table.

A.2 Dialogue-level Correlation Analysis of Turn-level Metrics

For each turn-level metric, we have applied four simple aggregation strategies to derive dialogue level scores from their respective constituent turn level scores: (1) Mean, (2) Sum, (3) Max and (4) Multiplication. The dialogue level correlation coefficients of USR, BERT-RUBER and GPT-2 based coherence metric are reported in Table 5, Table 6 and Table 7 correspondingly. Note that for turn-level metrics leveraging the language model objective, we don't consider token-level aggregation variants. Instead, we follow the same formulations in the original papers. For example, the GPT-2 based coherence metric (Pang et al., 2020) computes a turn-level score based on averaging the token-wise conditional log probabilities in the corresponding response.

It can be observed that all three metrics don't perform well at dialogue level evaluation. This further validates our statement in Section 1 that turn-level quality evaluation may be insufficient to assess the performance of open-domain dialogue systems as they don't specifically model the interaction over an entire dialogue.

Quality	Mean	Sum	Max	Prod
USR				
Coherence	**0.194**	*0.111*	*0.021*	*0.158*
Error Recovery	***0.170***	*0.083*	*0.075*	*0.130*
Consistency	*0.150*	***0.169***	*0.038*	*0.099*
Diversity	**0.242**	*0.167*	0.235	0.193
Topic Depth	**0.341**	*0.145*	0.255	0.295
Likability	**0.221**	0.193	*0.109*	*0.126*
Understanding	***0.172***	*0.112*	*0.004*	*0.124*
Flexibility	**0.209**	*0.151*	*0.164*	*0.129*
Informativeness	**0.288**	*0.157*	*0.171*	*0.237*
Inquisitiveness	0.148	*0.099*	**0.188**	*0.128*
Overall	**0.288**	*0.166*	*0.094*	0.212

Table 5: Dialogue level Spearman correlation coefficients of USR w.r.t different turn-level aggregation strategies on the FED dataset. Scores with p-values larger than 0.05 are italicized (indicating statistical insignificance). The best score for each category is highlighted in bold.

Quality	Mean	Sum	Max	Prod
BERT-R				
Coherence	0.222	0.221	**0.229**	*0.041*
Error Recovery	0.231	**0.242**	0.228	*0.005*
Consistency	*0.141*	***0.163***	*0.148*	*-0.030*
Diversity	0.180	**0.196**	*0.164*	*-0.051*
Topic Depth	0.181	**0.192**	*0.163*	*0.008*
Likability	0.256	**0.281**	0.249	*-0.037*
Understanding	0.189	**0.198**	0.189	*-0.023*
Flexibility	0.228	**0.253**	0.232	*-0.036*
Informativeness	0.194	**0.211**	0.186	*-0.023*
Inquisitiveness	0.326	**0.337**	0.331	*0.056*
Overall	0.231	**0.248**	0.224	*-0.021*

Table 6: Dialogue level Spearman correlation coefficients of BERT-RUBER w.r.t different turn-level aggregation strategies on the FED dataset.

B Reproducibility

B.1 Training Setup & Hyperparameters

For all the experiments involving training, we run the experiments five times with different random seeds for model weights initialization to reduce the risk of randomness. The experiments are performed on a single Tesla V100 32GB GPU with a batch size of 512. The model is trained for 20 epochs and its parameters are optimized using the Adam optimizer. The average run time for each epoch is around 8 hours and 15 minutes. The initial learning rate is set to 0.002 and decays by a factor

Quality	Mean	Sum	Max	Prod
GPT-2				
Coherence	*-0.002*	***0.123***	*-0.086*	*-0.120*
Error Recovery	*0.034*	***0.096***	*-0.057*	*-0.091*
Consistency	*-0.025*	***0.091***	*-0.048*	*-0.088*
Diversity	*0.092*	***0.147***	*-0.033*	*-0.145*
Topic Depth	*0.054*	***0.097***	*-0.036*	*-0.094*
Likability	*0.072*	***0.179***	*-0.047*	*-0.175*
Understanding	*-0.027*	***0.070***	*-0.062*	*-0.066*
Flexibility	*0.056*	***0.134***	*-0.032*	*-0.131*
Informativeness	*0.025*	***0.116***	*-0.100*	*-0.112*
Inquisitiveness	*-0.008*	***0.071***	*-0.071*	*-0.070*
Overall	*-0.002*	***0.123***	*-0.086*	*-0.120*

Table 7: Dialogue level Spearman correlation coefficients of GPT-2 based coherence metric w.r.t different turn-level aggregation strategies on the FED dataset.

of 0.5 per epoch. A dropout of 0.5 is also applied.

For Empathetic Dialogue and DailyDialog, the context window length, M is set to 4, because these two datasets contain relatively short conversations (4.31 and 7.90 average number of utterances per dialogue respectively). A context window size of 4 ensures each utterance is connected to all the remaining utterances in most of the dialogues. The utterances may provide important contextual information to each other within a dialogue. For ConvAI2, M is set to 2 to avoid introducing too much irrelavant context information. This is because most of the conversations in ConvAI2 are about two people getting to know each other and there are frequent topic changes in the conversations. M serves as an important hyperparameter to control the influence of an utterance on the rest in a dialogue.

For training DynaEval, we have filtered out dialogues of which the number of utterances is less than 4 or more than 30. We hypothesize that dialogues with less than 4 utterances containing little information for modeling speaker and utterance level interaction. Moreover, there are very few dialogues with more than 30 utterances in both datasets. Including them leads to large graphs and unnecessary paddings, which slow down the training process.