

D-Score: Holistic Dialogue Evaluation Without Reference

Chen Zhang[✉], *Student Member, IEEE*, Grandee Lee[✉], *Student Member, IEEE*,
Luis Fernando D'Haro[✉], *Member, IEEE*, and Haizhou Li[✉], *Fellow, IEEE*

Abstract—In artistic gymnastics, difficulty score or D-score is used for judging performance. Starting from zero, an athlete earns points from different aspects such as composition requirement, difficulty, and connection between moves. The final score is a composition of the quality of various performance indicators. Similarly, when evaluating dialogue responses, human judges generally follow a number of criteria, among which language fluency, context coherence, logical consistency, and semantic appropriateness are on top of the agenda. In this paper, we propose an automatic dialogue evaluation framework called D-score that resembles the way gymnastics is evaluated. Following the four human judging criteria above, we devise a range of evaluation tasks and model them under a multi-task learning framework. The proposed framework, without relying on any human-written reference, learns to appreciate the overall quality of human-human conversations through a representation that is shared by all tasks without over-fitting to individual task domain. We evaluate D-score by performing comprehensive correlation analyses with human judgement on three dialogue evaluation datasets, among which two are from past DSTC series, and benchmark against state-of-the-art baselines. D-score not only outperforms the best baseline by a large margin in terms of system-level Spearman correlation but also represents an important step towards explainable dialogue scoring.

Index Terms—Automatic Dialogue Evaluation, Holistic Framework, Multi-task Learning, Self-supervised Learning.

I. INTRODUCTION

RECENT studies in open-domain dialogue systems call for a good evaluation framework with objective evaluation

metrics to gauge the research progress [1], and to provide appropriate feedback for dialogue model training [2]. Dialogue evaluation seeks to evaluate machine responses, in a human-machine dialogue system, similarly to human judgment. However, machine evaluation performance is still far from human evaluation [3], as many objective evaluation metrics tend to focus on a narrow perspective, while humans usually judge from a holistic view across multiple aspects [4].

Dialogue evaluation studies are broadly classified into reference-based and reference-free techniques. Reference-based approaches compare lexical overlap between system responses and defined reference targets, which is rather restrictive [1], [5] considering there are a large variety of suitable responses for the same dialogue history. In addition, the process of manual annotation and creation of references is expensive and hard to scale [6]. On the other hand, reference-free metrics are based on the analysis of the dialogue context and how a given system response is considered appropriate, without requiring any annotation and allowing any kind of response to be evaluated. Therefore, a reference-free evaluation solution is highly sought after.

Deep learning has enabled, in the last few years, many reference-free evaluation metrics [7]–[10] with some limited success since they are trained with strong language modeling objectives. They are able to assess *language fluency* and *context coherence* of dialogue responses to certain extent. However, they are limited when assessing other aspects such as *logical consistency*, *semantic appropriateness* [11] and *user engagement* [12] systematically. For machine evaluation to approach human performance, we identify two research problems: a) to define evaluation metrics that describe different dialogue aspects, and b) to establish a holistic solution that considers the inter-dependence of the different aspects.

First, a model-based evaluation metric in machine learning is trained for one objective, such as minimizing a prediction error or maximizing a global reward [13]. In this way, multiple independently trained models are required for multiple objectives. Though acceptable performance can be attained by being laser-focused on individual tasks, shared information across individual tasks from the training data is not exploited [14].

Humans evaluate the quality of dialogue from different aspects [15], so does a system. We cannot expect a model optimized for one evaluation aspect to perform well on other aspects or to provide a good overall score [8]. We consider four high-level linguistic properties, *language fluency*, *context coherence*,

Manuscript received October 15, 2020; revised January 27, 2021 and April 3, 2021; accepted April 7, 2021. Date of publication April 20, 2021; date of current version August 6, 2021. This work was supported in part by National Research Foundation Singapore under its AI Singapore Programme Award AISG-GC-2019-002, in part by Human-Robot Interaction Phase 1 under Grant 19225 00054, National Research Foundation (NRF) Singapore under the National Robotics Programme, in part by Human Robot Collaborative AI for AME under Grant A18A2b0046, NRF Singapore, in part by Robert Bosch (SEA) Pte Ltd under EDB.s Industrial Postgraduate Programme II (EDB-IPP), Project Title: Applied Natural Language Processing, in part by the Spanish Projects: AMIC (MINECO, TIN2017-85854-C4-4-R), and in part by CAVIAR (MINECO, TEC2017-84593-C2-1-R) projects partially funded by the European Union. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hannes Schulz. (*Corresponding author: Chen Zhang.*)

Chen Zhang, Grandee Lee, and Haizhou Li are with Human Language Technology Group at Electrical and Computer Engineering Department of National University of Singapore, Singapore 117583, Singapore (e-mail: e0397123@u.nus.edu; grandee.lee@u.nus.edu; eleliha@nus.edu.sg).

Luis Fernando D'Haro is with the Speech Technology Group of Universidad Politécnica de Madrid (ETSIT, UPM), 28040 Madrid, Spain (e-mail: lfdharo@gmail.com).

Digital Object Identifier 10.1109/TASLP.2021.3074012

logical consistency and *semantic appropriateness*. They are identified as the properties that highly correlate with human’s perception of what constitutes a good conversation [15]–[18]. We hypothesize several relevant tasks pertaining to these linguistic properties that are naturally complementary to the task of dialogue evaluation with each account for a single evaluation aspect.

Second, to evaluate generative dialogue systems, a model-based evaluation framework is supposed to leverage information beyond what is known to the generative systems or a wider range of inductive biases. This can be illustrated with some real-life examples. For example, when scoring argumentative essays written by the students, teachers generally have a long list of references on the essay topic, either in terms of content or in terms of the methodologies to position the arguments. Similarly, gymnastics judge should have reviewed the performance by hundreds of gymnasts in advance as to how different aspects of views come together to make a good execution, such as the connection between moves and agility of the gymnasts. By the same token, a holistic framework, equipped with general linguistic knowledge and capable of connecting different dialogue aspects from human-human conversations, is desirable in automatic dialogue evaluation.

Furthermore, a dialogue aspect only provides a partial indication of the overall dialogue quality. When optimizing the evaluation metrics for a single aspect, the model tends to over-fit the task. In essence, each dialogue aspect constitutes a noisy data-dependent task that correlates partially with human judgement. Our aim is to learn a good overall metric while ignoring the data- and user-dependent noise and generalize well. For example, the natural language inference task from the perspective of the main evaluation task provides such inductive bias in terms of identifying the logical consistency of the context-response pair. The more tasks we are learning simultaneously, the more our model has to find a representation that captures information contributing to all the tasks and the less is our chance of over-fitting on individual aspects. As such, we are motivated to approach the multi-faceted evaluation with multi-task learning leveraging data amplification and eavesdropping [13], [14] to enhance the framework’s performance and learning capability.

The main contributions of this paper include,

- 1) A reference-free dialogue evaluation framework referred to as D-score, that provides an overall score derived from explainable dialogue aspects;
- 2) A framework that scales to different dialogue aspects, and applicable across dialogue domains.
- 3) To the best of our knowledge, D-score is the first multi-aspect scorer, with a shared dialogue knowledge representation for a holistic view.

In Section II, we will present the related work to set the stage for this study. In Section III, we will formulate the D-score framework in detail. In Section IV, the datasets and baselines will be presented. In Section V, we will discuss the experiment results. Finally, Section VI will conclude the paper. The codes and resources in this work are made available to the public.¹

II. RELATED WORK

We begin by looking into the human judging aspects for dialogue responses, and two types of evaluation metrics. We will then review related work where multi-task learning improves the generalization power of deep-learning-based models.

A. Aspects of Evaluation

In general, the effectiveness of an automatic evaluation framework is quantified by the extent to which it mimics human judgements [2]. Therefore, it is essential to consider the aspects human judges look into when evaluating the quality of dialogue. Finch & Choi [4] have conducted a comprehensive survey on 20 recent dialogue system papers and summarized the linguistic aspects that human judges generally consider when evaluating the dialogue systems. The most evaluated ones include *language fluency*, *context coherence*, *semantic appropriateness* and *logical consistency*.

By *language fluency*, we mainly evaluate the naturalness of the response in terms of syntactical quality or the adequacy of expression. *Semantic appropriateness* seeks to determine whether the response topically fit into its corresponding dialogue context. Many existing evaluation frameworks have focused on the two above-mentioned aspects [18]–[20]. *Context coherence* is concerned with the structure of the dialogue flow, such that the passage is expressed fluidly, clearly, and with sensible turn-taking [21]. *Logical consistency* is concerned with whether the turn-taking demonstrates clear logical exchange such that the knowledge or information of the context-response pair is not contradictory [11]. Evaluating *Logical consistency* is a long-standing goal for automatic dialogue evaluation because existing open-domain generative dialogue models tend to have a short memory and often contradict themselves. [22], [23].

B. Reference-Based Metrics

Many dialogue evaluation metrics are inspired by studies in machine translation, text summarization, and more recently image captioning. Most of these metrics belong to the reference-based category, which evaluates the generated dialogue responses within a constrained space, i.e., given a dialogue context, a list of reference responses are provided for semantical or syntactical comparisons. N-gram overlap metrics such as F1 score, BLEU [24], METEOR [25] and ROUGE [26] correlate poorly with the human annotation of dialogue responses [1], [6], [27]. This is largely due to the intrinsic inability of such metrics to handle the supposedly highly diverse dialogue responses [1]. Dinan *et al.* [28] also point out that F1 score can be gamed if one generates the most frequent tokens regardless of the context.

There are also embedding-based metrics, such as greedy matching [29], embedding average [30] and vector extrema [31], which are based on sentence-level embedding. The evaluation score is derived from the distance of embeddings between system responses and target references, which aims to avoid directly measuring lexical overlap. Embedding average and vector extrema are heuristic techniques that derive sentence-level embedding from token-level embedding, whereas greedy

¹<https://github.com/e0397123/D-score.git>

matching does not require the computation of sentence-level embedding. Instead, it matches the embedding of each token in the response sequence to that of a token in the reference based on the maximum cosine similarity.

Word embeddings are typically learned via distributional semantics. Examples include GloVe [32] and FastText [33]. More recently, contextualized embedding [34] is becoming increasingly popular for semantic representation. BERTScore [35] and BLEURT [36] are two recently proposed examples leveraging the contextualized representation. Although embedding-based metrics avoid directly comparing lexical overlap, they do not address the issue that there potentially exist many valid responses to one dialogue context. One solution is to come up with multiple reference responses [6]. However, such a solution is costly and hard to scale.

Another group of metrics, such as ADEM [27] and Alexa unified framework [17], directly adopt human-annotated response scores as the supervision signal to train a score prediction model. Such metrics are not scalable in general due to the high annotation cost. Essentially, all reference-based metrics are dependent on human annotations.

C. Reference-Free Metrics

There are two ideas for reference-free metrics. One is to continue with humans in the loop but make it more robust, faster and cheaper, such as ACUTE-EVAL [37] and SSA [38]. Another is to move towards fully automated evaluation. Typical automatic reference-free metrics take advantage of large human-human dialogue data and are designed with specialized learning objectives to mimic a specific aspect of human evaluation. Many existing works along with this line focus on evaluating language fluency. Typical examples include perplexity [28], [38], USR-MLM [8] and GPT-2 [39] based fluency metric [20]. USR-MLM leverages a fine-tuned RoBERTa model [40] to measure the token-wise log-likelihood and the negative of the sum of all the token-wise log-likelihoods within a response is used to determine the fluency aspect of that particular response. [20] measures the fluency aspect of a response via computing its log-likelihood with autoregressive GPT-2. These studies suggest that language fluency is an important component of the overall dialogue quality determination.

In addition, another group of metrics tries to measure context coherence. This aspect pertains to the idea of lexical cohesion [41], which belongs to the study of semantic field coherence. Foltz *et al.* [42] propose Latent Semantic Analysis (LSA) coherence approach, which is the first to use embeddings for modeling local coherence. The relationship between two consecutive sentences is determined by the cosine between their corresponding LSA sentence embedding vectors. More recent approaches [21], [43], [44] rely on deep neural network for learning context coherence. MAUDE [10], a recent state-of-the-art reference-free metric tailored for online dialogue evaluation, leverages a pre-trained DistilBERT [45] model to extract the semantic representation of dialogue turns and uses bidirectional LSTM to explicitly model the discourse structure with the intention of capturing the temporal transition between dialogue turns.

It seeks to evaluate the context coherence aspect of dialogue. These studies all point to the fact that context coherence should be an essential ingredient in designing reference-free evaluation metrics.

Besides context coherence and language fluency, the semantic appropriateness of dialogue is also extensively studied in the existing literature. It mostly concerns the relevance of the response with respect to its corresponding dialogue context. The reference-free component of RUBER is trained by bootstrapping negative responses directly from the dataset and a recurrent neural network (RNN) is used to predict the relatedness between these responses and the corresponding context. [7] improves RUBER with the contextualized representation derived from BERT [34]. The RUBER-based metrics are highly specialized in evaluating the semantic appropriateness of dialogue responses. The USR-DR metric [8] makes use of a response prediction model to discern whether the response is the true response or a randomly sampled one given the dialogue context. Hence, these studies on semantic appropriateness suggest that it is as important as both context coherence and language fluency.

Another school of thought is to evaluate the logical consistency aspect of dialogue. Welleck *et al.* [11] explore the use of natural language inference (NLI) models such as ESIM [46] and InferSent [47] for identifying contradictions within a dialogue. In addition, they manually annotate an NLI dataset based on relationship triplets derived from either persona sentences or dialogue turns extracted from the PERSONA-CHAT dataset [23]. They demonstrate the potential of applying NLI techniques for evaluating consistency within a dialogue. Dziri *et al.* [16] propose an NLI evaluation framework that does not require manual annotations. Rather, they rely on various sampling strategies to construct a conversational NLI dataset: the original on-topic response is regarded as entailment to the corresponding dialogue context. Grammatically-impaired sentences constructed by randomly selected words from the context as well as randomly sampled contradictory instances from MultiNLI corpus are labelled as contradictions. Lastly, random responses from other conversations and generic responses to the corresponding contexts are treated as neutral. Pang *et al.* [20] leverage a pre-trained RoBERTa model to determine the logical self-consistency of a dialogue model. Thus, these studies demonstrate that logical consistency is an essential contributing factor to dialogue quality evaluation.

Yet, all the above-mentioned metrics are highly optimized for a specific aspect of the evaluation. To evaluate the overall quality of a dialogue, a typical way is to apply multiple metrics, each describing one dialogue aspect. For example, Ghandeharioun *et al.* [48] adopt a learned ensemble metric containing three individual components to evaluate three dialogue aspects, sentiment coherence, engagement as well as semantic coherence. Pang *et al.* [20] apply different pre-trained models and data augmentation technique to evaluate language fluency, context coherence, logical self-consistency as well as response diversity. USR [8] combines the USR-MLM and USR-DR components for measuring semantic appropriateness and language fluency. This approach assumes there is no

interdependence or interaction among the dialogue aspects, therefore, each metric works independently.

D. Shared Knowledge Across Aspects

Studies show that humans have difficulties rating different aspects of text independently [49]. One of the main problems affecting human judgments across different aspects is due to the intrinsic human's inability to objectively separate information from different aspects, such as adequacy and fluency of a sentence [49]. As a result, human judgments across aspects are highly correlated. We have similar observations in the Persona-Chatlog evaluation dataset [15] where the average Pearson correlation of pairwise humans scores for different dialogue aspects is 0.724 at the system level and 0.393 at the conversation level. This suggests that when humans are judging the individual aspects of a dialogue, they tend to judge more from an overall perspective, and less from independent aspects objectively. We argue that a shared knowledge representation for the evaluation of various aspects would be advantageous for holistic dialogue evaluation, and a multi-task learning framework could serve as a suitable computational model.

E. Multi-Task Learning

Multi-task learning (MTL) improves generalization by leveraging the domain-specific information contained in the training signals of related tasks [14], [50]–[52]. Ruder in his overview of MTL [13] summarized the five advantages of MTL over the single-task training paradigm: implicit data augmentation, attention focusing, eavesdropping, representation bias and regularization. MTL has been effective in many natural language processing tasks, such as text classification [53], machine translation [54] and language modeling [34], [55]. Liu *et al.* [53] make use of related tasks such as query classification and web search to serve as an inductive bias for the main task of semantic classification. Luong *et al.* [54] make use of multiple languages to train a shared encoder for machine translation.

More related to our work, [56] uses part-of-speech tags, chunks, named entity tags, semantic roles, etc and combined with language modeling objective to demonstrate that all the shared task benefit from MTL. In [34], the authors adopt the masked language model and next sentence prediction objectives for joint optimization of the deep bidirectional transformer encoder (BERT) and [55] improves BERT by replacing the original next sentence prediction objective with sentence order prediction, but maintains the multi-task setup for training their model. While there is no prior work on MTL for automatic dialogue evaluation, we draw inspirations from other natural language processing tasks and propose a D-score framework, which will be elaborated in Section III.

III. D-SCORE FRAMEWORK

D-score framework features several unique evaluation strategies. (1) D-score arrives at an overall judgement from several aspects, while the aspect judgements are built on a shared knowledge representation about the dialogue; (2) We consider

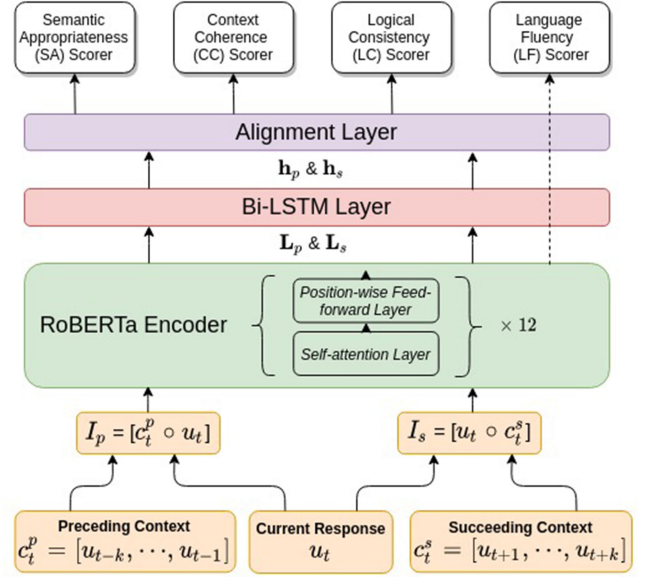


Fig. 1. D-score network architecture that is formulated under a multi-task learning framework with four optimization tasks, namely *semantic appropriateness*, *context coherence*, *logical consistency*, and *language fluency*.

that the quality of a dialogue turn should be judged not only by its preceding but also its succeeding context; (3) We propose a range of appropriate sampling strategies for multi-task learning from human-human conversation data.

The idea is to train a D-score framework from human-human dialogues, and use it to evaluate human-chatbot dialogues. To better understand how the idea works, let's formally define the evaluation task and the high-level architecture of D-score that is shown in Fig. 1.

A. Input Formulation

Let D denotes a human-human turn-taking dialogue that consists of N utterances,² where N is at least 2, i.e., $D = \{u_1, u_2, \dots, u_N\}$. Let the current response be u_t , where $t \in [1, N]$. All the previous dialogue context of u_t is $\{u_1, \dots, u_{t-1}\}$ and the entire succeeding dialogue context of u_t is $\{u_{t+1}, \dots, u_N\}$. In practice, a window size of k is selected to constrain the size of dialogue context to manage the computing load, and to keep only useful contextual information. We denote the preceding context as $c_t^p \in [u_{t-k}, \dots, u_{t-1}]$ and the succeeding context as $c_t^s \in [u_{t+1}, \dots, u_{t+k}]$ respectively.

We append an artificial sentence $\langle \text{start of dialogue} \rangle$ in front and $\langle \text{end of dialogue} \rangle$ behind to start and end a dialogue. c_t^p , u_t and c_t^s are sequences of tokens w , with x , y and z number of tokens respectively. A sequence of tokens can be obtained by a segmentation algorithm. In D-score, Byte-Pair Encoding (BPE) [57], which works for the RoBERTa model [40], is applied. The two inputs of D-score take the following form:

$$I_p = [c_t^p \circ u_t] \text{ or } [w_1^p, w_2^p, \dots, w_{(x+y)}^p] \quad (1)$$

$$I_s = [u_t \circ c_t^s] \text{ or } [w_1^s, w_2^s, \dots, w_{(y+z)}^s] \quad (2)$$

²A dialogue turn is referred to as an utterance in this paper.

I_p and I_s allows D-score to capture the bidirectional dialogue contextual information with respect to the current response u_t . This is inspired by the bidirectional language models [34], [58] which capture the contextualized information of the current word. This formulation is in contrast with that of most existing reference-free frameworks which only considers one-sided dialogue history. While such bidirectional formulation deprives D-score of the ability for online evaluation, it significantly helps the model to understand the intrinsic contextual dependency within the dialogues.

B. Framework Architecture

Large pre-trained language models such as [34], [39] have been successfully adapted to a wide range of downstream natural language processing tasks, such as question answering, text classification and text generation. D-score relies on such a pre-trained language model to provide good contextualized word representations that contain rich semantic information. In D-score, RoBERTa [40] is selected as the backbone pre-trained language model, because it is a highly optimized version of the original BERT model [34] and has been pretrained on a much larger English dataset as compared to BERT. Compared to GPT-2 [39], RoBERTa captures bidirectional contextual dependency within text sequences, making it more effective in identifying unnatural and incorrect responses [8]. Here, we denote the RoBERTa encoder as E and feed the two previously defined inputs, I_p and I_s into E .

The two input sequences are first transformed into their corresponding vector representations via an embedding layer, which is to encode their semantic and positional information. Each token, $w_i^p \in I_p$ or $w_j^s \in I_s$ is converted into a vector representation with a shared embedding lookup table $W \in \mathbb{R}^{V \times d}$, $\mathbf{w}_i^p \in \mathbb{R}^d$ or $\mathbf{w}_j^s \in \mathbb{R}^d$ respectively. V denotes the vocabulary size and d is the dimensionality of the embedding vector. In addition, the transformer positional encoding is applied to tokens in I_p and I_s separately. Thus, we have $\mathbf{wp}_i^p \in \mathbb{R}^d$ and $\mathbf{wp}_j^s \in \mathbb{R}^d$ respectively. The two types of embeddings are combined into a single vector via element-wise addition. The embedded vectors are then passed into a transformer encoder. E consists of 12 transformer encoders stacked in layers. Each transformer encoder contains a self-attention layer and a position-wise feed-forward layer as indicated in the RoBERTa encoder of Fig. 1:

1) **Self-Attention Layer:** The self-attention layer enables the capture of global dependency of a sequence via the compression of attentions toward itself. The embeddings of I_p and I_s can be represented by $\mathbf{I}_p \in \mathbb{R}^{(x+y) \times d}$ and $\mathbf{I}_s \in \mathbb{R}^{(y+z) \times d}$ respectively. The outputs of the self-attention layer, $\tilde{\mathbf{I}}_p \in \mathbb{R}^{(x+y) \times d}$ and $\tilde{\mathbf{I}}_s \in \mathbb{R}^{(y+z) \times d}$, are learned via the following equations:

$$\tilde{\mathbf{I}}_p = f_{comb}(f_{mh}(f_{lin}^Q(\mathbf{I}_p), f_{lin}^K(\mathbf{I}_p), f_{lin}^V(\mathbf{I}_p))) \quad (3)$$

$$\tilde{\mathbf{I}}_s = f_{comb}(f_{mh}(f_{lin}^Q(\mathbf{I}_s), f_{lin}^K(\mathbf{I}_s), f_{lin}^V(\mathbf{I}_s))) \quad (4)$$

$f_{lin}^{(\cdot)}$ linearly transform $\mathbf{I}_p$ and $\mathbf{I}_s$ to form h heads with $f_{lin}^{(\cdot)}(\mathbf{I}_p) \in \mathbb{R}^{(x+y) \times h \times d'}$ and $f_{lin}^{(\cdot)}(\mathbf{I}_s) \in \mathbb{R}^{(y+z) \times h \times d'}$ respectively, with $d' = d/h$. f_{mh} performs the following operation on queries and

key-value pairs for each head:

$$f_{mh}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d'}}\right)V \quad (5)$$

Finally, $f_{comb}(\cdot)$ combines the outputs of all h heads into a single vector representation.

2) **Position-Wise Feed-Forward Layer:** The self-attention layer is followed by a fully connected feed-forward layer, which is applied to each position separately and identically. The feed-forward operation [59] on $\tilde{\mathbf{I}}_p$ and $\tilde{\mathbf{I}}_s$ uses GELU activation [60] as follows:

$$FF(\tilde{\mathbf{I}}_p) = GELU(\tilde{\mathbf{I}}_p W_1 + b_1)W_2 + b_2 \quad (6)$$

$$FF(\tilde{\mathbf{I}}_s) = GELU(\tilde{\mathbf{I}}_s W_1 + b_1)W_2 + b_2 \quad (7)$$

where $W_1 \in \mathbb{R}^{(d \times \tilde{d})}$, $b_1 \in \mathbb{R}^{\tilde{d}}$ and $W_2 \in \mathbb{R}^{(\tilde{d} \times d)}$, $b_2 \in \mathbb{R}^d$. $\tilde{d}$ denotes the hidden size of the position-wise feed-forward layer (FF). Both self attention and feed-forward layers are followed by residual connection and layer normalization, which are omitted for brevity.

The activation of the second last transformer encoder layer of E with respect to I_p and I_s are extracted as their respective semantic vector representations. These representations contain rich semantic information. The second last layer is used since we need a more general contextualized representation and the embeddings corresponding to the last layer is more tuned towards the downstream tasks. We denote the outputs of the RoBERTa encoder as $\mathbf{L}_p \in \mathbb{R}^{(x+y) \times d}$ and $\mathbf{L}_s \in \mathbb{R}^{(y+z) \times d}$ as shown in the figure 1. The input and output dimensions of the RoBERTa encoder are the same. $\mathbf{L}_p$ and $\mathbf{L}_s$ can be seen as $[l_1^p, l_2^p, \dots, l_{(x+y)}^p]$ and $[l_1^s, l_2^s, \dots, l_{(y+z)}^s]$ respectively.

To encode the temporal flow at turn-level and token-level that represents the coherence of a dialogue, $\mathbf{L}_p$ and $\mathbf{L}_s$ are further processed by a bidirectional LSTM layer. We have the transitions for I_p and I_s as:

$$\mathbf{h}_i^p = LST\overleftrightarrow{M}(\mathbf{h}_i^p(+, -)1, \mathbf{l}_i^p) \quad (8)$$

$$\mathbf{h}_j^s = LST\overleftrightarrow{M}(\mathbf{h}_j^s(+, -)1, \mathbf{l}_j^s) \quad (9)$$

Next, the corresponding last hidden states, $\mathbf{h}_p \in \mathbb{R}^{d_l}$ and $\mathbf{h}_s \in \mathbb{R}^{d_l}$ of I_p and I_s respectively are used as the final context-level representations. d_l is the hidden state dimension of the bidirectional LSTM.

Finally, $\mathbf{h}_p$ and $\mathbf{h}_s$ pass through the alignment layer to form a single vector representation which is used for various scoring tasks covered in the subsequent subsection. Formally, the alignment takes the following form:

$$\mathbf{h}_{comb} = [\mathbf{h}_p \circ \mathbf{h}_p^T W_{align} \mathbf{h}_s \circ \mathbf{h}_s] \quad (10)$$

where $W_{align} \in \mathbb{R}^{d_l \times d_l}$ is a learned parameter. This concatenation strategy [9] is effective for the determination of the relevance between two context-level representations.

The four tasks in Fig. 1 can be seen as four scorers, each producing a quality indication for u_t from one linguistic aspect. The composition of the four scores forms a D-score for u_t . The

TABLE I
SEMANTIC SAMPLING EXAMPLES FROM PERSONA-CHAT DATASET

Original Preceding Context
Speaker A: Hi, how are you today?
Speaker B: I am great just about to head out fishing one of my favorite pastimes.
Speaker A: Sounds fun! i am going shoe shopping later.
Original Response
Speaker B: <i>On the island that i live on there are many places to shop</i>
Semantic Negative Response
Speaker B: <i>not much, just seen both my children off to school , you?</i>
Generic Negative Response
Speaker B: <i>i don't know what you're talking about.</i>
Original Succeeding Context
Speaker A: I would shop all day! what do you like to do in your spare time?
Speaker B: I love making boats in my spare time on weekends , what about you?
Speaker A: I love to go dancing. I love eating out.

SA, CC and LC scorers share the same RoBERTa encoder, Bi-LSTM layer and alignment layer whereas the LF scorer only shares the RoBERTa encoder. The entire framework is trained end-to-end in a multi-tasking manner.

The training of the D-score framework does not rely on any human annotation. It seeks to capture the underlying representation in human-human dialogues by learning from contrasting data instances. The question is how to construct negative samples that provide relevant information. Next, we discuss the sampling policy.

C. Response Sampling Policy

To have a wide coverage of negative sample distributions, we propose a holistic sampling policy, $P(\tilde{u}_t)$, that considers all the intended evaluation aspects. The contrasting response is defined as $\tilde{u}_t \sim P(\tilde{u}_t)$ where $P(\tilde{u}_t)$ can be grouped into the following categories:

1) **Semantic Sampling Policy, $P_s(\tilde{u}_t)$** : We adopt a widely used approach to randomly select an utterance from another conversation to replace the original response [61]. The crux of this strategy is that an utterance from a different conversation is topically incongruous with the current dialogue context. This can be illustrated with an example in Table I where the two speakers are talking about shopping. The semantic negative response breaks down the dialogue by suddenly moving into the discussion about children, therefore, deemed incongruous for the original dialogue context.

Another strategy used is to detect if a dialogue response has been replaced with an utterance randomly sampled from its context window. This strategy has been applied to negative context training of generative models [62]. Its purpose is to help the evaluation framework to detect repetitive utterances. Since purely detecting a repeated utterance is trivial, we permute the ordering of tokens in a sampled context sentence to alter its syntax and semantic ordering. In addition to repetition detection, generic response detection is also adopted. Many of the

sequence-to-sequence generative models optimized with maximum likelihood estimation suffer from the problem of outputting generic and short responses [63].

Transformer-based chatbots, that are trained on a large amount of dialogue data [38], [64], are capable of generating a diverse range of responses. To evaluate such chatbots, our framework considers the specificity of dialogue responses within their corresponding contexts. Adiwardana *et al.* [38] also stress the importance of specificity by proposing the sensibleness & specificity metric. As shown in the same dialogue example in Table I, even though the generic response from speaker B, ‘i don’t know what you’re talking about,’ does not break the conversational flow, the response is considered dull and not engaging at all. All the above strategies are to reinforce D-score’s capability in evaluating *semantic appropriateness*.

2) **Coherence Sampling Policy, $P_c(\tilde{u}_t)$** : As defined in [44], context coherence refers to whether the response occurs naturally within the context and guides the subsequent sentences. Dialogue is a form of discourse structure with explicit sentence orders as well as underlying conversation turn substructures, such as entity and dialogue act (DA) transitions. Hence, the integration of coherence assessment into automatic dialogue evaluation is indispensable.

One key aspect of coherence assessment is the sentence order discrimination task whereby a document formed by a coherent sequence of sentences is distinguished from the random permutation of those sentences. In D-score, since we examine the quality of each utterance within its respective context window, a swapping strategy is adopted whereby the current response is swapped with an utterance randomly sampled from either its preceding context or succeeding context. The evaluation model is supposed to determine the coherence of the conversational flow by discriminating the original dialogue context window with the one containing swapped utterances. This sampling strategy strengthens D-score’s ability to evaluate along the *context coherence* aspect.

3) **Sampling Policy for Logical Consistency, $P_l(\tilde{u}_t)$** : We cast the problem of evaluating whether the current response maintains logical consistency within the bidirectional context as natural language inference (NLI). However, this classification problem requires NLI labels, which are expensive to obtain. The sampling strategies discussed in the previous two subsections serve as an effective way for automatic NLI labelling. A similar strategy has been adopted in [16] to construct the InferConv dialogue NLI dataset.

First, the original utterance is labelled as entailment since the human-human conversation is assumed to be logically consistent. Next, an utterance chosen randomly from another conversation or a generic response is considered neutral. This is because there is a high probability that such responses are irrelevant to the current conversational flow and they are unlikely to contradict the dialogue context. Furthermore, the utterance randomly sampled from the context with word order shuffled changes the syntax and semantic information which are originally conveyed in the context. Hence, such negative samples are labelled contradiction. Lastly, the effect of swapped strategy on logical consistency is highly reliant on the position of the

sampled utterance. If the sampled utterance is close to the current response, they are labelled entailment since there is no change in the conveyed meaning by purely sentence order swapping of utterances in a local context. If the sampled utterance is far from the current response, most probably there is minimal dependency and information exchange between the sampled utterance and the current dialogue context. It is then labelled neutral.

With this automatic labelling mechanism, we can incorporate an NLI classifier into our D-score framework to measure the *logical consistency* of a dialogue. In addition, the NLI classification task will provide additional inductive bias to the training of our multi-task model.

D. Multiple Scoring Heads

Semantic and coherence scorers are both binary classifiers. Consistency scorer is an NLI classifier. All three classifiers are two-layer MLPs. The syntactic scorer is a causal language model (CLM) head built on top of the shared RoBERTa encoder. The masked self-attention mechanism [59] is applied to prevent the model from attending to future tokens.

1) **Semantic Appropriateness Scorer:** SA scorer seeks to determine *semantic appropriateness* of current response. It learns to distinguish u_t from $\tilde{u}_t \sim P_s(\tilde{u}_t)$ given the bidirectional contexts, c_t^p and c_t^s , i.e. to assign u_t the label **1** and $\tilde{u}_t$ the label **0**. The semantic metric score is defined as $F_{SA}(c_t^p, c_t^s, u_t') = \mathcal{P}(y = 1 | c_t^p, c_t^s, u_t')$ where $u_t' \in (u_t, \tilde{u}_t)$ and $F_{SA}(\cdot) \in [0, 1]$.

2) **Context Coherence Scorer:** This scorer aims to assess whether the current response together with its bidirectional context form a coherent flow in terms of explicit sentence ordering as well as implicit dialogue substructures. Hence, this classifier is to distinguish u_t from $\tilde{u}_t \sim P_c(\tilde{u}_t)$ given the bidirectional contexts. Similar to the semantic scorer, the coherent metric score is defined as $F_{CC}(c_t^p, c_t^s, u_t') = \mathcal{P}(y = 1 | c_t^p, c_t^s, u_t')$ where $u_t' \in (u_t, \tilde{u}_t)$ and $F_{CC}(\cdot) \in [0, 1]$.

3) **Logical Consistency Scorer:** A typical NLI classifier seeks to determine whether the hypothesis can be inferred from a premise. We frame u_t or $\tilde{u}_t \sim P_l(\tilde{u}_t)$ as the hypothesis and (c_t^p, c_t^s) as the premise. The metric score is defined as $F_{LC}(c_t^p, c_t^s, u_t') = \mathcal{P}(y = \text{entailment} | c_t^p, c_t^s, u_t')$ where $u_t' \in (u_t, \tilde{u}_t)$ and $F_{LC}(\cdot) \in [0, 1]$. The higher the entailment score, the more consistent the response is with respect to its context.

4) **Language Fluency Scorer:** To determine whether dialogue responses are fluent, we employ sentence-level log-likelihood adapted from the perplexity score of a language model. Perplexity is defined as the average inverse probability of a sentence over the tokens, and a lower perplexity suggests higher fluency. To relate higher score value to higher fluency, we define a syntactic score as,

$$F_{LF}(u_t) = \frac{1}{y} \sum_t^y \log \mathcal{P}(w_t | w_{<t}). \quad (11)$$

The response log-likelihood is an unbounded negative value. Hence, we apply min-max normalization to all the response scores in a test set to keep them between 0 and 1.

TABLE II
DATASET STATISTICS

DSTC6	training	development	test
#dialog	514,764	52,682	2,000
#turn	1,205,188	125,379	5,266
#word	20,735,566	2,192,418	99,389
#avg turn	2.34	2.38	2.63
DSTC7	training	development	test
#dialog	25,465	4,362	146,229
#turn	128,253	16,854	474,549
#word	1,648,998	224,723	6,900,597
#avg turn	5.03	3.86	3.24
PERSONA-CHAT	training	development	test
#dialog	17,878	1,000	-
#turn	262,626	15,566	-
#word	3,068,672	189,374	-
#avg turn	14.69	15.57	-

E. Evaluation With D-Score

During the evaluation stage, we obtain four aspect scores together with a D-score. This enables the framework to flexibly adapt to specific evaluation requirements. For example, if the assessment process focuses more on the relevance of a response to its corresponding context when judging the system performance, such as the relevance criteria specified in DSTC7 track 2 evaluation process, the metric score provided by the semantic and logical consistency scorers should be used for automatic evaluation accordingly.³ On the other hand, if the evaluation process focuses on a conversation's overall quality, a holistic perspective combining scores from all scorers should be considered. D-score is a weighted sum of four scores,

$$F_{final} = W_{SA}F_{SA}(\cdot) + W_{CC}F_{CC}(\cdot) + W_{LC}F_{LC}(\cdot) + W_{LF}F_{LF}(\cdot) \quad (12)$$

The weights can be either assigned empirically or according to scoring preference. In our experiments, we manually probe into the assessment criteria to determine the appropriate weights for each score. For dialogue-level evaluation, all turn-level scores belonging to the same dialogue are averaged to get an aggregate score. For system-level evaluation, the single scores of all dialogues in the test set are averaged.

IV. EXPERIMENT PRELIMINARIES

A. Datasets

In this work, the D-score framework is trained on three different human-human conversational datasets: DSTC6 customer service dialogue dataset [65], the DSTC7 knowledge-grounding dataset [66] and the PERSONA-CHAT dataset [23]. Then, the performance of D-score is assessed by conducting correlation analysis with respect to the human annotated scores. The statistics of the three training datasets are presented in Table. II

³Details on human evaluation criteria are provided in IV-A.

1) **DSTC6 Dataset:** In DSTC6’s end-to-end conversation modeling track, dialogue data are collected from multiple Twitter accounts for customer service. We followed the data collection procedure provided by the task organizer to collect the dialogue data.⁴ The collected dialogues are between customers and customer support agents on Twitter. Topics are generally about helping resolve customers’ complaints or answering customers’ questions. We only used the training and development set for D-score to learn the human-human conversations. The test set is used to perform dialogue evaluation since the human ratings are provided to generated responses with respect to the corresponding dialogue contexts in the test set.

The DSTC6 evaluation dataset contains generated responses from 20 systems for the 2000 dialogue contexts in the test set. Each generated response is annotated by 10 different Turkers using 5 points Likert Scale. The annotation is based on whether the responses are relevant to the respective dialogue context. This aligns to the *semantic appropriateness* aspect D-score intends to evaluate. For each dialogue context, there are 11 gold reference responses including the original response.

2) **DSTC7 Dataset:** The DSTC7 end-to-end conversational modeling task [66] aims to introduce generative models that go beyond chit-chat by producing more informative responses that are grounded by external knowledge following the framework proposed by Ghazvininejad *et al.* [67]. We followed the standard scripts provided by the task organizers⁵ to extract the official train, development and test data. The original dataset contains about 3 million conversational responses and 20 million facts respectively. Due to limited computation resources, only the development set and a small portion of the training set are used for training the D-score framework. In addition, the current version of D-score framework only considers dialogue contexts. Hence, we omit the information regarding external facts here and will incorporate grounding on external knowledge in our future research.

In the evaluation dataset, there are 1000 dialogue contexts and 10 responses per context. These 10 responses contain hypotheses from 9 different generative models plus the origin human response. For each dialogue response, three crowd-sourced annotators provided scores based on two criteria, relevance and informativeness. The scores for each criterion are based on the 5-point Likert scale. The overall score is obtained by combining the two judgments with equal weights.

3) **PERSONA-CHAT Dataset:** PERSONA-CHAT dataset is a crowd-sourced dataset, collected via Amazon Mechanical Turk, where each of the pair of interlocutors conditioned their conversation on a prior given persona profile, which is provided in advance [23]. The dataset is interesting since it contains more turns per dialogue, and more open-domain than the DSTC datasets. We obtained the dataset from ConvAI2 repository on the Facebook Parlai platform⁶ [68]. The statistics is provided in Table II. Here, we omit the information on persona profiles

since the current version of D-score framework only takes into consideration the dialogue contexts. For correlation analysis, we rely on the Persona-Chatlog human evaluation dataset collected by See *et al.* [15].

The Persona-Chatlog evaluation dataset contains 3316 conversations from 26 model configurations, which include a human agent. The annotation is performed at the conversation level whereby a crowd worker interactively chats with one model configuration for 6 conversational turns. At the end of the conversation, the crowd worker answers eight multiple-choice questions. Each question captures one aspect of conversational quality including avoiding repetition, interestingness, making sense, fluency, listening, inquisitiveness, humanness and engagingness. All the questions use a 1-4 Likert scale, the higher the better⁷ [15]. On average, there are 114 conversations per model configuration and each model configuration has been annotated with over 100 crowd workers.

B. Baselines

We compare D-score against representative reference-based metrics that are widely used, as well as existing state-of-the-art automatic reference-free metrics. For evaluation datasets from DSTC shared tasks, we compare D-score against reference-based baselines adopted by the task organizers. On the DSTC6 evaluation dataset, we also compare D-score against two more recently proposed reference-based metrics, AM-FM [19] and Deep AM-FM [18]. In addition, four state-of-the-art reference-free baselines, BERT NLI (BNLI) [11], contextualized RUBER (Ctr-R) [7], GPT-2 [20] and USR [8], are included. Lastly, for the Persona-Chatlog evaluation dataset, the recently MAUDE metric [69] is included on top of the four reference-free metrics for comparison. Scores provided by all reference-free baselines are normalized to between [0,1] using min-max scaling.

V. EXPERIMENT RESULTS & ANALYSIS

The effectiveness of D-score is assessed by conducting correlation analysis on dialogue evaluation datasets containing human judgements whereby the scores provided by D-score are compared against those given by human judges. Depending on the granularity of evaluation, there are three levels of correlation analysis: system level, conversation level, and turn level. For shared tasks such as DSTC6 and DSTC7, system-level Spearman correlation is particularly important as we would like to rank the systems. For cases where specific behaviours of a dialogue system need to be examined, conversation or turn level evaluation is more appropriate.

A. Correlation Analysis on DSTC6 Dataset

For the DSTC6 evaluation dataset, the human annotators only evaluated the relevance of a response given its preceding utterance, the language fluency scorer seems to be less relevant here. However, the evaluation criterion is closely related to the other

⁴<https://github.com/dialogtekgreek/DSTC6-End-to-End-Conversation-Modeling>.

⁵<https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling>

⁶<https://parl.ai/>

⁷Exceptions: Avoiding repetition is a 1-3 scale and inquisitiveness has an optimal score of 3 whereby 1 and 2 represent asking too little questions and 4 represents asking too much.

TABLE III
SYSTEM LEVEL PEARSON & SPEARMAN CORRELATION ON DSTC6
EVALUATION DATASET. THE HIGHEST CORRELATION SCORES OF METRICS ARE
HIGHLIGHTED IN BOLD

	Semantic Appropriateness (Overall @ DSTC6)	
	Pearson (p-value)	Spearman (p-value)
BLEU-4 [65]	-0.511 (2.14e-02)	-0.188 (4.27e-01)
ROUGE-L [65]	0.145 (5.42e-01)	0.054 (8.21e-01)
METEOR [65]	0.363 (1.16e-01)	0.069 (7.72e-01)
Skip-thought [65]	-0.461 (4.09e-02)	-0.355 (1.25e-01)
Embedding Avg [65]	0.775 (6.07e-05)	0.075 (7.52e-01)
AM-FM [19]	0.891 (< 0.05)	0.442 (> 0.05)
Deep AM-FM [18]	0.907 (< 0.05)	0.516 (< 0.05)
BNLI	0.012 (9.62e-01)	0.457 (4.27e-02)
Ctrl-R	-0.125 (6.00e-01)	0.451 (4.59e-02)
USR	0.144 (5.45e-01)	0.045 (8.50e-01)
GPT-2	-0.638 (2.49e-03)	-0.538 (1.43e-02)
D-score (SA)	0.442 (5.10e-02)	0.753 (1.26e-04)
D-score (LC)	0.439 (5.29e-02)	0.753 (1.26e-04)
D-score (CC)	0.754 (1.21e-04)	0.657 (1.26e-04)
Human	0.985 (< 0.05)	0.913 (< 0.05)

TABLE IV
TURN LEVEL PEARSON & SPEARMAN CORRELATION ON DSTC6 EVALUATION
DATASET. THE HIGHEST CORRELATION SCORES OF METRICS ARE
HIGHLIGHTED IN BOLD. P-VALUE < 0.05 INDICATES STATISTICAL
SIGNIFICANCE

	Semantic Appropriateness (Overall @ DSTC6)	
	Pearson (p-value)	Spearman (p-value)
BNLI	0.350 (0.00e-00)	0.376 (0.00e-00)
Ctrl-R	0.283 (0.00e-00)	0.302 (0.00e-00)
USR	0.221 (0.00e-00)	0.210 (0.00e-00)
GPT-2	-0.039 (8.44e-15)	-0.059 (2.96e-32)
D-score (SA)	0.415 (0.00e-00)	0.465 (0.00e-00)
D-score (LC)	0.419 (0.00e-00)	0.466 (0.00e-00)
D-score (CC)	0.159 (2.70e-224)	0.418 (0.00e-00)
Human	0.421 (< 0.05)	0.476 (< 0.05)

three scorers, namely *semantic appropriateness*, *context coherence* and *logical consistency*. Therefore, we report the respective correlations between the scores of each of the three scorers and the human relevance ratings. Table III and IV present the system-level and turn-level correlations respectively. In terms of system-level correlations, Deep AM-FM achieves the highest Pearson correlation. The reason why deep AM-FM performs exceptionally well in terms of system-level Pearson correlation is because: 1) the responses in the DSTC6 dataset are relatively simple. This phenomenon is unique to the DSTC6 dataset since it is about customer support. Therefore, due to customer privacy concerns, the task-oriented respondents (e.g., customer support) generally avoid having a conversation publicly and usually take the exchanges offline (e.g., by email or Twitter direct message) early in the conversation, well before the conversation has been completed [70]. Hence, responses like “we are sorry about that,

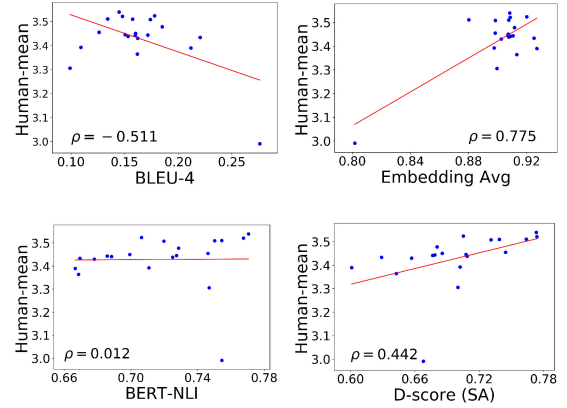


Fig. 2. Scatter plots of system level scores of automatic metrics against human annotations on DSTC6 dataset to investigate the large difference between Pearson and Spearman correlations.

please dm us” frequently occur in the dataset. 2) For each test case, there are 11 references. The presence of an abundant amount of references provides additional semantic information to the evaluation model. Nevertheless, D-score (SA) and D-score (LC) attain the highest Spearman correlation among all the automatic metrics. D-score (CC) attains the highest Pearson correlation among all the reference-free metrics.

In terms of turn level correlations, D-score (SA) and D-score (LC) also achieve close-to-human performance and they outperform all the reference-free metrics. For D-score (CC), the Spearman correlation is comparable to the other two scorers. However, the Pearson correlation is lower. We hypothesize this is because the DSTC6 evaluation data are all single-turn conversations. The responses may be semantically appropriate given the corresponding previous utterances, but not necessarily coherent. The human scores for such responses are high, but the scores provided by the CC scorer are low, resulting in a low Pearson correlation. Overall, the results prove D-score’s ability to provide accurate judgement along a single dialogue aspect.

It can also be noticed that most of the metrics have a large difference between Pearson and Spearman correlations, especially BLEU-4, Embedding Average, BERT-NLI, D-score (SA). This indicates the presence of out-liners in the data. As shown in figure 2, the out-liner is associated with the worst system in all the scatter plots. This suggests the non-uniformity of the human score distribution. The gap between the human score assigned to the worst system and those assigned to the rest is large. This indicates that all human annotators are good at identifying the worst system, but they have trouble evaluating the rest. The same observation is made in [65]. As explained earlier, the customer support related property of the DSTC6 dataset leads to generic and repetitive responses in the training data. Generative dialogue models are very good at producing such responses. Hence, it is noticed that the human scores for the generative systems other than the worst one are all very close in the scatter plots. If we remove the out-liner, the Pearson and Spearman correlations will be very close. For instance, the Pearson and Spearman correlation of D-score (SA) will then become 0.631 and 0.740 respectively. From the scatter plots, it can also be

TABLE V

SYSTEM LEVEL AND TURN LEVEL PEARSON & SPEARMAN CORRELATION ON DSTC7 EVALUATION DATASET. THE HIGHEST CORRELATION SCORES ARE HIGHLIGHTED IN BOLD. THE LAST ROW DENOTES THE MEAN INTER-ANNOTATOR PEARSON AND SPEARMAN CORRELATION, WHICH REPRESENT THE AGREEMENT BETWEEN AN ANNOTATOR AND THE REST

	Semantic Appropriateness (Relevance @ DSTC7)		Logical Consistency (Informativeness @ DSTC7)		Overall (Overall @ DSTC7)	
System Level	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)
BLEU-4	-0.410 (2.40e-01)	-0.304 (3.93e-01)	-0.499 (1.42e-01)	-0.430 (2.14e-01)	-0.457 (1.84e-01)	-0.365 (3.00e-01)
NIST-1	0.482 (1.59e-01)	0.505 (1.37e-01)	0.590 (7.23e-02)	0.636 (4.79e-02)	0.539 (1.08e-01)	0.584 (7.65e-02)
METEOR	0.506 (1.36e-01)	0.541 (1.06e-01)	0.603 (6.50e-02)	0.564 (8.97e-02)	0.557 (9.37e-02)	0.486 (1.54e-01)
BNLI	0.145 (6.90e-01)	-0.146 (6.88e-01)	0.350 (3.20e-01)	0.115 (7.51e-01)	0.244 (4.97e-01)	-0.103 (7.77e-01)
Ctrl-R	-0.313 (3.78e-01)	-0.285 (4.25e-01)	-0.294 (4.10e-01)	-0.213 (5.55e-01)	-0.326 (3.57e-01)	-0.212 (5.56e-01)
GPT-2	-0.332 (3.49e-01)	-0.146 (6.88e-01)	-0.464 (1.77e-01)	-0.273 (4.46e-01)	-0.389 (2.67e-01)	-0.212 (5.56e-01)
BNLI + Ctrl-R	0.356 (3.13e-01)	0.309 (3.85e-01)	0.320 (3.68e-01)	0.231 (5.20e-01)	0.388 (2.68e-01)	0.345 (3.28e-01)
USR	-0.712 (2.08e-02)	-0.754 (1.18e-02)	-0.614 (5.92e-02)	-0.721 (1.86e-02)	-0.672 (3.32e-02)	-0.770 (9.22e-03)
D-score (SA, LC)	0.765 (9.92e-03)	0.875 (9.05e-04)	0.556 (9.50e-02)	0.745 (1.33e-02)	0.674 (3.26e-02)	0.867 (1.17e-03)
Inter-annotator	0.995 (< 0.05)	0.988 (< 0.05)	0.983 (< 0.05)	0.967 (< 0.05)	0.991 (< 0.05)	0.978 (< 0.05)

Turn Level	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)	Pearson (p-value)	Spearman (p-value)
BLEU-4	-0.029 (4.27e-03)	-0.010 (3.13e-01)	-0.065 (1.84e-10)	-0.052 (4.43e-07)	-0.055 (7.54e-08)	-0.040 (1.11e-04)
NIST-1	0.081 (2.76e-15)	0.089 (2.61e-18)	0.097 (2.56e-21)	0.124 (5.84e-34)	0.096 (4.41e-21)	0.115 (3.09e-29)
METEOR	0.077 (5.46e-14)	0.088 (6.70e-18)	0.096 (7.06e-21)	0.112 (4.65e-28)	0.093 (1.29e-19)	0.106 (6.11e-25)
BNLI	0.016 (1.25e-01)	0.022 (3.49e-02)	0.016 (1.10e-01)	0.033 (1.24e-03)	0.016 (1.25e-01)	0.024 (1.74e-02)
Ctrl-R	-0.076 (1.35e-13)	-0.049 (1.84e-06)	-0.117 (2.21e-30)	-0.101 (5.62e-23)	-0.105 (1.65e-24)	-0.085 (1.22e-16)
GPT-2	-0.057 (3.26e-08)	-0.047 (5.51e-06)	-0.126 (5.45e-35)	-0.103 (5.17e-24)	-0.102 (1.30e-23)	-0.084 (3.11e-16)
BNLI + Ctrl-R	0.041 (6.70e-05)	0.062 (1.40e-09)	0.045 (1.07e-05)	0.070 (9.69e-12)	0.044 (1.77e-05)	0.071 (5.58e-12)
USR	-0.312 (8.71e-214)	-0.304 (1.72e-202)	-0.254 (2.70e-139)	-0.245 (6.95e-130)	-0.297 (9.62e-193)	-0.290 (5.44e-183)
D-score (SA, LC)	0.246 (3.64e-131)	0.306 (2.10e-205)	0.182 (8.23e-72)	0.247 (1.42e-132)	0.225 (7.90e-109)	0.289 (1.07e-182)
Inter-annotator	0.254 (< 0.05)	0.250 (< 0.05)	0.215 (< 0.05)	0.210 (< 0.05)	0.258 (< 0.05)	0.251 (< 0.05)

observed that BLEU-4 is a pessimistic metric whereas the other three are optimistic metrics since scores assigned by BLEU-4 are much lower as compared to those assigned by other metrics. This may be because, with diverse human-written references and the generic corresponding system hypotheses, there are few lexical overlaps between them, which lead to low BLEU scores whereas the other three metrics purely focus on semantics rendering them more suitable for dialogue evaluation.

The scatter plot of the embedding average sheds light on why it has a very low Spearman correlation. We can see that all the points are clustered at the top right region except the out-liner. This indicates that the metric has trouble discriminating the performance of these systems since it only tries to compare the semantics information in a response against a list of corresponding references without referring to the context. The semantic information is represented by static word embedding, which faces the limitation in capturing nuances within dialogue responses. On the contrary, D-score learns dialogue-specific contextualized representation and such information is useful in dialogue response quality evaluation.

B. Correlation Analysis on DSTC7 Dataset

In [66], relevance is defined as whether the system response is conversationally appropriate and relevant to the context. Informativeness is defined as the measurement of the degree to which the dialogue response is interesting and informative in the

context of a document provided by the URL. We hypothesize these two criteria best align with the *semantic appropriateness* and *logical consistency* dialogue aspects. The mapping between semantic appropriateness and relevance is obvious while that for logical consistency and informativeness may be less obvious. The logical consistency scorer is mapped to the informativeness criterion, because D-score framework is trained without external knowledge and we cannot directly determine the informativeness of a response based on whether it contains relevant information from external documents. Instead, we consider a response that entails the context as more interesting than the off-topic, neutral or contradictory responses. LC scorer is designed to perform this task. We intend to include a new scorer that can directly evaluate the relevancy of a response w.r.t to external information in future work. For correlation computation w.r.t the overall human score, we linearly combine SA score and LC score with $W_{SA} = W_{LC} = 1$ and the other weights are set to 0.

In Table V, we observe that D-score achieves close-to-human performance and the highest Spearman correlation across all three categories as well as the highest Pearson correlation in the relevance category over all baselines in terms of system-level correlation. Remarkably, D-score outperforms all the best-performing reference-free metrics across all categories by a large margin in terms of both Pearson and Spearman correlations. It can also be observed that word-overlap metrics actually outperforms the strong reference-free metrics, such as BNLI and GPT-2. This seems to contradict the finding in [1]. The better

TABLE VI
ABLATION STUDY OVER DSTC7 DATASET FOR RELEVANCE CRITERION.
*DENOTES P-VALUE < 0.05 .

W_{SA}	✓	✓	✓	✓
W_{LC}	✓	✓	✓	✓
W_{LF}		✓		✓
W_{CC}			✓	✓
Pearson	0.758*	0.500	0.703*	0.507
Spearman	0.912*	0.632*	0.815*	0.632*

results of the word-overlap metrics could be attributed to the presence of multiple references for each test case in the DSTC7 evaluation data. It has been shown in the previous work [70] that using multiple references actually improves the correlation of existing word-overlap metrics.

In addition, we observe that, at the turn level, the performance of D-score is close to average human performance in terms of Pearson correlation. As for the Spearman correlation, it consistently outperforms the average human performance across all categories. The results suggest that D-score not only excels in providing explainable judging for individual dialogue aspects but also in providing a holistic overview. USR performs exceptionally well in this case. This may be due to that we finetune the pre-trained USR models instead of training them from scratch. The domain of the pre-trained model is topical-chat, which is very similar to that of the DSTC7 dataset.

Furthermore, correlation results between a linearly combined score with equal weights assigned to BNLI and Ctr-R and human judgements help validate two statements: 1) Since Ctr-R is optimized to measure *semantic appropriateness* while BNLI is optimized to evaluate *logical consistency*, the fusion of scores from both models is supposed to yield better performance across all three evaluation categories. 2) A holistic multi-task model is more effective than the fusion of different models with each optimized towards evaluating a single dialogue aspect. These two statements are proved by the correlation results in Table. V. For example, for the overall category, both Spearman & Pearson correlations of BNLI + Ctr-R are higher than those of either BNLI or Ctr-R, but they are lower than respective correlations of D-score.

To further justify our score weight choice, we include an ablation study in Table VI to showcase that incorporating irrelevant scores into the score computation hurts the effectiveness of evaluation. It can be seen from the table that both correlations drop when incorporating scores from the other two aspects. Another point to note is that for training D-score on the DSTC7 dataset, we only consider the immediate context of the current response (set $k = 1$). This is to align with the description of the human evaluation process in [66] whereby the human judges are asked to consider only the immediately preceding utterance to reduce cognitive load.

C. Correlation Analysis on Persona-Chatlog

D-score’s strong ability to capture the dependency between the current response and its surrounding context is even more

advantageous in this case since the conversations used for training contain more turns compared to those for DSTC6 and DSTC7 evaluation. Its multitasking setup helps evaluate both the overall dialogue quality as well as individual dialogue aspects. We perform a correlation analysis on Persona-Chatlog to demonstrate the effectiveness of the D-score in evaluating multi-turn dialogues. The human evaluation process associated with the Persona-Chatlog evaluation dataset is different from that of the rest as explained in IV-A. Therefore, for system-level correlation analysis, we follow [15] and [69] to perform Bayesian calibration [71] to estimate the posterior mean of human ratings associated with each model based on the observed human ratings.

Many of the evaluation aspects in Persona-Chatlog, such as interestingness and humanness, actually correlate with all the four linguistic aspects that D-score is designed to evaluate. For simplicity purpose, we assume all four scorers contribute equally to the evaluation and thus, set their weights to be [$W_{SA} = W_{CC} = W_{LC} = W_{LF} = 1$]. The dialogue level and system level Spearman correlations⁸ are computed along each of the dialogue evaluation aspects as demonstrated in Table VII. It can be observed that D-score outperforms other metrics in terms of both system and conversation level Spearman correlation for most of the dialogue aspects. Especially at the system level, D-score performs significantly better.

Furthermore, to demonstrate the advantage of multi-task learning over a linear combination of individual metrics, we compare D-score against USR, a state-of-the-art reference-free metric as well as a fusion method, which is a linear combination of Ctr-R, BNLI and GPT-2. It can be observed that D-score outperforms these two metrics across all the evaluation categories at both the system and conversation level. USR consists of two sub-components, USR-MLM and USR-DR. USR-MLM is designed to evaluate aspects such as understandability and fluency while USR-DR is suitable for evaluating making sense and listening. Similarly, for the fusion method, each of the three metrics is optimized to evaluate a specific dialogue aspect with Ctr-R focusing on *semantic appropriateness*, BNLI focusing on *logical consistency* and GPT-2 focusing on *language fluency*. This highlights the advantage of learning a shared dialogue knowledge representation for holistic evaluation, over the fusion of individual models, each having a narrow optimization focus. The fusion approach tries to reduce the variance of the scores by averaging multiple scores from independent models to gain better generalization. This is usually not as effective as directly learning a shared representation via multi-task learning since the knowledge contained in one task can be effectively leveraged by other tasks to make a well-informed decision.

To understand how D-score framework scores individual aspects, we compute the pairwise Pearson correlation among the four aspect scores as shown in Fig. 3. The plots suggest that there exist a high level of Pearson correlation for LC-SA, CC-SA, and LC-CC pairs, and a low level of correlation for LF-SA, LF-CC, LF-LC pairs. The high correlation suggests that the scorers are not all mutually exclusive. SA and CC may indeed

⁸Due to page limitations, only Spearman Correlation is reported here.

TABLE VII

COMPARISON OF BOTH SYSTEM AND DIALOGUE LEVEL SPEARMAN CORRELATIONS AMONG STATE-OF-THE-ART REFERENCE-FREE METRICS ON PERSONA-CHATLOG EVALUATION DATASET. THE RESULTS ARE REPORTED FOR THE EIGHT UNIQUE DIALOGUE CATEGORIES. P-VALUES ARE INCLUDED IN THE BRACKETS. THE BEST SCORE FOR EACH CATEGORY IS HIGHLIGHTED IN BOLD. THE CORRELATION RESULTS OF MAUDE IS ADAPTED FROM [69]

Conversation Level Spearman Correlation							
Dialogue Aspects	Ctr-R	BNLI	GPT-2	MAUDE [69]	USR	Fusion	D-score
Avoid Repetition	0.032 (6.53e-02)	-0.069 (7.06e-05)	0.035 (4.57e-02)	0.142 (1.98e-16)	0.055 (1.66e-03)	0.050 (3.98e-03)	0.131 (3.56e-14)
Enjoyment	0.101 (5.32e-09)	-0.109 (3.28e-10)	0.034 (4.73e-02)	0.134 (7.51e-15)	0.075 (1.69e-05)	0.088 (3.83e-07)	0.168 (2.19e-22)
Fluency	0.154 (4.89e-19)	-0.170 (8.35e-23)	0.116 (2.03e-11)	0.100 (8.45e-09)	0.167 (4.14e-22)	0.176 (1.46e-24)	0.237 (1.77e-43)
Inquisitiveness	0.178 (5.30e-25)	-0.165 (1.26e-21)	0.049 (5.08e-03)	0.033 (5.74e-02)	0.086 (8.23e-07)	0.161 (8.56e-21)	0.170 (5.90e-23)
Interestingness	0.076 (1.14e-05)	-0.077 (1.05e-05)	-0.018 (3.02e-01)	0.163 (4.28e-21)	0.045 (9.57e-03)	0.040 (2.14e-02)	0.121 (3.37e-12)
Listening	0.154 (6.27e-19)	-0.166 (7.81e-06)	0.058 (7.86e-04)	0.144 (7.66e-17)	0.167 (4.43e-22)	0.149 (6.34e-18)	0.234 (1.59e-42)
Making-sense	0.211 (1.03e-34)	-0.217 (1.26e-36)	0.146 (2.60e-17)	0.172 (2.11e-23)	0.164 (2.45e-21)	0.240 (8.49e-45)	0.273 (7.20e-58)
Turing	0.083 (1.87e-06)	-0.090 (2.22e-07)	0.070 (6.13e-05)	0.086 (7.70e-07)	0.084 (1.13e-06)	0.097 (1.99e-08)	0.161 (1.33e-20)
Average	0.124 (< 0.05)	-0.133 (< 0.05)	0.061 (< 0.05)	0.122 (< 0.05)	0.105 (< 0.05)	0.125 (< 0.05)	0.187 (< 0.05)
System Level Spearman Correlation							
Avoid Repetition	-0.041 (8.31e-01)	-0.024 (9.03e-01)	0.362 (5.32e-02)	-0.035 (8.55e-01)	-0.047 (8.08e-01)	0.154 (4.25e-01)	0.401 (3.13e-02)
Enjoyment	0.149 (4.41e-01)	-0.232 (2.25e-01)	0.054 (7.80e-01)	0.232 (2.27e-01)	0.264 (1.66e-01)	0.079 (6.84e-01)	0.527 (3.30e-03)
Fluency	0.476 (9.07e-03)	-0.502 (5.53e-03)	0.331 (7.99e-02)	0.370 (4.82e-02)	0.477 (8.83e-03)	0.448 (1.49e-02)	0.790 (3.53e-07)
Inquisitiveness	0.698 (2.56e-05)	-0.769 (1.13e-06)	0.039 (8.39e-01)	0.728 (7.74e-06)	0.722 (9.75e-06)	0.464 (1.12e-02)	0.713 (1.43e-05)
Interestingness	0.113 (5.58e-01)	-0.247 (1.97e-01)	-0.008 (9.66e-01)	0.240 (2.09e-01)	0.181 (3.47e-01)	0.010 (9.60e-01)	0.485 (7.63e-03)
Listening	0.187 (3.32e-01)	-0.214 (2.65e-01)	0.182 (3.45e-01)	0.112 (5.64e-01)	0.209 (2.76e-01)	0.189 (3.26e-01)	0.524 (3.52e-03)
Making-sense	0.361 (5.43e-02)	-0.375 (4.48e-02)	0.315 (9.63e-02)	0.208 (2.79e-01)	0.359 (5.61e-02)	0.374 (4.54e-02)	0.637 (2.03e-04)
Turing	0.061 (7.53e-01)	-0.086 (6.57e-01)	0.305 (1.07e-01)	0.095 (6.26e-01)	0.066 (7.34e-01)	0.149 (4.40e-01)	0.424 (2.18e-02)
Average	0.251 (> 0.05)	-0.306 (> 0.05)	0.198 (> 0.05)	0.244 (> 0.05)	0.279 (> 0.05)	0.233 (> 0.05)	0.563 (< 0.05)

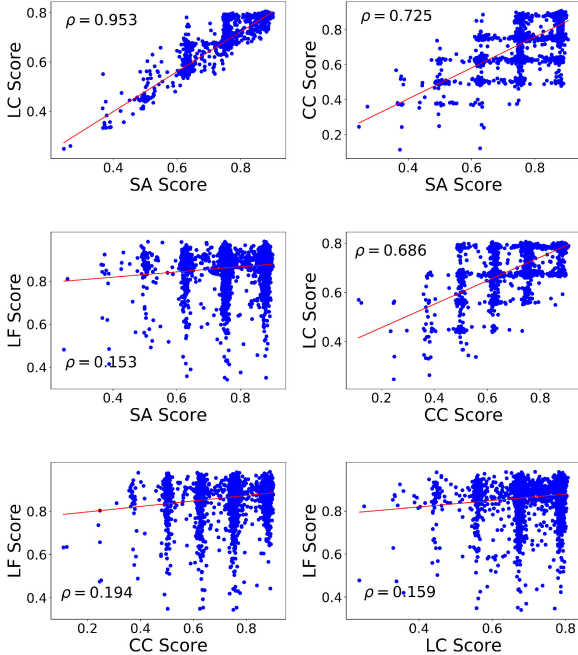


Fig. 3. Scatter plots of pairwise Pearson correlation among scores provided by the four scorers on Persona-Chatlog dataset.

be two closely related concepts and one can infer SA from CC. However, from the perspective of multi-task learning (MTL), the tasks are related, but not necessarily required to be all mutually exclusive [14]. Moreover, D-score explicitly models these two scorers to look into different aspects based on different sampling strategies. SA focuses more on the turn-level context relevancy of the generated response while CC pertains to the ordering of a span of utterances. The extremely high correlation between the LC-SA pair is due to the overlap in their sampling strategies.

The random utterance replacement strategy described in III-C1 plays an important role in forming negative samples for SA and LC scorers correspondingly.

On the other hand, the low correlations of LF-SA, LF-CC, LF-LC pairs indicate that D-score is capable of performing effective evaluation along part of the individual aspects independently, despite that the scorers are built on top of a shared knowledge space. To reinforce this point, we also compute the turn level correlation between human scores and CC, LF and SA scores respectively.⁹ It is found that: 1) LF-Human correlates the best along with the fluency criterion with a Pearson of 0.356 and a Spearman of 0.174. 2) CC-Human correlates the best along listening with a Pearson of 0.247 and a Spearman of 0.251. 3) SA-Human correlates the best with making sense with a Pearson of 0.231 and a Spearman of 0.227.

Lastly, the correlation results on all three datasets suggest that turn level evaluation is much more challenging than system-level evaluation. Even for human judges, it is difficult to agree upon the quality of individual utterances. A similar observation has been made in [27] where a Pearson correlation of 0.954 is reported at the system level on their self-collected evaluation dataset. However, at the utterance level, their ADEM framework only achieves a Pearson correlation of 0.41. This suggests an important future research direction to analyze the more fine-grained factors that contribute to turn level evaluation and study the dynamics within a conversation.

VI. CONCLUSION

D-score serves as an important step into the exploration of holistic dialogue evaluation via a single model setting. Extensive empirical results have proven the generalization capability

⁹Since we are analyzing individual dialogue attributes, turn-level correlation is more appropriate.

of D-score in mimicking human evaluation. D-score is trying to serve the role of an expert judge containing knowledge that is well-suited for all-rounded dialogue evaluation. The self-supervised paradigm enables D-score to leverage contextual information and intrinsic dialogue transition pattern within human-human conversations for quality assessment of dialogue responses within human-machine conversations.

REFERENCES

- [1] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2122–2132.
- [2] J. Deriu *et al.*, "Survey on evaluation methods for dialogue systems," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 755–810, 2021.
- [3] T. Hashimoto, H. Zhang, and P. Liang, "Unifying human and statistical evaluation for natural language generation," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1 (Long and Short Papers), 2019, pp. 1689–1701.
- [4] S. E. Finch and J. D. Choi, "Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols," in *Proc. 21st Ann. Meeting Special Interest Group Discourse Dialogue*, 2020, pp. 236–245.
- [5] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. 55th Ann. Meeting Assoc. Comput. Linguistics*, (Volume 1: Long Papers), Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 654–664.
- [6] P. Gupta, S. Mehri, T. Zhao, A. Pavel, M. Eskenazi, and J. Bigham, "Investigating evaluation of open-domain dialogue systems with human generated multiple references," in *Proc. 20th Annu. SIGDIAL Meeting Discourse Dialogue*, Stockholm, Sweden: Association for Computational Linguistics, Sep. 2019, pp. 379–391.
- [7] S. Ghazarian, J. Wei, A. Galstyan, and N. Peng, "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings," in *Proc. Workshop Methods Optim. Evaluating Neural Lang. Gener.*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 82–89.
- [8] S. Mehri and M. Eskenazi, "USR: An unsupervised and reference free evaluation metric for dialog generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 681–707.
- [9] C. Tao, L. Mou, D. Zhao, and R. Yan, "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, pp. 722–729, 2018.
- [10] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau, "Learning an unreferenced metric for online dialogue evaluation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 2430–2441.
- [11] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3731–3741.
- [12] S. Ghazarian, R. Weischedel, A. Galstyan, and N. Peng, "Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 7789–7796, 2020.
- [13] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," 2017.
- [14] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
- [15] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1702–1723.
- [16] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Evaluating Coherence in Dialogue Systems Using Entailment," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3806–3812.
- [17] A. Venkatesh *et al.*, "On evaluating and comparing open domain dialog systems," 2018, *arXiv:1801.03625*.
- [18] C. Zhang, L. F. D'Haro, R. E. Banchs, T. Friedrichs, and H. Li, "Deep AM-FM: Toolkit for automatic dialogue evaluation," in *Conversational Dialogue Syst. Decade*, Singapore, Springer, 2020, pp. 53–69, doi: [10.1007/978-981-15-8395-7_5](https://doi.org/10.1007/978-981-15-8395-7_5).
- [19] L. F. D'Haro, R. E. Banchs, C. Hori, and H. Li, "Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics," *Comput. Speech Lang.*, vol. 55, pp. 200–215, 2019, doi: [10.1016/j.csl.2018.12.004](https://doi.org/10.1016/j.csl.2018.12.004).
- [20] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 3619–3629.
- [21] P. Xu *et al.*, "A cross-domain transferable neural coherence model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 678–687.
- [22] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proc. 54th Ann. Meeting Assoc. Comput. Linguistics* (Volume 1: Long Papers), 2016, pp. 994–1003.
- [23] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," in *Proc. 56th Annual Meeting Assoc. Comput. Linguistics* (Volume 1: Long Papers), 2018, pp. 2204–2213.
- [24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.
- [25] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Statist. Mach. Transl.*, Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 376–380.
- [26] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.
- [27] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," in *Proc. 55th Ann. Meeting Assoc. Comput. Linguistics*, (Volume 1: Long Papers), Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1116–1126.
- [28] E. Dinan *et al.*, "The second conversational intelligence challenge (ConvAI2)," *The Springer Ser. Challenges Mach. Learn.*, pp. 187–208, Nov. 2019.
- [29] V. Rus and M. Lintean, "An optimal assessment of natural language student input using word-to-word similarity metrics," in *Int. Conf. Intell. Tutoring Syst.*, Springer, 2012, pp. 675–676.
- [30] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *Proc. ACL-08: HLT*, 2008, pp. 236–244.
- [31] G. Forgues, J. Pineau, J.-M. Larchevêque, and R. Tremblay, "Bootstrapping dialog systems with word embeddings," in *Nips, Modern Mach. Learn. Natural Lang. Process. Workshop*, vol. 2, 2014.
- [32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process.* (EMNLP), 2014, pp. 1532–1543.
- [33] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017, doi: [10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051).
- [34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *8th Int. Conf. Learn. Represent.*, 2020.
- [36] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892.
- [37] M. Li, J. Weston, and S. Roller, "Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons," 2019, *arXiv:1909.03087*.
- [38] D. Adiwardana *et al.*, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.
- [39] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, 1(8), p. 9, 2019.

- [40] Y. Liu *et al.*, “A robustly optimized bert pretraining approach,” 2019, *arXiv:1907.11692*.
- [41] A. Halliday Michael and H. Ruqaiya, “Cohesion in English,” vol. 9, 2014.
- [42] P. W. Foltz, W. Kintsch, and T. K. Landauer, “The measurement of textual coherence with latent semantic analysis,” *Discourse Processes*, vol. 25, no. 2-3, pp. 285–307, 1998.
- [43] J. Li, R. Li, and E. Hovy, “Recursive deep models for discourse parsing,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 2061–2069.
- [44] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, “Incremental transformer with deliberation decoder for document grounded conversations,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 12–21.
- [45] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter,” 2019, *arXiv:1910.01108*.
- [46] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced LSTM for Natural Language Inference,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1657–1668.
- [47] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark: Association for Computational Linguistics, Sep., 2017, pp. 670–680.
- [48] A. Ghandeharioun *et al.*, “Approximating interactive human evaluation with self-play for open-domain dialog systems,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13658–13669.
- [49] R. E. Banchs, L. F. D’Haro, and H. Li, “Adequacy-fluency metrics: Evaluating MT in the continuous space model framework,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 3, pp. 472–482, Mar. 2015.
- [50] T. Evgeniou and M. Pontil, “Regularized multi-task learning,” in *Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, ser. KDD ’04*, New York, NY, USA: Association for Computing Machinery, 2004, p. 109–117.
- [51] R. K. Ando and T. Zhang, “A framework for learning predictive structures from multiple tasks and unlabeled data,” *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Dec. 2005.
- [52] A. Kumar and H. Daume III, “Learning task grouping and overlap in multi-task learning,” 2012, pp. 1723–1730.
- [53] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-y. Wang, “Representation learning using multi-task deep neural networks for semantic classification and information retrieval,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Denver, Colorado: Association for Computational Linguistics, May-Jun. 2015, pp. 912–921.
- [54] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, “Multi-task sequence to sequence learning,” 2015, *arXiv:1511.06114*.
- [55] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite BERT for self-supervised learning of language representations,” in *8th Int. Conf. Learn. Representations*, 2020.
- [56] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.
- [57] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proc. 54th Annual Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.
- [58] M. Peters *et al.*, “Deep Contextualized Word Representations,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1 (Long Papers), New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.
- [59] A. Vaswani *et al.*, “Attention is all you need,” in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [60] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” 2016, *arXiv:1606.08415*.
- [61] S. Sato, R. Akama, H. Ouchi, J. Suzuki, and K. Inui, “Evaluating dialogue generation systems via response selection,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Assoc. Comput. Linguistics, Jul. 2020, pp. 593–599.
- [62] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” in *8th Int. Conf. Learn. Representations*, 2020.
- [63] O. Vinyals and Q. Le, “A neural conversational model,” 2015, *arXiv:1506.05869*.
- [64] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, “Can you put it all together: Evaluating conversational agents’ ability to blend skills,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online: Assoc. Comput. Linguistics, Jul. 2020, pp. 2021–2030.
- [65] C. Hori *et al.*, “Overview of the sixth dialog system technology challenge: DSTC6,” *Comput. Speech Lang.*, vol. 55, pp. 1–25, 2019, doi: [10.1016/j.csl.2018.09.004](https://doi.org/10.1016/j.csl.2018.09.004).
- [66] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, “Grounded response generation task at DSTC7,” in *Proc. AAAI Dialog Syst. Tech. Challenges Workshop*, 2019.
- [67] M. Ghazvininejad *et al.*, “A knowledge-grounded neural conversation model,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, pp. 5110–5117, 2018.
- [68] A. H. Miller *et al.*, “Parlai: A dialog research software platform,” in *Proc. Conf. Empirical Methods Nat. Lang. Process.: Syst. Demonstrations*, 2017, pp. 79–84, 2017.
- [69] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau, “Learning an unreferenced metric for online dialogue evaluation,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2430–2441.
- [70] P. Gupta, S. Mehri, T. Zhao, A. Pavel, M. Eskenazi, and J. P. Bigham, “Investigating evaluation of open-domain dialogue systems with human generated multiple references,” in *Proc. 20th Annu. SIGDIAL Meeting Discourse Dialogue*, 2019, pp. 379–391.
- [71] I. Kulikov, A. Miller, K. Cho, and J. Weston, “Importance of search and evaluation strategies in neural dialogue modeling,” in *Proc. 12th Int. Conf. Natural Lang. Gener.*, Tokyo, Japan: Association for Computational Linguistics, Oct.-Nov. 2019, pp. 76–87.



Chen Zhang (Student Member, IEEE) is currently a third year Ph.D. candidate with Electrical & Computer Engineering Department of National University of Singapore (NUS) under the Bosch-NUS Industrial Ph.D. Program. His research interests include automatic dialogue evaluation, machine learning and natural language processing.



Grandee Lee (Student Member, IEEE) received the B.Eng. (Highest Distinction) degree in electrical engineering from the National University of Singapore (NUS), Singapore in 2017. He is currently a research scholar working towards the Ph.D. degree with NUS. His current research interests include cross-lingual language modeling and representation, and dialogue system evaluation.



Luis Fernando D'Haro (Member, IEEE) is an Associate Professor with the Universidad Politécnica de Madrid (ETSIT, UPM) in Spain and a member of the Speech Technology Group. His current research mainly focuses on spoken dialogue and natural language processing systems, he has written more than 15 international journal publications specifically on dialog systems. He Co-Led the International Dialog State Tracking Challenges (DSTC) in 2015 and 2016 and is a member of the program and steering committees from DSTC6 to DSTC9 challenges. Since 2015,

Prof. D'Haro has Co-Organized the WoChat and DBDC series of workshops, which have the common goal of advancing chatbot systems and their automatic evaluation. He was also a Member of the local organizers for Interspeech in 2014, Human Agent Interaction conference (HAI2016), and the International Workshop on Spoken Dialog System Technology (IWSDS) in 2018, and General Chair for IWSDS 2020. He is Senior Member for the Chanel workshop at the Johns Hopkins Summer school (JSALT2020).



Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degree in electrical and electronic engineering from the South China University of Technology, Guangzhou, China in 1984, 1987, and 1990, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager at the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003–2016). He was the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of Computer Speech and Language (2012–2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014 and ASRU 2019. Dr. Li is a Fellow of the ISCA. He was the recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.