

# FlowEval: A Consensus-Based Dialogue Evaluation Framework Using Segment Act Flows

Jianqiao Zhao<sup>1\*</sup>, Yanyang Li<sup>1\*</sup>, Wanyu Du<sup>3\*</sup>, Yangfeng Ji<sup>3</sup>,  
Dong Yu<sup>4</sup>, Michael R. Lyu<sup>1</sup>, Liwei Wang<sup>1,2†</sup>

<sup>1</sup>Department of Computer Science and Engineering, The Chinese University of Hong Kong

<sup>2</sup>Shanghai AI Laboratory <sup>3</sup>Department of Computer Science, University of Virginia

<sup>4</sup>Tencent AI Lab, Bellevue

{jqzhao, yyli21, lyu, lwwang}@cse.cuhk.edu.hk

## Abstract

Despite recent progress in open-domain dialogue evaluation, how to develop automatic metrics remains an open problem. We explore the potential of dialogue evaluation featuring dialog act information, which was hardly explicitly modeled in previous methods. However, defined at the utterance level in general, dialog act is of coarse granularity, as an utterance can contain multiple segments possessing different functions. Hence, we propose *segment act*, an extension of dialog act from utterance level to segment level, and crowdsource a large-scale dataset for it. To utilize *segment act flows*, sequences of segment acts, for evaluation, we develop the first consensus-based dialogue evaluation framework, *FlowEval*. This framework provides a reference-free approach for dialog evaluation by finding pseudo-references. Extensive experiments against strong baselines on three benchmark datasets demonstrate the effectiveness and other desirable characteristics of our *FlowEval*, pointing out a potential path for better dialogue evaluation.

## 1 Introduction

Dialogue evaluation plays a crucial role in the recent advancement of dialogue research. While human evaluation is often considered as a universal and reliable method by the community (Smith et al., 2022), automatic dialogue evaluation metrics draw growing attention as they can assess dialogues with faster speed and lower cost (Tao et al., 2018; Huang et al., 2020; Mehri and Eskénazi, 2020).

Traditional word-overlap metrics, like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), lose some of their effectiveness in the dialogue setting as reliable references are hard to obtain (Liu et al., 2016). Recent works tackle this problem by leveraging more sophisticated architectures (Zhang et al., 2021; Li et al.,

\*Equal contributions. Wanyu participated in building the dataset, while doing her internship with Prof. Liwei Wang.

† Corresponding author.

Dialogue Section	Segment Act Flow
Speaker1: How are you? May I have a cup of coffee?	greeting, directive
Speaker2: Hmm. Certainly. What kind of coffee do you like? We have espresso and latte.	backchannel- success, commissive, question, inform

Table 1: A snippet of an open-domain dialogue and its segment act flow. Each segment is marked with the same color as its corresponding segment act label.

2021) and harnessing the power of large models (Mehri and Eskénazi, 2020). Although these recent metrics claim to show some progress towards higher correlation with humans, the gap between automatic metrics and human evaluation is still noticeable (Yeh et al., 2021). Automatic open-domain dialogue evaluation is still an open question, and extensive efforts have been made to improve performance from different angles (Pang et al., 2020; Ghazarian et al., 2019; Mehri and Eskenazi, 2020; Phy et al., 2020).

Among those newly released metrics (Zhang et al., 2021; Mehri and Eskénazi, 2020; Li et al., 2021; Pang et al., 2020; Tao et al., 2018; Mehri and Eskenazi, 2020; Phy et al., 2020), hardly any explicitly employs dialog act, one of the pillars of dialogue study, in their methods. Intuitively, introducing dialog act into open-domain dialogue evaluation should be beneficial: a sequence of dialog acts distills the core function of each utterance and can potentially reveal how speakers interact in general. However, directly using preexisting dialog act definitions (Stolcke et al., 2000; Hemphill et al., 1990) seems undesirable, as an utterance can contain several segments that possess different conversational functions. We show our observation in Table 1. By saying “*Hmm. Certainly. What kind of coffee do you like? We have espresso and latte.*”,

the participant first acknowledges the conversation with backchannel, then commits to finish the request by saying “*Certainly*”. Later, the speaker asks for a more concrete order and offers all the options. In human conversations, it is common to have more than one function for each utterance, which means using a single dialog act to express the core function of an utterance inevitably suffers from information loss. To solve this issue, we extend the concept of dialog act from the utterance level to the segment level. We name these segment level dialog act *segment act* and a sequence of segment acts a *segment act flow*.

One difficulty of using segment act for open-domain dialogue evaluation is the lack of related data. Since there is no dataset for segment act, we follow the ISO 24617-2 annotation criteria (Bunt et al., 2019) and propose a simplified ISO-format segment act tagset. We crowdsource large-scale segment act annotations on two popular open-domain dialogue datasets: ConvAI2 (Dinan et al., 2019) and DailyDialog (Li et al., 2017). We name our dataset *ActDial*.

Another challenge of incorporating segment act into open-domain dialogue evaluation lies in finding a suitable way to assess dialogues with the segment act feature. Modeling segment act flow is not trivial. On the one hand, dialogues have different numbers of turns and, thus, have varying lengths of segment act sequences. On the other hand, defining and finding the ground-truth segment act flow for a dialogue are almost infeasible, discouraging the development of any reference-based methods. To overcome this challenge, we design the first consensus-based reference-free open-domain dialogue evaluation framework, *FlowEval*.

For a dialogue to be evaluated, our *FlowEval* first obtains the segment act flow, e.g., from a trained classifier. Then, we harvest segment act features, from a dedicated BERT-like (Devlin et al., 2019) masked segment act model, and content features, from RoBERTa-large (Liu et al., 2019). We retrieve pseudo-references from the training set, according to the segment act features as well as content features. Last, we evaluate the dialogue with the consensus of the pseudo-references, fusing metrics from both segment act and word-overlap perspectives. The essence of our consensus-based framework lies in retrieving pseudo-references and using the consensus of pseudo-references to assess a new dialogue. This process can be regarded as

reference-free, since no additional dialogue evaluation label is required. Not limited to segment act feature, our proposed consensus-based framework is compatible to a wide range of features and metrics, such as sentiment features, engagingness feature, etc.

Extensive experiments are carried out against the state-of-the-art baselines on Controllable Dialogue dataset (See et al., 2019), FED dataset (Mehri and Eskénazi, 2020), and DSTC9 dataset (Gunasekara et al., 2020). The result supports that segment act flow is effective in dialogue evaluation: our consensus-based method achieve the best or comparable correlation with human evaluation. Additionally, segment act flow can bring complementary information to metrics that heavily focus on the raw text of dialogues.

In summary, the contributions of this work are three-fold:

1. We propose to model the segment level act as the dialog flow information for open-domain dialogue evaluation.
2. We are the first to propose a consensus-based framework for open-domain dialogue evaluation. Our studies show that the consensus approach can work efficiently even when the size of the search set, i.e., the number of dialogues in the training set, is around ten thousand. This attainable size shows the promise of our consensus approach for dialogue evaluation and other natural language evaluation tasks.
3. Our method can reach the best or comparable performance when compared with state-of-the-art baselines. Additional experiments are conducted to examine detailed properties of our method and consensus process.

We will release all code and data once the paper is made public.

## 2 Related Works

### 2.1 Automatic Dialog Evaluation Metrics

RUBER (Tao et al., 2018) combines a reference-based metric and a reference-free metric where the reference-free metric is learned by an RNN-based model to judge if a response is appropriate for the dialogue history. GRADE (Huang et al., 2020) adopts the graph structure to represent dialogue topics and enhances the utterance-level contextualized representations with topic-level graph representations to better evaluate the coherence of

dialogues. DynaEval (Zhang et al., 2021) reaches the highest correlation with human evaluation on FED dataset (Mehri and Eskénazi, 2020), by utilizing the Graph Convolutional Network to capture the dependency between dialogue utterances. FED (Mehri and Eskénazi, 2020) measures 18 different qualities of dialogues by computing the likelihood of DialoGPT (Zhang et al., 2019) generating corresponding handwritten follow-up utterances. In addition to commonly-used dialogue context, Flow score (Li et al., 2021) takes into account the semantic influence brought by each utterance, which is defined to be the difference between the dense representations of two adjacent dialogue histories. Flow score employs special tasks to promote a better modeling of semantic influence during pretraining. It correlates best with human evaluation on the DSTC9 dataset (Gunasekara et al., 2020).

Different from Flow score and other related works, our method explicitly models the segment acts of a dialog, which deliver clear and interpretable functions for each utterance segment, rather than dense representations.

## 2.2 Dialog Act in Dialogue Systems

Dialog act (Stolcke et al., 2000; Shriberg et al., 2004) and similar concepts, like intent (Hemphill et al., 1990; Larson et al., 2019), have been widely studied in the past decades. Walker and Passonneau (2001) construct a dialog act tagging scheme and evaluate travel planning systems (Walker et al., 2001) based on standalone dialog acts, rather than dialog act sequences. This tagging scheme, while provides more detailed information compared with previous works, only focuses on the system side in a task-oriented setting and may need major modifications when applied to open-domain dialogues. After the initial flourish, recent works come with their own purposes and tagsets for dialog act, tailored for different scenarios or special needs (Budzianowski et al., 2018; Yu and Yu, 2019; Cervone and Riccardi, 2020).

In this work, we propose segment act, an extension of dialog act to the utterance segment level, and design its corresponding tagset. Our segment-focused arrangement can not only cover the diverse scenarios of open-domain dialogues, but also provide finer-grained information for dialogue evaluation than prevailing dialog act designs.

## 2.3 Consensus-Based Methods

Consensus-based methods have been adopted in image captioning (Devlin et al., 2015; Wang et al., 2017; Deshpande et al., 2019) and evaluation (Vedantam et al., 2014; dos Santos et al., 2021).

Devlin et al. (2015) retrieve nearest neighbors in the sense of image feature and use them as consensus. They later take the caption that has the best word overlap with the consensus as the generation results and achieve competitive performance against other caption generation techniques. Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2014) measures the similarity of a generated caption against a set of human-written sentences using a consensus-based protocol. Our proposed method shares similar element as it also involves evaluating by the closeness to a consensus of human sentences. However, to the best of our knowledge, this is the first work that adapts the consensus-based evaluation to dialogues.

## 3 ActDial: A Segment Act Dataset on Open-Domain Dialogues

We propose the new concept of segment act, extracting the core function of each segment in an utterance. We then crowdsource a large-scale open-domain dialogue dataset with our proposed segment act labels, called *ActDial*.

### 3.1 Our Segment Act Tagset

We design an open-domain segment act tagset based on the ISO 24617-2 annotation criteria (Bunt et al., 2019). We define a segment act as a functional label that expresses the communicative goal of participants in a conversation, which is irrelevant to syntactic or sentiment details. Based on this definition, we conduct combination operations, like merging Choice-Question, Check Question, etc. into question, on the original 56 labels proposed by Bunt et al. (2019) and eventually obtain 11 labels as our tagset. These combination operations guarantee a robust coverage on diverse dialogue expressions and mutual exclusiveness between different segment act labels. From our later experiments, these 11 labels capture key information from dialogues while remain simple enough to enable large-scale accurate annotations. Detailed definition and examples of each segment act can be found in Appendix A.1

### 3.2 Datasets and Segmentation

We crowdsourced segment act annotations on ConvAI2 (Dinan et al., 2019) and DailyDialog (Li et al., 2017). The details of the crowdsourcing process are in Appendix A.2

The **ConvAI2** dataset is based on the PersonaChat dataset (Zhang et al., 2018), where all dialogues are constructed by asking two crowdworkers to chat with each other based on randomly assigned persona profiles. ConvAI2 is a widely-used benchmark for many state-of-the-art dialogue systems (Golovanov et al., 2019; Bao et al., 2020; Shuster et al., 2020; Roller et al., 2020).

The **DailyDialog** dataset (Li et al., 2017) is constructed by crawling raw data from various English-learning websites. Note that DailyDialog already has 4 dialog act labels: question, inform, directive, and commissive. Our finer-grained annotation which takes social chit-chat and simple feedback into account can better cover diverse dialogue scenarios and provide extra information.

Following our definition of segment acts, we split each utterance into multiple segments using NLTK (Bird and Loper, 2004) sentence punctuation tokenizer (Kiss and Strunk, 2006). The resulting segments will have their own segment act labels during annotation. Each segment is annotated by three different crowdworkers. With our special tagset design and the segmentation process, annotators can easily reach substantial agreement and deliver a high-quality dataset: Fleiss’ kappa (Fleiss, 1971) achieves 0.754 for DailyDialog and 0.790 for ConvAI2. Detailed statistics of our ActDial dataset is documented in Appendix A.3. Note that the majority of the segments are labeled as question and inform. This is common in dialog act datasets (Stolcke et al., 2000; Yu and Yu, 2019) as most of the dialogues consist of asking for information and stating fact or opinion.

## 4 FlowEval: A Segment-Act-Flow Aware Evaluation Metric

In this section, we describe the details of our proposed dialogue evaluation framework, *FlowEval*. FlowEval is implemented in three stages: segment act harvesting, retrieval, and assessment.

### 4.1 Segment Act Harvesting

In order to utilize the segment act flow, we first need to harvest the segment act labels for an unseen raw dialogue  $U$ . In our experiments unless

specified, the segment act labels are acquired by a text classification model, based on RoBERTa-large (Liu et al., 2019) and fine-tuned on ActDial. The accuracy of this classifier is 90% on unseen data. In the end, we will have the annotated segment act flow  $A_U = \{a_1, \dots, a_i, \dots, a_n\}$  for the dialogue  $U$ , where  $a_i$  is the segment act label for  $i$ -th segment and  $n$  is the number of segments in  $U$ .

### 4.2 Retrieval

For the retrieval process, FlowEval retrieves two sets of dialogues based on segment act features and content features respectively. The search space for FlowEval is our ActDial dataset and the unseen raw dialogue  $U$  serves as query. FlowEval first extracts segment act features from a masked segment act model, and retrieves  $k^a$  nearest neighbors for  $U$  based on our defined similarity function. Then, FlowEval extracts content features from a RoBERTa-large model, and retrieves  $k^c$  nearest neighbours for  $U$  based on another similarity function. The final outcome of this retrieval stage is  $k = k^a + k^c$  relevant dialogues for the unseen dialogue  $U$ . Figure 1 illustrates this process in detail.

**Segment Act Flow Features.** To extract segment act flow features, we treat every segment act label as a word and a segment act flow of a dialogue as a sequence. We then train a masked language model (Devlin et al., 2019) called *ActBERT* on all segment act flows in our ActDial datasets. Detailed implementation of ActBERT is documented at Appendix D. ActBERT has an accuracy of 81% for predicting the masked segment act on unseen data, which is significantly higher than guessing the majority segment act label (67%). This means that our ActBERT indeed captures reliable features from the segment act flow. ActBERT will be used to extract segment act features for any dialogue that has segment act flow.

Given a dialogue  $D$ , we first pass  $D$ ’s segment act flow  $A_D$  into ActBERT. The output of  $h$ -th intermediate layer of ActBERT,  $H_D^h \in \mathbb{R}^{n \times d}$ , will be chosen, where  $h$  is a hyper-parameter,  $n$  is the number of segments in  $D$  and  $d$  is the hidden size of ActBERT.  $H_D^h$  is then max-pooled along the  $n$  dimension to construct a fixed length vector  $\bar{H}_D^h \in \mathbb{R}^d$  as the segment act feature of  $D$ .

We further employ TF-IDF features to constrain the retrieved dialogues to have a similar topic as  $U$ . We collect the word count statistics from our ActDial dataset and compute the TF-IDF feature

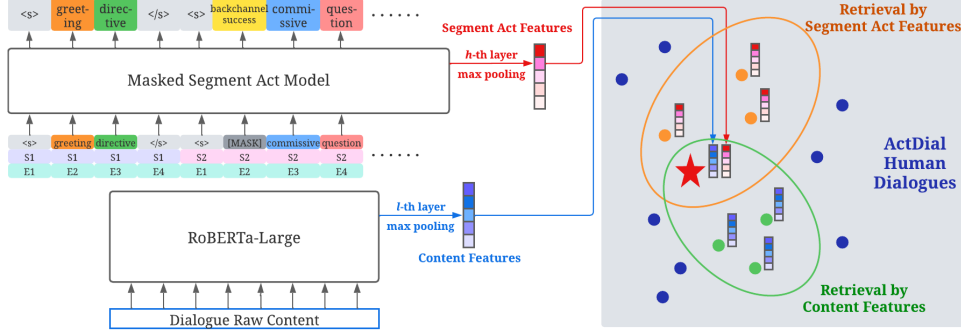


Figure 1: Extract segment act and content features. Retrieve closest human dialogues from ActDial dataset.

vector  $T_D \in \mathbb{R}^v$  for any dialogue  $D$ , where  $v$  is the vocabulary size.

Having the feature set  $\{\bar{H}_U^h, T_U\}$  of  $U$  and  $\{\bar{H}_R^h, T_R\}$  of a human dialogue  $R$  in ActDial, we define an segment-act-based similarity metric  $S^a$  to retrieve  $k^a$  nearest neighbors  $\{R_i\}_{k^a}$ :

$$S^a(U, R) = (1 + \cos(\bar{H}_U^h, \bar{H}_R^h))(1 + \cos(T_U, T_R)) \quad (1)$$

where  $\cos$  is the cosine similarity.  $S^a$  in Eq. 1 will only score high if  $R$  has a segment act flow as well as a topic closed to  $U$ .

**Content Features.** Retrieval with segment act features only might miss dialogues that discussed similar contents as  $U$  but speakers communicated in a different way to  $U$ . Therefore, we retrieve from ActDial again but using features with regard to the content of  $U$ .

We use RoBERTa-large (Liu et al., 2019), a pre-trained language models, to extract the content feature of any dialogue  $D$ . We first feed the raw text of  $D$  into RoBERTa and take the  $l$ -th layer representation  $L_D^l \in \mathbb{R}^{m \times d}$  of RoBERTa, where  $l$  is a hyper-parameter,  $m$  is the number of tokens in  $D$  and  $d$  is the hidden size of RoBERTa.  $L_D^l$  is then max-pooled along the  $m$  dimension to obtain a fixed-length content feature vector  $\bar{L}_D^l \in \mathbb{R}^d$  for  $D$ . Having the content feature  $L_U^l$  of  $U$  and  $L_R^l$  of  $R$  in ActDial, we define a content-based similarity metric  $S^c$  for the second-round retrieval to retrieve  $k^c$  nearest neighbors  $\{R_i\}_{k^c}$ :

$$S^c(U, R) = \cos(L_U^l, L_R^l) \quad (2)$$

$S^c$  in Eq. 2 will output a high score if  $R$ 's content is closed to  $U$ . The final retrieved set of dialogues will be  $\{R_i\}_k = \{R_i\}_{k^a} \cup \{R_i\}_{k^c}$ .

### 4.3 Assessment

We define a metric to find the closest  $R^* \in \{R_i\}_k$  to  $U$  by treating this small retrieved set  $\{R_i\}_k$  as pseudo-references. The distance between  $R^*$  and  $U$  will be the final score of  $U$ . Concretely, we have the following scoring function  $F$ :

$$F(U) = \max_{R \in \{R_i\}_k} wF^a(U, R) + (1 - w)F^c(U, R) \quad (3)$$

$$F^a(U, R) = S^a(U, R) \cdot \text{BLEU}(A_U, A_R) \quad (4)$$

$$F^c(U, R) = \text{BERTScore}(U, R) \quad (5)$$

where  $w$  is a hyper-parameter between 0 and 1. Eq. 3 assess  $U$  from two aspects:  $F^a$ , computed by Eq. 4, indicates whether speakers in  $U$  interact naturally and is evaluated by ActBERT in Eq. 1 and BLEU score (Papineni et al., 2002) of the raw segment act flow  $A_U$ ;  $F^c$ , on the other hand, measures how natural sentences in  $U$  are using BERTScore (Zhang et al., 2020) in Eq. 5.

## 5 Experiments and Analysis

### 5.1 Benchmark Datasets

**Controllable Dialogue Dataset** contains the human-to-bot conversation data collected by See et al. (2019). These conversations are based on the ConvAI2 dataset (Dinan et al., 2019). We extend the original dataset by crowdsourcing segment act labels and human evaluation scores. Details of human evaluation procedural are documented in Appendix C. There are 278 dialogues coming from 3 generative models. 28 dialogues are sampled randomly to form a validation set for hyperparameter tuning, while the rest make up the test set.

**FED Dataset** (Mehri and Eskénazi, 2020) contains 125 human-to-bot conversations coming from three systems. We take the mean of the 5 overall scores

Metric	Controllable Dialogue			FED Dataset			DSTC9 Dataset		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
BLEU	0.132	0.136	0.104	N/A					
BERTSCORE	<u>0.282</u>	<u>0.214</u>	<u>0.162</u>						
CONSENSUS-BERTSCORE	0.284	0.240	0.183	0.0874*	-0.037*	-0.023*	0.060	0.054	0.039
FED Metric	-0.025*	0.010*	0.007*	0.134*	0.126*	0.088*	0.117	0.108	0.078
DynaEval_daily	0.050*	0.051*	0.039*	0.390	0.396	0.278	0.084	0.085	0.061
DynaEval_emp	0.026*	-0.007*	-0.005*	<u>0.464</u>	<u>0.489</u>	<u>0.341</u>	<u>0.029*</u>	0.046	0.033
Flow	-0.065*	-0.029*	-0.020*	-0.073*	-0.003*	-0.002*	<u>0.154</u>	<u>0.148</u>	<u>0.106</u>
FlowEval (Our)	0.301	<b>0.256</b>	<b>0.193</b>	0.246	0.212	0.152	0.088	0.096	0.070
FlowEval (Our) + <u>SOTA</u>	<b>0.327</b>	0.250	0.190	<b>0.468</b>	<b>0.493</b>	<b>0.342</b>	<b>0.165</b>	<b>0.161</b>	<b>0.116</b>
FED Metric + <u>SOTA</u>	0.032*	0.058*	0.042*	0.403	0.411	0.284	0.103	0.093	0.067
DynaEval + <u>SOTA</u>	0.117*	0.109*	0.084*	N/A			0.054	0.059	0.042
Flow + <u>SOTA</u>	0.207	0.140	0.107	0.460	0.471	0.327	N/A		

Table 2: Correlations between different metrics and human evaluation on Controllable Dialogue (test set), FED and DSTC9 datasets. All values are statistically significant to  $p < 0.05$ , unless that are marked by \*. SOTA refers to the previous best performing methods (except our FlowEval) in each dataset and is underlined.

for each dialogue as the human evaluation score in our experiments. We annotate all the segment act labels using the trained classifier described in Section 4.1.

**DSTC9 Dataset** (Gunasekara et al., 2020) contains 2200 human-to-bot conversations from eleven chatbots. We take the mean of the 3 human ratings as the final score. All the segment act labels are predicted by a trained classifier.

## 5.2 Methods

We describe all the baselines used for comparison and the implementation details of our method.

**FED metric** (Mehri and Eskénazi, 2020), leveraging the ability of DialoGPT-large (Zhang et al., 2019) and the use of follow-up utterances, is an automatic and training-free evaluation method widely used by the community (Gunasekara et al., 2020).

**DynaEval** (Zhang et al., 2021) adopts the graph convolutional network to model dialogues, where the graph nodes are dialogue utterances and graph edges represents the relationships between utterances. DynaEval\_emp and DynaEval\_daily denote two variants trained on Empathetic Dialogues (Rashkin et al., 2019) and DailyDialog (Li et al., 2017) respectively. DynaEval\_emp reaches the best correlation on FED dataset.

**Flow score** (Li et al., 2021), considering the semantic influence of each utterance and modeling the dynamic information flow in dialogues, becomes the best evaluation method on DSTC9 dataset.

**BLEU** (Papineni et al., 2002) and **BERTScore** (Zhang et al., 2020) are two popular reference-

based metrics. The performance of BLEU and BERTScore are tested on Controllable Dialogue dataset only, as finding suitable reference is unfeasible on FED and DSTC9 dataset. The process of how to find references on Controllable Dialogue and the implementation of BLEU and BERTScore are documented in Appendix E.

**FlowEval (our method)** tune its hyperparameters on the validation set of Controllable Dialogue dataset and directly apply to the test set of Controllable Dialogue, FED and DSTC9. Besides, since Controllable Dialogue dataset is constructed on top of ConvAI2 (See et al., 2019), we only use the DailyDialog part of ActDial for all the training and retrieval to prevent any data leakage.

## 5.3 Results and Analysis

The common practice to show the effectiveness of a dialogue evaluation metric is to calculate the Pearson, Spearman’s, and Kendall correlation between human evaluation and the automatic evaluation (Mehri and Eskénazi, 2020; Zhang et al., 2021; Li et al., 2021; Yeh et al., 2021), as shown in Table 2. From these results, the following four conclusions can be drawn.

**FlowEval Reaches Comparable Performance.** Across three datasets, our FlowEval achieves the best or comparable correlations with human evaluation. On Controllable Dialogue dataset, all baseline metrics fail to reach meaningful correlation, while FlowEval becomes the top performer. On the other two datasets, the results of FlowEval are comparable with most baselines, though the gap to the

best method is obvious. We perform an ablation study on Controllable Dialogues to further demonstrate the effectiveness of segment acts and our consensus-based framework. Detailed description and results are documented in the Appendix B. We also list one success case and one failure case in the Appendix F to enable a closer observation of our approach.

**Automatic Evaluation Metrics Lack Transferability.** We can observe that the best method on one dataset becomes mediocre on the other datasets, including our FlowEval. FlowEval outperforms all other methods on Controllable Dialogue dataset, but can only get to the second tier on the other two datasets. DynaEval, the best method on FED dataset, loses its advantage when tested on other datasets. The same story also happens for Flow score, a state-of-the-art metric in the DSTC9 dataset. This observation is consistent with study from previous work (Yeh et al., 2021).

One reason for the brittleness of these methods is that their calculations rely on large models. The data used to train these large models plays a decisive role, as we can see from the performance difference between DynaEval\_emp and DynaEval\_daily. In addition, FlowEval depends on the segment act labels and these labels on FED dataset and DSTC9 dataset are annotated by a trained classifier. Even though the classifier has relatively high accuracy (90%), it still injects some errors to the segment act flow, which hinders the application of FlowEval on new datasets. These observations indicate that how to construct a robust dialogue evaluation metric remains a problem for the community.

**FlowEval Can Provide Complementary Information to Other Methods.** Similar to Yeh et al. (2021), we test different combinations of metrics by directly averaging one metric with the previous best metrics on the three datasets, which are BERTScore on Controllable Dialogue dataset, DynaEval\_emp on FED dataset, and Flow score on DSTC9 dataset. The last 4 rows of Table 2 show that FlowEval can consistently push the current correlation ceiling to a new level the most, while many other combinations improve little or even hurt performance. These results imply that segment act is an important missing aspect in dialogue evaluation that worth even further exploration in the future.

**Our Consensus-Based Framework Shows Potential.** In our consensus-based framework, the retrieval step of FlowEval could find pseudo-

references for other reference-based metrics like BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) and make them reference-free.

Here we experiment with BERTScore, as it is the best performing reference-based metric on Controllable Dialogue. The reference-free form of BERTScore, called *Consensus BERTScore*, is similar to our FlowEval, except that we do not employ segment act features in the retrieval step and we exclude the segment act score, i.e., Eq. 4, in the assessment step. As shown in the third row of Table 2, Consensus BERTScore slightly outperforms BERTScore in all three correlations (0.284 vs. 0.282, 0.240 vs. 0.214, 0.183 vs. 0.162).

This promising result shows the potential of our consensus-based framework. It leads a new way to rethink the usability of reference-based metrics in dialogue evaluation.

#### 5.4 What Does Segment Act Bring to Dialogue Evaluation?

Compared with semantic-meaning-focused metrics, what does segment act bring to dialogue evaluation? We hypothesize the explicit involvement of segment acts can bring useful information, complementary to semantic-meaning-focused metrics.

We illustrate our hypothesis in Figure 3. If segment act is useful, the segment-act-based evaluation  $v_p$  should be positively correlated to human evaluation  $v_o$ , i.e.,  $v_p$  has roughly the same direction as  $v_o$  but with a small angle  $\theta_2$ . If segment act is complementary to semantic-meaning-focused metrics, the segment-act-based evaluation  $v_p$  should be almost orthogonal to the semantic-meaning-focused evaluation  $v_m$ , i.e.,  $v_m$  falls into the other side of  $v_o$  so that  $v_m$  is also positively correlated to  $v_o$  with a small angle  $\theta_1$  but almost orthogonal to  $v_p$  with a large angle  $\theta_3 = \theta_1 + \theta_2$ . These angles  $\theta_1$ ,  $\theta_2$  and  $\theta_3$  could be characterized by the correlation of two evaluation results. A higher correlation implies a smaller angle.

We conduct experiments on the test set of Controllable Dialogue dataset to validate our hypothesis. Two of the popular semantic-meaning-focused metrics are BERTScore (Zhang et al., 2020) and BLEU (Papineni et al., 2002). We modify the retrieval and assessment parts of our FlowEval, so that only segment act information is utilized. We denote this variant as *FlowEval\_seg*.

As we could observe from the first three rows of Table 3 that the FlowEval\_seg, BLEU and

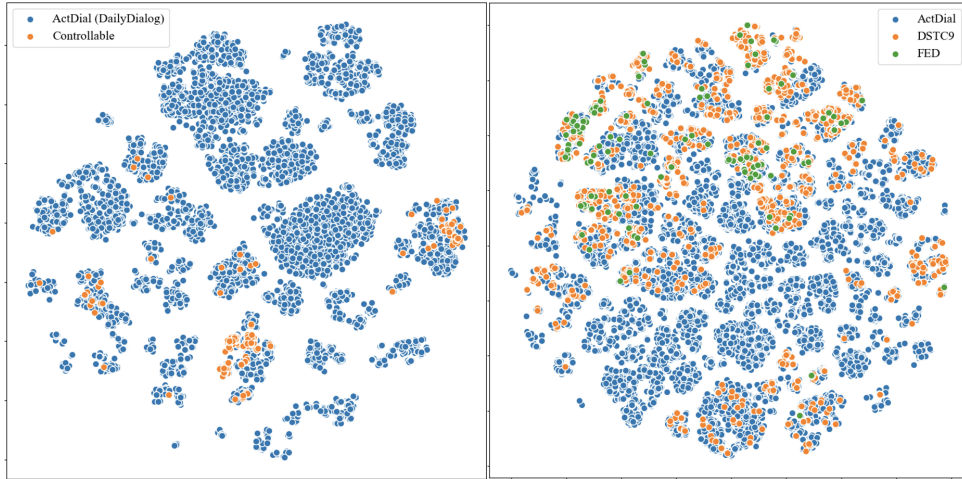


Figure 2: Segment act feature space of Controllable Dialogue, FED, DSTC9 dataset and the retrieval set ActDial. We have a separate plot for Controllable Dialogue because the ActDial we used are different (See Section 5.2).

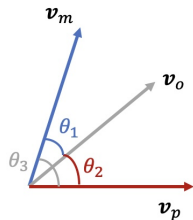


Figure 3: The relationships, in our hypothesis, between human evaluation  $v_o$ , semantic-meaning-focused evaluation  $v_m$ , and segment-act-based evaluation  $v_p$ .

Metric 1	Metric 2	Pearson	Spearman	Kendall
FlowEval_seg	Human	0.191	0.151	0.113
BLEU	Human	0.132	0.136	0.104
BERTScore	Human	0.282	0.214	0.162
FlowEval_seg	BERTScore	0.014*	-0.030*	-0.022*
FlowEval_seg	BLEU	0.067*	0.042*	0.026*
BLEU	BERTScore	0.576	0.637	0.460

Table 3: Inter-correlations between FlowEval\_seg, BERTScore, BLEU, and human evaluation. FlowEval\_seg is a version of FlowEval using segment act flow only for assessment. All values are statistically significant to  $p < 0.05$ , unless that are marked by \*.

BERTScore all exhibit strong correlation to human evaluation. Unsurprisingly, BLEU and BERTScore are highly correlated (the last row of Table 3), since both of them focus on the semantic meaning of dialogues. In line with our hypothesis, the BLEU-FlowEval\_seg correlation and BERTScore-FlowEval\_seg correlation is far smaller (rows 4-5 of Table 3), which indirectly shows that segment act can evaluate dialogues from a complementary perspective. These findings resonate with the theory

from Bender and Koller (2020), where the *meaning* and the *communicative intent*, i.e., segment act here, are considered to be two decoupled and complementary dimensions.

## 5.5 Why Does Consensus Work?

We investigate why consensus-based framework can perform well in dialogue evaluation by visualizing the segment act feature space, an essential aspect in the retrieval process of FlowEval. We compare the segment act feature distribution between the three test sets and their corresponding retrieval sets, projecting these features to 2-dimensional space by t-SNE (van der Maaten and Hinton, 2008) as shown in Figure 2. We did not tune any hyperparameter to obtain these results, in consideration of the sensitivity of t-SNE plots.

The core idea of consensus lies on using the nearest neighbors as references to measure a newcomer. Only if the suitable nearest neighbors consistently exist, will the consensus of them have meaningful indication to evaluate a new subject. We can observe from Figure 2 that, even though dialogues in three test sets are diverse, every datapoint from the test sets is surrounded by datapoints from the retrieval sets. We can always reliably find good references for a new dialogue, which explains why using consensus in dialog evaluation is promising. Moreover, this desirable coverage is achieved by an attainable amount of datapoints. It only needs 10,494 and 31,993 dialogues as retrieval sets in our experiments to get good results. The power of the consensus may go stronger and more reliable if the size of retrieval set grows, which could be a favor-



able property in many of industrial applications.

## 6 Conclusion

In this work, we propose a consensus-based reference-free framework for open-domain dialog evaluation with segment act flows. From extensive experiments against the state-of-the-art baselines, our method can reach the best or comparable correlation with human evaluation. Our segment-act-based methods complement well to previous semantic-meaning-focused methods, pushing the ceiling of correlations. Moreover, the promise of our consensus-based framework encourages us to step further in the direction of dialog evaluation.

## Limitations

Our segment act dataset, ActDial, is constructed based on two widely-adopted open-domain dialogue datasets, ConvAI2 (Dinan et al., 2019) and DailyDialog (Li et al., 2017). Despite of various benefits, ActDial also inherits some of the limitations from ConvAI2 and DailyDialog. The scale of the dataset could be larger. The nature of ConvAI2 dialogues, learning personas of each other, pushes the segment act distribution towards question and inform slightly. These limitations do not interfere too much with our methods and our extensive experiments still show significant results. We will potentially improve our dataset in the future.

This work also brings the consensus-based framework into open-domain dialogue evaluation. We show the effectiveness of this framework when incorporating segment act flow and content information. Yet, the full potential of the consensus-based framework still needs more exploration. We will leave this as future work.

## Ethics Statement

A big part of this work contains (1) the data annotation on two existing benchmark datasets of conversation modeling: the ConvAI2 dataset and the DailyDialog dataset and (2) human evaluation on the overall quality of generated conversations. As our ActDial is built upon the existing datasets, we follow the original copyright statements of these two datasets and will further release our segment act annotations to the research community. During annotation, we only collected the segment act annotations, and no demographic or annotator’s identity information was collected. In addition, we provide

a detail description of human evaluation design in Appendix C.

## Acknowledgements

This work was supported by Centre for Perceptual and Interactive Intelligence Limited, UGC under Research Matching Grant Scheme, the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14210920 of the General Research Fund), and Tencent Rhino-Bird Research Award.

## References

- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. [PLATO: Pre-trained dialogue generation model with discrete latent variable](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. [Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling](#). *CoRR*, abs/1810.00278.
- Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Chengyu Fang, and Kars Wijnhoven. 2019. [The dialogbank: dialogues with interoperable annotations](#). *Lang. Resour. Evaluation*, 53(2):213–249.
- Alessandra Cervone and Giuseppe Riccardi. 2020. [Is this dialogue coherent? learning from dialogue acts and entities](#). *CoRR*, abs/2006.10157.
- Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-

- of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. [Exploring nearest neighbor approaches for image captioning](#). *CoRR*, abs/1505.04467.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. 2019. [The second conversational intelligence challenge \(convai2\)](#). *CoRR*, abs/1902.00098.
- Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. [Cider-r: Robust consensus-based image description evaluation](#). *CoRR*, abs/2109.13701.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2019. [Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems](#). *CoRR*, abs/1911.01456.
- Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyril Truskovskiy, Alexander Tselousov, and Thomas Wolf. 2019. [Large-scale transfer learning for natural language generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.
- R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David R. Traum, Maxine Eskénazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. [Overview of the ninth dialog system technology challenge: DSTC9](#). *CoRR*, abs/2011.06486.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. [The ATIS spoken language systems pilot corpus](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
- Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. [GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9230–9240. Association for Computational Linguistics.
- Tibor Kiss and Jan Strunk. 2006. [Unsupervised multilingual sentence boundary detection](#). *Computational Linguistics*, 32(4):485–525.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. [An evaluation dataset for intent classification and out-of-scope prediction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. [Conversations are not flat: Modeling the dynamic information flow across dialogue utterances](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.

- Shikib Mehri and Maxine Eskénazi. 2020. [Unsupervised evaluation of interactive dialog with dialogpt](#). *CoRR*, abs/2006.12719.
- Shikib Mehri and Maxine Eskenazi. 2020. [USR: An unsupervised and reference free evaluation metric for dialog generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.
- Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. [Towards holistic and automatic evaluation of open-domain dialogue generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. [Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. [Recipes for building an open-domain chatbot](#). *CoRR*, abs/2004.13637.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? how controllable attributes affect human judgments](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. [The ICSI meeting recorder dialog act \(MRDA\) corpus](#). In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. [The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.
- Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. [Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents](#). *CoRR*, abs/2201.04723.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *CoRR*, cs.CL/0006023.
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. [RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. [Cider: Consensus-based image description evaluation](#). *CoRR*, abs/1411.5726.
- Marilyn Walker and Rebecca Passonneau. 2001. [DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems](#). In *Proceedings of the First International Conference on Human Language Technology Research*.
- Marilyn A. Walker, John S. Aberdeen, Julie E. Boland, Elizabeth Owen Bratt, John S. Garofolo, Lynette Hirschman, Audrey N. Le, Sungbok Lee, Shrikanth S. Narayanan, Kishore Papineni, Bryan L. Pellom, Joseph Polifroni, Alexandros Potamianos, P. Prabhu, Alexander I. Rudnicky, Gregory A. Sanders, Stephanie Seneff, David Stallard, and Steve Whittaker. 2001. [Darpa communicator dialog travel planning systems: the june 2000 data collection](#). In *INTERSPEECH*.
- Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. [Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space](#). *Advances in Neural Information Processing Systems*, 30.
- Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. [A comprehensive assessment of dialog evaluation metrics](#). *CoRR*, abs/2106.03706.

Dian Yu and Zhou Yu. 2019. [MIDAS: A dialog act annotation scheme for open domain human machine spoken conversations](#). *CoRR*, abs/1908.10023.

Chen Zhang, Yiming Chen, Luis Fernando D’Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. [Dynaeval: Unifying turn and dialogue level evaluation](#). *CoRR*, abs/2106.01112.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. [Dialogpt: Large-scale generative pre-training for conversational response generation](#). *CoRR*, abs/1911.00536.

## A Details of Our ActDial Dataset

### A.1 Segment Acts definitions

For the formal definitions and examples of segment act, please refer to Table 4. The eleven segment act labels cover three major communication activities: (i) general task, which includes information-transfer activities and action-discussion activities; (ii) social obligation management, which includes typical social conventions in communication; and (iii) simple feedback, which includes simple non-informative feedback about the processing of previous utterances.

We segmented all the dialogue utterances using the NLTK sentence punctuation tokenizer (Kiss and Strunk, 2006) that mainly consists of a set of rule-based regular expressions on punctuation.

### A.2 Crowdsourcing Segment Act Annotation

We crowdsourced segment act annotation from annotators whose native language is Mandarin Chinese (zh-cmn), but more importantly, they are proficient in English (en-US). More than 50 annotators participated after rigorous training to ensure data quality. Each segment is annotated by three different annotators. If the initial three annotations are all different, further round(s) of annotation on this

segment would be conducted until it got a majority vote (at least two annotations are the same).

Besides Fleiss’ kappa (Fleiss, 1971) mentioned in Section 3, we here report Fleiss’ kappa in a new setting and the overall sample accuracy to show the quality of our annotations.

Since the segment act distribution is unbalanced, we calculated another Fleiss’ kappa excluding all the annotations with the most dominant segment act, i.e., `inform`, to eliminate potential bias. In this setting, the new kappa is 0.768 for DailyDialog and 0.775 for ConvAI2, staying roughly the same as the overall ones. These results prove the robustness of our annotations.

Although it is impossible to check the correctness of every single annotation, we do perform sampling inspection when collecting the annotations everyday. In total, We sampled 8,000 segments randomly and annotated these segments by ourselves. Since we have a deeper understanding than our annotators and our annotations are examined multiple times by ourselves, our annotations on these 8,000 segments can be considered as ground truth. The majority votes of crowdsourced annotations are later compared with the ground truth labels to obtain sample accuracy. The sample accuracy in DailyDialog annotation is 0.90 and that in ConvAI2 is 0.93. The small gap of the accuracy is due to the difference in dialogue complexity.

### A.3 Dataset Statistics and Distributions

For the ConvAI2 dataset, we collected 481,937 segment acts on the training set, and 29,232 segment acts on the validation set. Since the testing set is not publicly available, we did not annotate it.

For the DailyDialog dataset, we gathered 178,604 segment acts on the training set, 16,500 segment acts on the validation set, and 16,028 segment acts on the testing set.

Note that even though ConvAI2 and DailyDialog split their data for training, validation, and testing purpose, it is not always necessary to mechanically follow the splits. Our annotations on ConvAI2 and DailyDialog can be used as a unity, *ActDial*, depending on the research problems.

Table 4 shows the distribution of all segment acts on our dataset. The segment act distribution is unbalanced. Specifically, the distribution is highly skewed to `inform` and `question`, which is not surprising because ConvAI2 and DailyDialog are chitchat datasets and the majority of com-

General Dimension	Segment Act	Definition	Examples	Distribution
General Task	inform	The sender makes the addressee know some information which he assumes to be correct.	“The train is leaving.”, “The meeting starts in 5 minutes.”	65.702%
	question	The sender asks the addressee to provide some information which he assumes the addressee knows.	“What time is it?”, “Where is the nearest bank?”	16.529%
	directive	The sender asks the addressee to perform an action.	“Please don’t do this ever again.”	2.880%
	commissive	The sender considers to perform an action which he believes would be in addressee’s interest, or he has been requested/suggested to perform by the addressee.	“I will not do that any more.”, “May I offer you an upgrade?”	0.517%
Social Obligation Management	greeting	The speakers inform the presence of each other.	“how are you?”, “I’m fine.”	6.023%
	goodbye	The speakers inform the end of the dialog.	“Bye”, “See you.”	0.172%
	apology	The speakers express or mitigate the feelings of regret.	“Sorry.”, “No problem.”	0.542%
	thanking	The speakers express or mitigate the feelings of gratitude.	“Thanks.”, “You are welcome.”	1.049%
Simple Feedback	backchannel-success	The speakers succeed in processing the previous dialog.	“Okay”, “Uh-huh”	6.543%
	backchannel-failure	The speakers fail in processing the previous dialog.	“Sorry?”, “Excuse me?”	0.030%
	check-understanding	The sender wants to check whether the addressee succeed in processing the previous dialog.	“Do you get what I just said?”	0.013%

Table 4: Our ISO-format open-domain segment act tagset: the definition, examples, and distribution

munication activities is exchanging information. In addition, few written dialogues between two strangers, the setting of ConvAI2 and DailyDialog, involve apology or encounter communicative difficulties, which results in the rare occurrences of apology, backchannel-failure, and check-understanding segment acts. However, it is still essential to include these segment acts as they take place more commonly in spoken dialogues in the real world.

## B Ablation Study on Controllable Dialogues

We perform ablation study on Controllable Dialogues and obtained positive results. This experiment is designed to reveal the effectiveness of segment act, so content-related information and features are excluded from the whole process. Specifically, we remove the content feature and only used the segment act flow feature during the retrieval (Section 4.2). We later assessed each dialogue on this shrunk retrieval set. The Pearson, Spearman’s, and Kendall correlations in this setting are 0.298, 0.252, and 0.189 respectively. These results de-

crease slightly from our full version of FlowEval (0.301, 0.256, and 0.193) but remain higher than the previous SOTA (0.282, 0.214, and 0.162). This ablation study strengthens our claim on the effectiveness of segment acts and our consensus-based framework.

## C Human Evaluation for Controllable Dialogue

We collected human judgements from Amazon Mechanical Turk (AMT). The crowd-workers are provided with the full multi-turn conversation for evaluation. We ask crowd-workers to evaluate the *relevancy*, *avoiding contradiction*, *avoiding repetition*, *persona consistency* and *overall quality* of the conversation. The reason for designing the human evaluation on different aspects is that we assume a good conversation between human and a dialogue system should satisfy the following properties: (1) generating relevant and non-repetitive responses (*relevancy* and *avoiding repetition*), (2) memorizing the dialogue history and generating non-contradictory information (*avoiding contradiction*), (3) maintaining a consistent persona/topic

	Controllable Dialogue Dataset				
	RELEVANCY	NO CONTRADICTION	NO REPETITION	CONSISTENCY	OVERALL
Seq2Seq	0.7802	0.8791	0.3846	0.9010	3.4505
Seq2Seq + Repetition	<b>0.8437</b>	<b>0.9062</b>	0.7812	0.8541	3.6250
Seq2Seq + Specificity	0.8351	0.8681	<b>0.8351</b>	<b>0.8681</b>	<b>3.8791</b>

Table 5: Controllable Dialogue (See et al., 2019) evaluation results by AMT crowd-workers.

(*persona/topic consistency*), (4) formulating a natural conversation (*overall quality*).

The first four aspects are formulated as binary-choice questions, and the overall quality is formulated as Likert question on a 1-5 scale, where higher is better. During evaluation, we did not distinguish whether an utterance is generated by human or by dialogue model, because we want the evaluation is about the full conversation, rather than just utterances generated by the dialogue model.

To ensure better data quality, Turkers are selected by their job success rate and geographic location (only admits turkers from English speaking countries). Before starting our evaluation job, turkers must read through our detailed guideline. For each dialogue, a turker is asked to evaluate the dialogue from the following perspectives:

1. **Irrelevant response (binary)**: Whether or not the speaker generates a response which seems to come out of nowhere according to the conversation history. Binary score.
2. **Contradictory information (binary)**: Whether or not the speaker generates a response which contradicts to the common sense or to what himself just said in the previous conversation. Binary score.
3. **Repetitive response (binary)**: Whether or not the speaker generates a response which has the same meaning as his previous utterance(s). Binary score.
4. **Inconsistent with persona (binary)**: Whether or not the speaker generates a response which is not consistent with his persona profile. **Only used if the dialogues-to-evaluate follow ConvAI2 setting and are generated with personas.** Binary score.
5. **Topic shifts (binary)**: Whether or not the speaker generates a response which belongs to a completely different topic compared with the previous conversation history. **Only used**

**if the dialogues-to-evaluate follow Daily Dialogue setting and are not generated with personas.** Binary score.

6. **Overall score (1-5)**: An overall impression of the dialogue quality, not necessary to have any relationship with the aspects above. Score is between 1 to 5 inclusive, all integer. The higher the better.

The evaluation results are examined by ourselves. Incorrect annotation would be rejected and re-evaluated by another turker. The final evaluation results are shown as Table 5.

## D Implementation of ActBERT

ActBERT follows the architecture of RoBERTa (Liu et al., 2019). The vocabulary size is relatively small as it only contains 11 segment acts and other special tokens. It has 4 hidden layers, 4 attention heads, and a hidden dimension size of 256. Speaker information is included using different input token type. Similar to the masked language model task, we use a masked segment act task during the training.

## E Implementation of BLEU and BERTScore

Controllable Dialogue (See et al., 2019) are trained on ConvAI2 dataset whose setting is two participants talking about their own personas. This unique characteristics makes it feasible to find references for BLEU, BERTScore, or other reference-based metrics.

We take dialogues, from the testing set of ConvAI2, that have the most overlapping personas as the references for a dialogue. Although not as convincing as references in machine translation task, references obtained in this way prove to be helpful to dialogue evaluation. Both BLEU and BERTScore reaches relatively high correlations on Controllable Dialogue.

The smooth function of the BLEU score is NIST geometric sequence smoothing (Bird and Loper,

<b>Human:</b> My motorbike is giving me trouble. What a bad start to the day!
<b>Chatbot:</b> Sorry to hear that. What do you do for a living?
<b>Human:</b> I'm a waitress. What about you?
<b>Chatbot:</b> I am old, old, old, old, old, old, old, old, old, old, old, old.
<b>Human:</b> Ha ha! Ok. I'm young with big brown eyes always looking for fun!
<b>Chatbot:</b> I am old, old, old, old, old, old, old, old, old, old, old, old, old, old.
<b>Human:</b> I get it!! So do you like mtorbikes?
<b>Chatbot:</b> I am old, old, old, old, old, old, old, old, old, old, old, old, old, old, old, old.
<b>Human:</b> Again. Can't you say anything else?
<b>Chatbot:</b> Old, old, old, old, old, old, old, old, old, old, old, old, old, old, old, old.
<b>Human:</b> You already said that.
<b>Chatbot:</b> Old, old, old, old, old, old, old, old, old, old, old, old, old, old, old, old.

Table 6: A conversation from Controllable Dialogue with human rating as 1 (lowest). FlowEval can successfully give it the lowest score, while DynaEval, Flow score fail to distinguish the low quality.

2004). BERTScore are calculated by using the package from its authors (Zhang et al., 2020).

## F Cases Analysis

In this section, we exhibit two real cases from Controllable Dialogue to show the strengths and weaknesses of our FlowEval.

**A Successful Case** Table 6 shows a poor conversation with human rating 1 point, the lowest point possible. Our FlowEval can successfully rank this dialogue as the worst among 278 dialogues, while DynaEval and Flow score consider it not bad but mediocre. To be exact, DynaEval ranks it at the top 170th and Flow score places it at the 114th.

Repetition is the main flaw of this conversation. From the segment act feature perspective, the abnormal long inform flows caused by a series of “old” provide FlowEval enough evidence to penalize this behavior. From content feature aspect, the repeated words do not contribute well to our word overlapping assessment in FlowEval either. Other methods do not have specific mechanism that help them distinguish this repetition, which makes them vulnerable in similar cases.

<b>Human:</b> Hello how are you today
<b>Chatbot:</b> Doing well, how about you?
<b>Human:</b> I'm doing well just getting ready to go to work
<b>Chatbot:</b> Nice, where do you work?
<b>Human:</b> I am a librarian what about you
<b>Chatbot:</b> I work as an administrator at the office of a local construction company.
<b>Human:</b> That sounds like a lot of fun
<b>Chatbot:</b> Oh, believe me, it's not. So, do you have any hobbies?
<b>Human:</b> I love being in the library
<b>Chatbot:</b> It's good to really enjoy your work.
<b>Human:</b> Thank you very much
<b>Chatbot:</b> So, have you ever travelled outside the country?

Table 7: A conversation from Controllable Dialogue with human rating as 4 (high quality). FlowEval and Flow score fail to give it a high score, while DynaEval considers it one of the best.

**A Failure Case** Here we show an example where FlowEval fails to deliver the correct evaluation. The dialogue in Table 7 is of high quality: annotators give it 4 points, the highest points among 278 dialogues. However, FlowEval ranks it at the 253rd. and Flow score ranks it at the 178th. DynaEval gets it right, reviewing it as the 3rd best dialogue.

The segment act flow of this dialogue is relatively natural. The contents appear not so frequently in the ActDial dataset. These characteristics make it hard for FlowEval to output a correct ranking.

Just like our competing baselines, more analysis and case study are needed to determine a more concrete pattern.

## G Computational Cost

All of our experiments are run on a single NVIDIA V100 GPU. Note that our method does not require excessive computational power and GPU with lower computational ability can reproduce our results in reasonable amount of time.