Weakly Supervised Turn-level Engagingness Evaluator for Dialogues

SHAOJIE JIANG, University of Amsterdam, The Netherlands
SVITLANA VAKULENKO*, Amazon, Spain
MAARTEN DE RIJKE, University of Amsterdam, The Netherlands

Engagingness is an important measurement for evaluating open-domain conversational systems. The standard approach to evaluating dialogue engagingness is by measuring conversation turns per session (CTPS), which implies that the dialogue length is the main predictor of the user engagement with a dialogue system. The main limitation of CTPS is that it can only be measured at the session level, i.e., once the dialogue is over. But a dialogue system has to continuously monitor user engagement throughout the dialogue session as well. Existing approaches to measuring turn-level engagingness require human annotations for training. We pioneer an alternative approach, Weakly Supervised Engagingness Evaluator (WeSEE), which uses the remaining depth for each turn as a heuristic weak label for engagingness. WeSEE does not require human annotations and also relates closely to CTPS, thus serving as a good learning proxy for this metric. We show that WeSEE achieves the new state-of-the-art results on the *Fine-grained Evaluation of Dialog* dataset (0.38 Spearman correlation coefficient) and the *DailyDialog* dataset (0.62 Spearman correlation coefficient).

CCS Concepts: • **Information systems** → *Users and interactive retrieval.*

Additional Key Words and Phrases: Conversation analysis, engagingness, user experience

ACM Reference Format:
Shaojie Jiang, Svitlana Vakulenko, and Maarten de Rijke. 2023. Weakly Supervised Turn-level Engagingness Evaluator for Dialogues. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3576840.3578319

1 INTRODUCTION

Engagingness, the quality of being engaging, is an important aspect of an open-domain conversational system. At the session level, it reflects user satisfaction with the system [37]. At the turn level, it also measures how willing the user is to continue the conversation [16]. We are particularly interested in engagingness of open-domain conversational systems, such as conversational search and recommendation systems [24] as opposed to task-oriented dialogue systems. Task-oriented dialogue systems are usually optimized for utility, and high engagingness there may actually reflect poor performance of the system [29]. Engagingness is typically measured using the *conversation turns per session* (CTPS) since more engaging conversations tend to have more turns than less engaging ones [8, 32]. CTPS values can easily be obtained *off-line* to compare engagingness levels of different systems. However, performing *online* turn-level engagingness evaluation is of even greater importance since it can be used to guide the dialogue generation process directly or to choose between different candidate responses [37]. Figure 1 provides examples of turn-level engagingness evaluation.

*Research conducted when the author was at the University of Amsterdam.

Manuscript submitted to ACM

Engaging responses:	
1	**Context:** Yes yes. I've been to Tokyo as well. It's so nice!
1	**Response:** What did you do here? WeSEE: 1.00
2	**Context:** no not yet she is a little to young
2	**Response:** i love to go running, its my favorite passion. i wish i ran for a living WeSEE: 0.93
Not Engaging responses:	
3	**Context:** of course. a two-week paid vacation a year, a five-day workweek.
3	**Response:** good. anything else? WeSEE: 0.32
4	**Context:** what am i supposed to do then?
4	**Response:** a good rest is all you need, and drink more water. i'll write you a prescription. WeSEE: 0.18

Fig. 1. Illustrations of turn-level engagingness evaluation. The predictions of our proposed model (WeSEE) for the responses are at the end of each response line, in the range of $[0, 1]$. The context and response examples are taken from the *DailyDialogue* (1), *PersonaChat* (2), and *Fine-Grained Evaluation of Dialog* (3 and 4) datasets. See Figure 6 for more examples.

Recent work on engagingness has focused on training neural models to predict turn-level engagingness [5, 6, 17, 37], which is an important step towards online evaluation of the performance of conversational systems. However, existing approaches to predicting turn-level engagingness exhibit important limitations. For example, the most common approach is to address engagingness prediction as a binary classification task [6, 37]. The main reason is the need for human labels for training the models. While labeling turns as engaging or not engaging is conceptually simple, the approach lacks scalability. In addition, the produced binary labels may not reflect differences between engagingness levels sufficiently well. As a reasonable and scalable alternative, we propose a simple approach of using weak supervision for evaluating the engagingness of a conversational system. Our experiments show that this approach has better correlation with human judgments of engagingness than previously proposed approaches.

1.1 Proposed approach

We first use the *remaining depth* (RD) as heuristic weak labeling for turn-level engagingness; RD is defined as the number of conversation turns in a session following the current one. Then, we train a regression model for turn-level engagingness prediction. There are multiple advantages to our approach. First, RD labels for the training data can be interpreted as the CTPS of the sub-dialogue starting from the current turn onward, and intuitively, highly engaging responses are likely to result in large RD values. This is not always the case in reality, but RD labels can serve as noisy indicators of engagingness, and can be easily inferred for existing dialogue data, which saves extra annotation efforts. Second, we show that this weak signal can be used to train a BERT-based [1] regressor to be an engagingness evaluator and achieve state-of-the-art correlation with human engagingness judgments on two dialogue datasets. Our proposed model, *Weakly Supervised Engagingness Evaluator* (WeSEE), can not only output real numbers that reflect fine-grained engagingness levels, but it can also use as little input as a single-turn text to make predictions, thus making it broadly applicable.

In our experiments, WeSEE achieves Pearson and Spearman correlation coefficients of 0.36 and 0.38 with human annotations, respectively, on the Fine-grained Evaluation of Dialog (FED) dataset [17], and 0.58 and 0.62 on the DailyDialog-Human dataset [6], which is the new state-of-the-art performance in engagingness prediction on both datasets.

1.2 Main contributions

Our main contributions are: (i) We propose to use *remaining depth* (RD) as weak labels for turn-level engagingness, which avoids the need for explicit human annotations. (ii) We formulate engagingness prediction as a regression task, therefore, the predicted scores can distinguish different magnitudes of engagingness. (iii) We show that a BERT-based model can produce decent predictions with only single dialogue turns, while the use of more turns can lead to improved correlation coefficients with human annotations. (iv) We share our source code (also in the supplementary material), the datasets used, implemented baselines, and trained parameters at https://github.com/ShaojieJiang/lit-seq.

We provide a brief overview of the related work in Section 2. Then, we introduce the WeSEE model in Section 3. In Section 4 and 5, we explain our experimental setup and analyze the results of our experiments, respectively. We conclude with a summary and directions for future work in Section 6. Ethical considerations with respect to using dialogue data in this work are made explicit in Appendix A.

2 RELATED WORK

We start by providing a summary of the state-of-the-art in automatic dialogue quality evaluation. We then zoom in on the challenge of measuring engagement and characterize the main limitations related to measuring dialogue engagingness that motivate our work.

2.1 Dialogue quality

Dialogue quality is a multi-faceted phenomenon and cannot be evaluated along a single dimension only [23, 27, 36]. However, most evaluation approaches proposed to date evaluate either the overall dialogue quality or the response quality on the turn-by-turn level [5, 12, 17, 18, 21, 23, 28, 37, 38]. Being versatile also means sacrificing performance as well as interpretability with respect to the individual aspects of the dialogue quality, such as dialogue engagingness [36]. Our experiments show that such general-purpose quality evaluators do not achieve a high correlation with manually-labeled engagingness scores.

2.2 Measuring user engagement

User engagement is a quality of user experience that is characterized by the depth of a user's investment when interacting with a digital system [19]. Different methods have been used to assess engagement [9], including (i) behavioral metrics such as web page visits and dwelling time; (ii) neurophysiological techniques including eye tracking; and (iii) self-reports such as questionnaires, interviews, diary entries and verbal elicitation [20]. We are interested in evaluating engagement with dialogue systems. While extensive descriptive studies of interaction behavior with dialogue systems exist (for example, in the context of conversational search [31]), the evaluation of engagingness has been less well studied than overall dialogue quality evaluation. One line of work builds on user studies. For instance, Fergencs and Meier [4] recently conducted a controlled interactive information retrieval experiment with 10 participants to compare a chatbot to a graphical search user interface in terms of engagement and usability. And Papenmeier et al. [22] formulate design guidelines for product search assistants in e-commerce, including switching between reactive and proactive roles depending on user engagement, based on conversations in a user study (with 24 participants), where experts engage with users to help find the right product for their needs.

In addition to engagingness measurements that are based on user studies, there is some prior work on automatically measuring engagingness. But the few approaches to automatically measuring engagingness that exist, have several

drawbacks. First, training supervised models that predict engagingness requires manual labels, which are difficult to obtain [6, 37]. Recently, Liang et al. [13] proposed to use heuristic rules for automatic engagingness annotation, but only restricted to unengaging user responses. Second, defining annotation guidelines for measuring dialogue engagingness has proved to be a hard task. For example, Yi et al. [37] resorted to binary labels (engaging/not engaging) that are easier to acquire but are not very descriptive. Ghazarian et al. [6] grouped the original samples annotated with five engagingness levels into two because of the highly imbalanced training data. Third, formulating the problem of measuring engagingness as a classification task limits the models' ability to distinguish between different levels of engagingness.

2.3 Predicting engagingness

The main novelty of our work is that we establish a simple heuristic that allows us to train a reliable turn-level dialogue engagingness evaluator that shows a high correlation with human judgments. Instead of using manual labels, we employ an automatic approach to deducing remaining depth (RD) as weak labels for engagingness. This approach can be applied to any multi-turn dialogue dataset, allowing one to extract engagingness signals that are naturally embedded in the dialogue data itself, thus no extra annotation is needed.

We also argue in favor of formulating the problem of dialogue engagingness prediction as a regression task, instead of a classification task as in prior work, which brings several important benefits. First, our proposed model WeSEE trains on continuous-valued labels in $[0, 1]$ rather than discrete class labels. Thereby, it does not suffer from the class imbalance problem. Second, WeSEE can also better exploit ordinal relations between engagingness levels and distinguish between them on a fine-grained scale.

To the best of our knowledge, the only other approach to engagingness prediction that does not require human annotations is due to Mehri and Eskénazi [17]. They use the log-likelihood of a curated pool of the follow-up utterances produced by DialoGPT [40] as their engagingness scores. Log-likelihood is not bounded and is influenced by utterance length. In contrast, the normalized WeSEE scores fall in the range $[0, 1]$ and allow one to compare the engagingness of candidate responses of different lengths.

3 OUR APPROACH: AN ENGAGINGNESS EVALUATOR TRAINED ON WEAK LABELS

We use $D = (X_1, X_2, \ldots, X_{|D|})$ to represent a dialogue session in the dataset that has $|D|$ turns, with one turn denoting the message from one speaker at a time. Consecutive messages from the same speaker are considered as a single turn. We assume that there are at least two dialogue speakers, and each turn contains a response to the previous turn. Each turn i may consist of up to n tokens: $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$.

The *remaining depth* (RD) of X_i normalized to $[0, 1]$ is calculated as:

$$\mathrm{RD}_i = \frac{|D| - i}{|D| - 1}, \tag{1}$$

which we subsequently use as weak engagingness labels when formulating the RD prediction problem as a regression task. In this manner, each pair (X_i, RD_i) is treated as a single data point for training the prediction model.

Our WeSEE model is based on BERT as illustrated in Figure 2. The dialogue turns are embedded with BERT and then averaged before making predictions. More concretely, we first use the pretrained BERT model [1] to get a vector representation of the turn X_i. To use the context available from the dialogue history, we also embed up to $k \geq 0$ turns

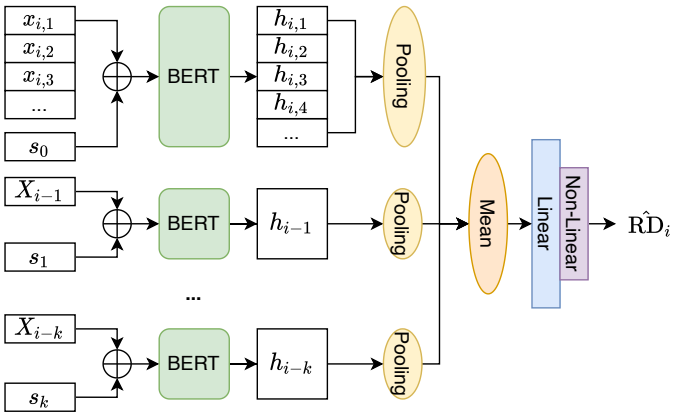

Fig. 2. WeSEE model architecture.

that occur before the i-th turn:

$$\begin{aligned} h_i = \mathrm{Mean}(\ & \mathrm{BERT}(X_i \oplus s_0), \\ & \mathrm{BERT}(X_{i-1} \oplus s_1), \\ & \vdots \\ & \mathrm{BERT}(X_{i-k} \oplus s_k)), \end{aligned} \tag{2}$$

where Mean denotes mean pooling and $h_i \in \mathbb{R}^{hid\_sz}$ is a hid_sz-dimensional contextualized vector representation for turn X_i; s_0, s_1, ..., s_k are segment embeddings, and $\oplus$ denotes element-wise addition. The representation for each turn is a vector obtained by pooling the BERT positional outputs. We consider four different pooling methods in our experiments: class-token pooling uses the output of the special [CLS] token; and *mean*, *max* and *min* pooling take the element-wise average, maxima and minima of the BERT outputs produced for each of the input tokens, respectively.

Finally, we use a linear layer to project h_i to a scalar as the predicted engagingness level and use a simple cut-off to normalize it to $[0, 1]$ range:

$$\hat{\mathrm{RD}}_i = \min(\max(\mathrm{Linear}(h_i), 0), 1). \tag{3}$$

WeSEE is then trained by minimizing the mean squared error (MSE) between the weak labels RD_i and the predicted labels $\hat{\mathrm{RD}}_i$:

$$\mathcal{L} = \frac{1}{|D|} \sum_{1}^{|D|} (\mathrm{RD}_i - \hat{\mathrm{RD}}_i)^2. \tag{4}$$

Up to now, we have loosely assumed that RD labels well represent the turn-level engagingness, and simply train WeSEE to predict RD labels. However, there might be some occasions where our assumption is violated. For example, in the PersonaChat dataset [39], the annotators creating the data were instructed to chat for 12–14 turns, which may have the problems of: (i) an unengaging turn appears in the beginning of the session but the conversation lasts long; and (ii) an engaging turn is not followed by a response because of the length restriction. Other high-quality datasets are created similarly in a lab environment [3, 11, 25]. If we fit the model merely on the weak RD labels, its prediction might not well correlate with human engagingness labels. To make sure that our model predicts engagingness rather than remaining depth, we use a small calibration set of dialogues annotated with engagingness labels at the validation phase. We save only the model parameters that peak on the Pearson correlation with engagingness labels. In this manner, our

model can use relatively few turn-level engagingness labels (that are expensive to obtain) only for validation, while being trained on RD labels that can be automatically generated from any dialogue dataset.

4 EXPERIMENTAL SETUP

We design our experiments to answer the following research questions: (RQ1): Are the RD labels predictable? (RQ2): When trained on the weak RD labels, how do the predictions produced by WeSEE correlate with human engagingness scores? (RQ3): How does each component, such as training on RD labels, regression formulation, different numbers of historical turns, and pooling method, contribute to the performance of WeSEE? And (RQ4): What can we learn by checking WeSEE's predictions?

4.1 Datasets

In order to infer the RD labels for training and validation, the datasets we use should have multiple turns in each dialogue session. We use the most popular English language open-domain dialogue datasets that meet this requirement: DailyDialog [DD, 11], PersonaChat [PC, 39], Empathetic Dialogues [ED, 25], Wizard of Wikipedia [WoW, 3], and BlendedSkillTalk [BST, 30]. We use only the dialogue text without other additional attributes, such as the persona descriptions in PC. Statistics for the datasets we use to train WeSEE are shown in Table 1. Since these datasets are relatively small and are different in style and average dialogue length, we combine them for training WeSEE to better generalize to real application scenarios. We note that although these datasets are created in a lab environment, there are still noticeable patterns of using engaging/unengaging responses as desired in the dialogue sessions. For example, dialogue participants tend to say greetings, start topics, or ask questions in the beginning of a conversation, and express farewells, use more generic responses towards the end of a dialogue. The ConvAI [CA, 14] dataset is only used for comparing the effects of weak RD labels and noisy human labels in Section 5.3 but not in our final model.

For ground-truth engagingness labels, we use the Fine-grained Evaluation of Dialog [FED, 17] and DailyDialog-Human [DD-H, 6] datasets, the only publicly available datasets that contain turn-level, Likert-scale engagingness labels annotated by human. We use DD-H (the smaller of the two datasets) as our validation set and FED as our test set. Both datasets contain 5 labels per turn with high inter-annotator agreement scores. We use the average of the 5 scores for each data sample as the ground truth for turn-level engagingness.

4.2 Baselines

For checking the predictability of RD labels, we compare WeSEE with the following methods: (i) a random baseline that randomly predicts a score between 0 and 1; (ii) an average baseline that uses the average dialogue length instead of $|D|$ in Eq. 1 for making predictions; (iii) the WeSEE-U model with the prediction layer **u**ntrained; and (iv) the WeSEE-S model that is trained using **s**huffled RD labels. For the task of explicitly predicting dialogue-turn engagingness we consider the following prior work as our baselines: FED-metric [17] and PredictiveEngagement (PredEnga) [6].[1] There are some models that were *not* proposed for explicit engagingness evaluation but that were reported to have a good correlation with human engagingness judgments [36], such as DialogRPT [5], USL-H [23] and DynaEval [38], which we also adopt as baselines.

[1]We also considered the approach proposed in [37] but excluded it from our evaluation due to difficulties in reproducing their results. Unfortunately, neither their implementation nor their trained checkpoints are available at the time of writing.

Table 1. Statistics for the datasets used in this paper.

DD:	Train	Val	Test
#Dialogues	11,118	1,000	1,000
#Turns total	87,170	8,069	7,740
#Turns avg	7.84	7.74	8.07
#Turns std	4.01	3.84	3.88
#Tokens	1,186,046	108,933	106,631
PC:	Train	Val	Test
#Dialogues	8,938	999	967
#Turns total	131,424	15,586	15,008
#Turns avg	14.70	15.60	15.52
#Turns std	1.74	1.04	1.10
#Tokens	1,534,258	186,055	176,903
ED:	Train	Val	Test
#Dialogues	17,780	2,758	2,540
#Turns total	76,609	12,025	10,941
#Turns avg	4.31	4.36	4.30
#Turns std	0.71	0.73	0.73
#Tokens	1,025,120	175,231	169,778
WoW:	Train	Val	Test
#Dialogues	18430	981	965
#Turns total	166,787	8,909	8,715
#Turns avg	9.05	9.08	9.03
#Turns std	1.04	1.02	1.02
#Tokens	2,730,760	145,995	142,896
BST:	Train	Val	Test
#Dialogues	4,819	1,009	980
#Turns total	54,881	11,467	11,154
#Turns avg	11.39	11.36	11.38
#Turns std	2.41	2.35	2.42
#Tokens	730,351	154,437	154,335
CA:	Train	Val	Test
#Dialogues	2,099	–	–
#Turns total	25,319	–	–
#Turns avg	12.06	–	–
#Turns std	9.44	–	–
#Tokens	171749	–	–

4.3 Metrics

To show the predictability of RD labels, we report the MSE, Pearson and Spearman correlation with the ground-truth RD labels for DD, PC, ED, WoW and BST. To compare with the baselines and evaluate a model's performance on the target task of turn-level engagingness prediction, we report the Pearson and Spearman correlation coefficients between the models' predictions and human annotations for FED and DD-H.

Table 2. Mean squared error (MSE) results (multiplied by 100) for predicting weak RD labels on the test sets for all datasets. Lower is better. Model weights are selected according to minimum MSE on the validation sets.

	DD	PC	ED	WoW	BST
Random	19.40	17.92	21.85	18.56	18.00
Average	5.02	0.14	2.86	0.80	0.79
WeSEE-U	35.71	32.04	40.50	38.15	38.61
WeSEE-S	10.94	9.47	13.42	10.38	9.98
WeSEE	7.22	5.81	6.10	6.96	9.89

4.4 Parameter settings and implementation

We chose the BERT base uncased model [1] as implemented in the Transformers library[2] as our turn encoder. The parameters for the linear projection layer of WeSEE are randomly initialized. The WeSEE model contains 109M trainable parameters (weights), in total. We select hyper-parameters using two different criteria, as described in the end of Section 3. We also evaluated (i) four alternative pooling methods, (ii) two activation functions mentioned in Section 3, and (iii) $k \in \{1, 2, 3, 4, 5\}$ for deciding upon the best history size. In our preliminary experiments, we trained the WeSEE model using an SGD optimizer with a learning rate (LR) chosen from the set $\{5e{-}2, 5e{-}3, 5e{-}4, 5e{-}5, 5e{-}6\}$, and found out that $5e{-}2$ works best according to the MSE loss on the validation set, and $5e{-}5$ works best when validated on DD-H. All WeSEE variants were trained for 50,000 steps. A fixed LR scheduler with 5,000 warmup steps was used. During training, we use a batch size of 20 and clip the gradient L2 norm to 0.1. The training finishes within 6 hours on a single TITAN Xp GPU with 5 history turns used as input (our computationally most intensive setting). For the single-turn model, in which only the current turn is used as input without any dialogue history, the training takes only 1.5 hours.

The source code to reproduce our experiments can be found at https://github.com/ShaojieJiang/lit-seq and the supplementary material. Our implementation is based on Hugging Face Transformers [34], PyTorch Lightning [33], and Hydra [35]. The data downloading and preprocessing steps are automatically taken care of in our training scripts, parameter settings included. Reproducing the best-performing model requires only a single line of code. Please refer to the README in the above link.

5 RESULTS AND ANALYSIS

In this section, we address our research questions. We first report on experiments to check the predictability of RD labels (RQ1). Then we see how well the model predictions can be used as engagingness scores, in terms of correlating with human annotations (RQ2). Then we report on ablation studies to understand how each component contributes to the model performance (RQ3). And, finally, we include a case study to show the interpretation of model predictions, as well as some error analysis (RQ4).

5.1 RQ1: Predictability of remaining depth

To answer RQ1, we report the MSE loss of predicting RD labels and calculate the correlation coefficients of predicted labels and ground truth labels.

5.1.1 Main findings. The MSE results and correlation with RD labels for WeSEE are shown in Table 2 and Table 3, respectively. Unsurprisingly, Random and WeSEE-U both perform badly on both MSE and correlation with RD labels.

[2] https://huggingface.co/transformers/model_doc/bert.html

Table 3. Correlation of model predictions with RD labels evaluated on the test sets. P: Pearson; S: Spearman. Results that are not statistically significant ($p\text{-}value < 0.05$) are in *italics*. Higher is better. Model checkpoints are the same as for Table 2.

	DD		PC		ED		WoW		BST	
	P	S	P	S	P	S	P	S	P	S
Random	*0.00*	*0.00*	*0.00*	*0.00*	*−0.01*	*−0.01*	*0.01*	*0.01*	*0.02*	*0.02*
Average	0.78	0.80	0.99	0.99	0.95	0.96	0.97	0.98	0.96	0.96
WeSEE-U	*−0.02*	−0.02	−0.05	−0.06	0.07	0.06	−0.04	−0.06	*0.01*	*0.00*
WeSEE-S	0.13	0.13	0.09	0.10	*0.00*	*0.01*	0.08	0.12	*0.01*	*0.01*
WeSEE	0.59	0.56	0.62	0.56	0.74	0.71	0.59	0.55	0.21	0.18

After training on normal RD labels, WeSEE achieves much lower MSE and high correlation coefficients on most datasets. On the other hand, WeSEE-S trained on the shuffled RD labels has much higher MSE than WeSEE and shows almost no improvement on correlation coefficients, which suggest that the shuffling breaks the meaningful correspondence of the textual content and the RD labels. The Average baseline achieves much lower MSE and higher correlation coefficients than WeSEE. This is due to the fact that Average baseline makes prediction in an *oracle* manner. However, as we discuss in Section 5.2, merely predicting RD labels is not helpful in a scenario that requires more content awareness, such as predicting engagingness. One reason is the noisy nature of RD labels. Semantic awareness is needed for learning meaningful patterns from the RD labels and denoising outliers. For example, in the training data we can sometimes observe short and generic responses (such as "I see. OK.") appear early in the dialogue. These messages are usually considered as unengaging responses by humans [27], thus not helpful with extended conversations. But in our weak labeling schema, they can be assigned with high RD values, which acts as noise. When we train WeSEE on RD labels, it learns to denoise by seeing more examples. Since WeSEE is trained to employ textual content to make predictions, and the generic responses are likely to be followed by fewer dialogue turns, WeSEE learns to assign lower values to them. There are presumably other types of noise as suggested by the correlation coefficients of WeSEE being lower than 1 in Table 3.

Among the datasets reported in Table 2 and 3, BST is an outlier. On BST, the MSE of WeSEE is almost identical to that of WeSEE-S. And in terms of correlation coefficients, WeSEE achieves Pearson correlation ≥ 0.59 and Spearman ≥ 0.55 on other datasets; on BST the coefficients are only 0.21 and 0.18, respectively. The level of noise of RD labels on BST is too high; indeed, in our preliminary experiments, we observed that training on BST with RD labels is detrimental to human correlation. Deeper investigation revealed that the BST dataset consists of human-machine dialogues [30]; machine generated messages are prone to be generic [27], which can result in more noisy RD labels according to our earlier analysis. There might be other reasons; we nevertheless exclude the BST dataset from our training data. For our experiments below, we train WeSEE by mixing the DD, PC, ED and WoW datasets together, to achieve better generalization.

5.1.2 Correlations with first and last turns. Next we consider only the first and last turns to determine whether we can observe different results. The WeSEE correlations with first and last k turns of each dialogue, compared to considering all turns is illustrated in Figure 3. WeSEE's predictions of the remaining depth tend to be more accurate closer to the beginning and the end of a dialogue session. By considering only the first and last k turns for each of the dialogues, we observe even higher correlations of the WeSEE predictions with the ground-truth RD labels. Figure 3 visualizes this effect in our data. When removing the predictions for intermediate turns, the correlation consistently increases. The first and last dialogue turns are often more similar across dialogues than the central part. People usually greet each

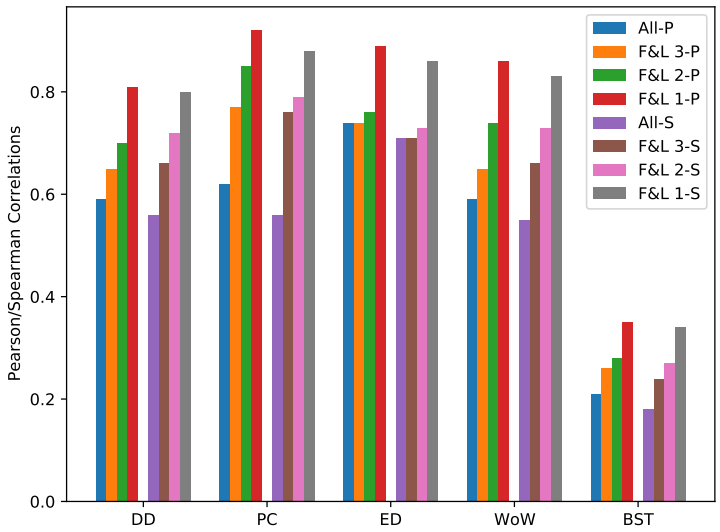

Fig. 3. WeSEE correlations with RD for all turns and first & last k (F&L k) turns only. -P: Pearson, -S: Spearman.

Table 4. Correlation between model predictions and human engagingness annotations. All correlation results that are not statistically significant (with $p\text{-}value < 0.05$) are *italicized*. Higher is better. Best results in each column are **bold faced**. WeSEE uses DD-H as the validation set.

	FED		DD-H	
	Pearson	Spearman	Pearson	Spearman
Average	*0.03*	*0.03*	–	–
FED-metric	0.16	0.18	0.23	0.27
DialogRPT	0.23	0.22	0.30	0.30
PredEnga	0.18	0.25	0.51	0.55
USL-H	0.24	0.26	0.55	0.56
DynaEval	0.25	0.26	*0.09*	*0.07*
WeSEE	0.29	0.33	**0.58**	**0.62**
WeSEE-H3	**0.36**	**0.38**	0.52	0.53

other and ask a few customary questions in the beginning of a dialogue, and say farewells and express gratitude at the end. WeSEE successfully captures these patterns, which are clearly very important to detect the user intent to continue or conclude the dialogue.

5.1.3 Upshot. From the experiments in this section, we can conclude that there are predictable patterns in high-quality dialogue datasets between the textual messages and our automatically generated RD labels. The pattern is even stronger around the beginning and ending of each dialogue session. When we use random shuffling to break potential patterns, or when the quality of the dataset is not good enough, the predictability of the RD labels is compromised as per the MSE scores and correlation coefficients.

5.2 RQ2: Predictability of dialogue engagingness

The correlation of WeSEE and baseline models with human engagingness annotations is reported in Table 4. Due to the noisy nature of RD labels, fitting WeSEE too well to RD labels can harm its ability for human correlation. We provide more insights in Section 5.3, but in this subsection, we calibrate WeSEE using human annotations by selecting the model weights with the highest correlation on the DD-H dataset, effectively using DD-H as a validation set. All baseline results are reproduced by us using their official source code and trained model weights to ensure a fair comparison.

5.2.1 Main findings. Utilizing heuristics to accurately predict RD labels, as done by the Average baseline, does not yield a good correlation with human engagingness scores; see Table 4. This indicates that the RD signal is *not equal* to turn-level engagingness, which is why we only treat RD as a weak supervision signal. Besides, we cannot use the Average baseline on datasets with a fixed number of history turns such as DD-H. WeSEE trained to use only a single dialogue turn outperforms all baseline methods on the FED and DD-H datasets, w.r.t. Pearson and Spearman coefficients. When using 3 history turns, WeSEE-H3 performs even better on FED with a slight decrease on DD-H. This is because DD-H has only two turns for each annotation, therefore, WeSEE-H3 trained with a longer history does not help to improve the performance on this dataset. The best-performing WeSEE outperforms the second best baseline models by 0.11 (0.12) of Pearson (Spearman) on the FED dataset, and 0.03 (0.06) of Pearson (Spearman) on the DD-H dataset. However, we note that although our approach performs the best, its performance is still far from the conventional definition for a "high" correlation. Similar observations are also reported by other work on evaluation metrics in the dialogue task, where a typical correlation is around 0.2–0.5 [6, 7, 10, 15, 17].

Although the FED-metric relies entirely on the pretrained DialoGPT, which smartly avoids training, it performs poorly on both datasets. Our reproduced results for the FED-metric on the FED dataset are different from the original work [17], but are consistent with later work [36]. The reason for its poor performance is due mainly to the underlying DialoGPT model, which is trained on Reddit data, which is quite different from real conversations in style. This is supported by DialogRPT, another model relying on DialoGPT as well as being trained on Reddit data. Compared to PredEnga and USL-H, which are trained on real dialogue data, DialogRPT has a much worse performance on the DD-H dataset. Since DialogRPT is trained on the depth information of Reddit comments, which is similar to our RD labels, it performs better than the FED-metric, especially on the FED dataset. Because DialogRPT also relies on other features (for example, the width and up-/down-votes of user comments), none of which are common in real dialogue data, DialogRPT only achieves mediocre performance on both datasets. In contrast, WeSEE is trained on dialogue data and uses RD as weak labels for engagingness. RD labels have an intuitive connection with engagingness, thus serving as a main contributing factor to WeSEE's superior performance. In Section 5.3 we show that WeSEE trained on RD labels shows higher human correlation than when trained on some noisy human engagingness annotations.

PredEnga and USL-H have a similar performance on both datasets. Both are BERT-based models, trained on dialogue data, and rely on binary classification except that USL-H also utilizes a BERT-MLM score. Training as a classifier loses much fine-grained information such as different engagingness levels, which restricts their ability for engagingness prediction. We train WeSEE as a regression model, allowing it to capture subtle differences of RD labels. Our ablation study in Section 5.3 shows that this regression formulation is more suitable than classification with RD labels.

DynaEval outperforms other baseline models on FED. DynaEval is trained on dialogue datasets (i.e., ED, ConvAI2 [2] and DD), and is able to make use of the graph structure of dialogue turns from the same dialogues. Due to this second aspect, DynaEval is not applicable to the datasets that do not containing dialogue sessions, which explains its poor performance on DD-H. The main reason for DynaEval's inferior performance on the FED dataset compared to WeSEE

Table 5. Model performances when using only a single dialogue turn. All correlation results that are not statistically significant (with $p\text{-}value < 0.05$) are *italicized*. Higher is better. Best results in each column are **bold faced**. WeSEE uses DD-H as the validation set.

	FED		DD-H	
	Pearson	Spearman	Pearson	Spearman
FED-metric	*0.09*	0.12	0.12	0.14
DialogRPT	0.23	0.32	**0.58**	0.59
PredEnga	0.13	0.26	0.46	0.59
DynaEval	*−0.07*	*−0.06*	0.17	0.19
WeSEE	**0.29**	**0.33**	**0.58**	**0.62**

is that it was not trained on engagingness labels. Acquiring enough high-quality engagingness (class) labels is itself a difficult task, while WeSEE circumvents this problem with weak supervision.

5.2.2 Single turn input. All baseline approaches need multiple dialogue turns as input. To understand how they perform when only a single turn is given, we compare their performance in Table 5. Most baseline approaches experience significant performance drops on the FED and DD-H datasets; USL-H does not work in this setting due to its requirement for the dialogue context. DialogRPT sees a performance increase, especially on the DD-H dataset. We hypothesize that this is because DialogRPT uses the transformer output for the last token as the utterance representation. In batch processing (padding tokens added to the left), this shifts the positional ids of shorter utterances in the batch to the right, which causes inaccurate predictions. When more dialogue turns are used, the shifting effect increases, hence deteriorating the prediction. WeSEE does not suffer from this problem, as we use mean pooling of all tokens excluding padding tokens as the turn representation.

5.2.3 Upshot. When calibrated on human engagingness scores, WeSEE trained on noisy RD labels can be prevented to fit to the data noise. Thanks to the relationship between RD labels and engagingness, and the calibration using human engagingness annotation, WeSEE achieves the new state-of-the-art on dialogue engagingness prediction. Since WeSEE is trained to utilize the language understanding ability of BERT, WeSEE can still perform well when only single-turn dialogue texts are provided.

5.3 RQ3: Ablation study

We ablate the core components of WeSEE to better understand their impact on the overall performance; see Table 6. These components are: (i) training on RD labels; (ii) regression formulation instead of classification; (iii) history size; and (iv) pooling methods. For ease of reference, at the top of the table we repeat the performance of WeSEE trained with a single turn, mean pooling, and with model weights selected according to the best performance on DD-H (i.e., used as a validation set).

5.3.1 Training on RD labels. Table 3 shows that WeSEE-S trained with shuffled RD labels performs poorly. In the -Shuffle row of Table 6, we confirm this using correlation with human annotations. Thus, although RD labels are used as noisy engagingness labels, there is useful information for training an engagingness evaluator. Due to the noisy nature of RD labels, we cannot totally rely on them for training WeSEE. As can be seen from the -ValLoss row, if we allow WeSEE to fit well on RD labels, it achieves sub-optimal correlation with human engagingness labels. To provide another angle of how noisy RD labels can be, we calculated their correlation with human engagingness annotations on the FED dataset; the results are -0.03 Pearson and -0.01 Spearman, both are not statistically significant. This does not

Table 6. Ablation study results. Correlation results that are not statistically significant ($p\text{-}value < 0.05$) are *italicized*. Higher is better.

	FED		DD-H	
	Pearson	Spearman	Pearson	Spearman
WeSEE	0.29	0.33	0.58	0.62
-Shuffle	*0.09*	*0.08*	−0.15	−0.14
-ValLoss	0.26	0.28	0.35	0.34
-FT-CA1	0.29	0.33	0.51	0.53
-FT-CA3	0.37	0.39	0.46	0.48
-SC-CA1	0.27	0.32	0.54	0.59
-SC-CA3	0.36	0.37	0.43	0.45
-Class2	*0.07*	*0.05*	*0.07*	*0.06*
-Class5	0.13	0.12	*−0.01*	*−0.02*
-Class10	0.15	0.16	0.13	*0.10*
-H2	0.35	0.38	0.52	0.53
-H3	0.36	0.38	0.52	0.53
-Flat-H2	0.33	0.35	0.51	0.53
-Flat-H3	0.32	0.33	0.51	0.53
-cls	0.23	0.22	0.41	0.41
-max	0.37	0.37	0.35	0.35
-min	0.25	0.29	0.25	0.26

mean that RD labels are useless, as the FED dataset has only 375 annotated examples. The positive correlation of the -ValLoss experiment confirms the value of using RD labels as a weak engagingness supervision signal. To understand the importance of training on RD labels, we trained/fine-tuned WeSEE on the engagingness labels of the CA dataset; see the -SC-CA* (training from scratch) and -FT-CA* (fine-tuning) rows. The CA dataset contains 1 human engagingness annotation for each dialogue participant in a session of human-bot dialogue, which we use as turn-level engagingness labels following Ghazarian et al. [6]. During training/fine-tuning WeSEE on the CA dataset, we also used DD-H as the validation set. As shown in Table 6, WeSEE trained on CA with 1 (-CA1) or 3 (-CA3) turns performs worse than trained only on RD labels, suggesting that weak RD labels are more useful than low-quality human engagingness labels for training WeSEE.

5.3.2 Regression instead of classification. Next, to see the importance of our regression formulation, we modify WeSEE to be a classifier, and map the RD labels to (i) binary labels $\{0, 1\}$ using a threshold 0.5, (ii) 5 class labels using thresholds of $\{0.2, 0.4, 0.6, 0.8\}$, and (iii) 10 class labels using thresholds of $\{0.1, 0.2, \ldots, 0.9\}$. Then we train the modified WeSEE classifiers with Cross Entropy loss. The results in the -Class* rows (Table 6) show that, although this classification formulation shows some positive correlation especially with finer-grained label buckets, the correlation is much weaker than the WeSEE regression model. This suggests that our formulating the engagingness prediction as a regression task is more suitable than a classification formulation.

5.3.3 History size. By training and testing WeSEE with more than one historical turn (-H* rows, in Table 6), we observe that the single-turn WeSEE model (top row) performs the best on DD-H, while -H3 with 3 dialogue turns performs the best on FED. Using more than 3 turns showed similar results as -H3. We design WeSEE to encode each dialogue turn separately to preserve the speaker information. To see how this design influences the prediction, we also consider

	Single-turn text	WeSEE-H1
1.	hey!. nice to meet you. me and my folks are currently in arkansas. you?	1.00
2.	hello, where can i buy an inexpensive cashmere sweater?	1.00
3.	hello there, how are you today?	1.00
4.	my dear, what's for supper?	1.00
5.	hi buddy, what you think about cinematography	1.00
6.	where'd you get those?	0.82
7.	i like to run, create art, and take naps! how about you?	0.80
8.	i love italian cuisine	0.56
9.	jeez! its so unfortunate... very sad really.	0.50
10.	it has 10 provinces	0.42
11.	thanks for all your help / info today	0.38
12.	well you sleep well goodnight	0.00
13.	i wish you the best of luck, you will be fine!	0.00
14.	thank you, bye - bye.	0.00
15.	thank you. good luck to your son	0.00

Fig. 4. Successful cases of WeSEE-H1. Only single turns sampled from the datasets listed in Section 4 are displayed here. The turns are ordered according to the predicted scores.

using *flat* history by concatenating history dialogue turns into one utterance, with separator tokens to indicate the switch of speaker. Their performance for using 2 and 3 turns is shown in the -Flat-H* rows. Using flat history performs consistently worse; the difference between using more dialogue turns is bigger as can be seen from the FED results on -Flat-H3 and -H3. This is because when using flat history, messages from different speakers get mixed during the encoding process, making the prediction more difficult.

5.3.4 Pooling method. The last three rows in Table 6 show that using *cls*, *max* or *min* pooling (with 3 dialogue turns) negatively influences performance on the DD-H dataset, which is also true on the FED dataset except that max pooling shows no noticeable difference.

5.3.5 Upshot. The most important lesson we learn from this set of experiments is the effect of RD labels. Compared to low-quality human engagingness annotation, RD labels are much cheaper (almost free) to acquire, but are still more useful for training an engagingness predictor. Other experiments justify our design of the best-performing model, i.e., using three history turns, with each turn independently encoded and pooled by mean pooling, trained as a regression task and calibrated on human engagingness annotation.

5.4 RQ4: Result analysis

In this section, we list several case studies of the single-turn WeSEE model selected according to minimum validation loss.

5.4.1 Successful single-turn examples. Figure 4 shows some representative good examples. It shows that WeSEE gives highest scores to dialogue starters and lowest scores to dialogue endings. With the content shifts from greetings to questions and statements, and then to farewells, the WeSEE model can accurately detect the dialogue progress: the lower the prediction, the nearer towards the end. We observe such interesting patterns from more examples: WeSEE is most accurate with clear greetings and farewells, and usually gives an inquisitive utterance a high score; it is often the case when an utterance starts a new topic, WeSEE predicts longer conversations will happen. There may be other

Dialogue turns	RD	WeSEE-H1	WeSEE-H3
1. is there anything else i can do for you?	0.08	0.66	0.19
2. that's ok.	0.00	0.35	0.17
3. it'll be worth it in the end. just think of the freedom you'll have!	0.29	0.02	0.48
4. enjoy your visit and safe travels.	0.53	0.00	0.57
5. i like the sound of that	0.56	0.16	0.39
6. thank you.	0.62	0.11	0.40
7. yes, you did.	0.73	0.17	0.49

Fig. 5. Cases in which WeSEE-H1 deviates from the RD labels and WeSEE-H3 aligns better. Only single turns sampled from the datasets listed in Section 4 are displayed here.

Dialogue turns	Human	WeSEE-H1
1. everything is going extremely well. how are you?	0.90	0.89
2. what is the meeting about?	0.80	0.76
3. try me. what is your problem?	1.00	0.61
4. not that much more, no.	0.40	0.27
5. i did not want to hear that now	0.80	0.33

Fig. 6. WeSEE-H1 predictions versus human annotations from the FED dataset.

interesting patterns that are less obvious to discover or more complicated to describe. We will release the annotated files for all the test sets we use in this paper.

5.4.2 Single-turn failure examples. There are also some tricky cases that the single-turn WeSEE model fails to cope with. One biggest type of such errors usually happen on generic utterances, such as the 2nd, 6th and 7th examples shown in Figure 5. While we can argue that many generic responses fit naturally in the end of a conversation, it takes longer context and heavier reasoning to decide whether the conversation actually dies. Indeed, our best-performing WeSEE-H3 using 3 turns of history can make more accurate predictions in such cases, however, the overall predictions from the -H3 model is less comprehensible than the -H1 model. We also note that there are cases that are easy for us to decide in real-life. For example, a "Thank you." together with a leaving body-language clearly shows that the conversation is ending. In the purely textual setting, without additional signals, this is sometimes impossible to accurately predict. There is another tendency that the WeSEE model responds too much to questions, such as the first example in Figure 5. While the utterance itself already shows a good sign of conversation ending, the single-turn WeSEE model thinks it is a normal question and predicts a medium score for it. This is improved by using longer contexts as can be seen from the WeSEE-H3 prediction.

5.4.3 Comparison with human annotations. Comparisons with human annotations from the FED dataset are shown in Figure 6. In many cases, our model's prediction correlates well with human annotations (normalized to $[0, 1]$), and there are also some cases that WeSEE makes arguably better predictions than human annotations, such as the last example when the participant is trying to end the conversation/topic, but human annotators still think it is engaging.

5.4.4 A full example. We also show a randomly-chosen complete dialogue from the DD dataset in Figure 7, from which we can see that our WeSEE model can not only detect when the conversation starts and ends, but also reflects where the conversation can end prematurely, such as the 5th and 7th rows.

5.4.5 Upshot. We summarize the main insights gained from the case studies presented in this section: (i) WeSEE can distinguish conversation starters and endings by assigning higher scores to the former and lower scores to the latter.

	Dialogue	WeSEE-H1
1.	what can i do for you today?	1.00
2.	i have a question.	1.00
3.	what do you need to know?	0.64
4.	i need to take the driver's course. how many hours do i need?	0.85
5.	it depends on what you're trying to do with the completion of the course.	0.21
6.	i need to get my license.	1.00
7.	you're going to need to complete six hours.	0.42
8.	how many hours a day can i do?	0.62
9.	you can do two hours a day for three days.	0.43
10.	that's all i need to do to finish?	0.37
11.	yes, that's all you need to do.	0.17
12.	thanks. i'll get back to you.	0.00

Fig. 7. A complete dialogue randomly sampled from the DD dataset and labeled by WeSEE-H1.

This does not mean that WeSEE is only responsive to conversation starters and endings. A closer analysis where we split WeSEE's predictions into three buckets, representing the conversation *starter*, *middle* and *ending*, reveals that the predictions fall into these three buckets for 24.5%, 57.6% and 17.8% of the times, respectively. This is expected, as the middle of a dialogue is usually the most content-rich and dynamic section. (ii) When an utterance contains a question, starts a new topic, or becomes more detailed, WeSEE usually assigns a higher score, which concurs with the identified factors that facilitates engagingness [26, 27]. (iii) WeSEE struggles to predict correct labels for short and uninformative responses, and questions that terminate the conversation (for example, "Anything else I can do?"). This is probably due to the data bias in the dataset used, because a turn containing questions is usually engaging and appears in the early stage of a dialogue session. We expect that adding some hard negatives to the training data can alleviate this problem, although it is not yet clear to us how to effectively mine such negative examples.

6 CONCLUSION

We have studied the problem of predicting turn-level dialogue engagingness and proposed a novel approach, *Weakly Supervised Engagingness Evaluator* (WeSEE), for this task. Using *remaining depth* (RD) labels for weak supervision is the main novelty of the proposed approach. We formulate the engagingness prediction problem as a regression task using the automatically generated RD labels. This formulation allows us to take advantage of the implicit signals in multi-turn dialogue data because RD can be deduced automatically. We can use any multi-turn dialogue dataset for training our predictive model.

When trained on a mixture of four popular dialogue datasets, the proposed WeSEE model with a single dialogue turn already outperforms existing approaches, establishing the new state-of-the-art performance on the FED and DD-H datasets. When using three history turns, WeSEE-H3 achieves the highest performance on FED, but lower on the DD-H dataset. We hypothesize that this is due to DD-H's having only two turns for each data point, which is too short for WeSEE-H3.

Human curated dialogue datasets have been shown to be very important sources for training well-performing open-domain dialogue models [26]. Our work confirms this point. Our work underlines the value of human-curated dialogue datasets for as carriers of turn-level engagingness signals in the number of turns in each dialogue session. Making use of such meaningful signals can save us from expensive human annotations and, more importantly, help us understand and improve models for automatic engagingness evaluation and prediction.

Our proposed WeSEE model has several limitations. It does not handle well some generic messages that need long contexts or several modalities to decide the engagingness. It is currently also over responsive to questions, which we expect can be alleviated by adding hard negatives to the training data.

The WeSEE model developed in this work can be applied to evaluate engagingness of conversational systems, or serve as a ranker for selecting more appropriate candidate responses. Further study needs to be done for checking how well WeSEE can cope with such tasks. We also note that engagingness is not the only gold measurement one should optimize for open-domain conversational systems. In the future, more work needs to be done to combine WeSEE with evaluation metrics focusing on other aspects, such as coherence, specificity and consistency, etc.

ACKNOWLEDGMENTS

This research was supported by the China Scholarship Council, and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

A ETHICAL CONSIDERATIONS

All the training/validation/test data used in this work is publicly available. To the best of our knowledge, the creators of these data sets have taken ethical issues into consideration when creating the data sets. We manually checked some predictions from WeSEE, and did not observe any noticeable traces of concern, such as scoring biased or rude utterances high. The WeSEE models are trained on English, open-domain dialogue data. Therefore, we are not yet clear whether unexpected predictions may appear when WeSEE is used on other tasks/languages. We share our source code and trained model weights to support its correct use. However, we note that when incorrectly used, such as training the WeSEE model to rank discriminative utterances high, it may also pose harm to users of conversational applications into which WeSEE is integrated.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] 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 (2019).

[3] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-powered Conversational Agents. *arXiv preprint arXiv:1811.01241* (2018).

[4] Tamás Fergencs and Florian Maximilian Meier. 2021. Engagement and Usability of Conversational Search – A Study of a Medical Resource Center Chatbot. In *Proceedings of iConference 2021 (LNCS 12645)*. Springer.

[5] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 386–395.

[6] Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 7789–7796.

[7] Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, 379–391.

[8] Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. 2018. Alexa Prize - State of the Art in Conversational AI. *AI Mag.* 39, 3 (2018), 40–55.

[9] Mounia Lalmas, Heather L. O'Brien, and Elad Yom-Tov. 2014. Measuring User Engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 6, 4 (2014), 1–132.

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

[11] 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, IJCNLP 2017, Taipei, Taiwan, November 27-December 1, 2017 - Volume 1: Long Papers*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, 986–995.

[12] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the intrinsic information flow between dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

[13] Weixin Liang, Kaihui Liang, and Zhou Yu. 2021. HERALD: An Annotation Efficient Method to Detect User Disengagement in Social Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3652–3665.

[14] Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Polulyakh, and Aleksandr Seliverstov. 2018. ConvAI Dataset of Topic-oriented Human-to-chatbot Dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 47–57.

[15] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1116–1126.

[16] Xiaojuan Ma. 2018. Towards Human-Engaged AI. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 5682–5686.

[17] Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 225–235.

[18] Shikib Mehri and Maxine Eskénazi. 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, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 681–707.

[19] Heather L. O'Brien. 2016. Theoretical Perspectives on User Engagement. In *Why Engagement Matters: Cross-Disciplinary Perspectives and Innovations on User Engagement with Digital Media*. Springer, 1–26.

[20] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human Computer Studies* 112 (2018), 28–39.

[21] 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, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 3619–3629.

[22] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..." – Conversational Strategies for Product Search Assistants. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. ACM.

[23] 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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 4164–4178.

[24] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[25] 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5370–5381.

[26] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19-23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 300–325.

[27] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar

Solorio (Eds.). Association for Computational Linguistics, 1702–1723.

[28] Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2430–2441.

[29] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding User Satisfaction with Task-Oriented Dialogue Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 2018–2023.

[30] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2021–2030.

[31] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2021. A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. *ACM Transactions on Information Systems* 39, 4 (August 2021), Article 49.

[32] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On Evaluating and Comparing Open Domain Dialog Systems. *arXiv preprint arXiv:1801.03625* (2018).

[33] Falcon William and The PyTorch Lightning team. 2019. PyTorch Lightning. https://github.com/PyTorchLightning/pytorch-lightning.

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 38–45.

[35] Omry Yadan. 2019. Hydra - A Framework for Elegantly Configuring Complex Applications. https://github.com/facebookresearch/hydra.

[36] Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. *CoRR* abs/2106.03706 (2021).

[37] Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, Kees van Deemter, Chenghua Lin, and Hiroya Takamura (Eds.). Association for Computational Linguistics, 65–75.

[38] 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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5676–5689.

[39] 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2204–2213.

[40] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, Asli Çelikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 270–278.