Weakly Supervised Turn-level Engagingness Evaluator for Dialogues

Anonymous ACL submission

Abstract

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 (RD) 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* (FED) dataset (0.38 Spearman) and the *DailyDialog* dataset (0.62 Spearman).

1 Introduction

Engagingness is an important aspect of an open-domain dialogue system. It reflects user satisfaction with the dialogue system (Yi et al., 2019). At the turn level, it also measures how willing the user is to continue the conversation. Engagingness is typically measured using the *Conversation Turns Per Session* (CTPS) since engaging conversations tend to have more turns than less engaging ones (Venkatesh et al., 2018; Khatri et al., 2018). CTPS values can be easily obtained off-line to compare engagingness levels of different systems. However, performing an online turn-level engagingness evaluation is of even greater importance since it can be also used to guide the dialogue generation process directly or to choose between different candidate responses (Yi et al., 2019). Fig. 1 provides an example of turn-level engagingness evaluation.

Engaging responses:
C: Yes yes. I've been to Tokyo as well. It's so nice!
R: What did you do here?   WeSEE: 1.00
...
C: no not yet she is a little to young
R: i love to go running, its my favorite passion. i wish i ran for a living   WeSEE: 0.93

Not Engaging responses:
C: of course. a two - week paid vacation a year, a five - day workweek.
R: good. anything else?   WeSEE: 0.32
...
C: what am i supposed to do then?
R: a good rest is all you need, and drink more water. i'll write you a prescription.   WeSEE: 0.18

Figure 1: An illustration of turn-level engagingness evaluation. The predictions of our model (WeSEE) for the responses are at the end of each response line, in the range of $[0, 1]$. **C:** context; **R:** response. Examples are from the DD, PC and FED datasets. For more examples, see Fig. 6 in the Appendix.

Recent work has focused on training neural models to predict turn-level engagingness (Yi et al., 2019; Ghazarian et al., 2020; Gao et al., 2020; Mehri and Eskénazi, 2020a), which is an important step towards online evaluation of dialogue system performance. However, existing approaches exhibit important limitations. E.g., the most common approach is to address engagingness prediction as a binary classification task (Yi et al., 2019; Ghazarian et al., 2020). The main reason is the need for human labels for training the models. While labelling turns as engaging or non-engaging is conceptually simple, the approach lacks scalability. The produced binary labels may also not sufficiently well reflect differences between engagingness levels. As a reasonable and scalable alternative, we propose a simple approach of using weak supervision for the engagingness evaluation. Our experiments show that this approach has better correlation with human judgements of engagingness than previously proposed approaches. Importantly, we only study the engagingness evaluation for open-domain dialogue systems, not for task-oriented dialogue systems; task-oriented dialogue systems are usually optimised for quick task completion, and having an engaging system there can mean a negative thing.

We first use the *remaining depth* (RD) as heuristic weak labelling for turn-level engagingness; RD is defined as the number of conversation turns 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. Therefore, RD labels can serve as noisy indicators of engagingness, and can be easily obtained for existing dialogue data, which saves extra annotation efforts. Second, we show that this weak signal can be used to train a BERT- based (Devlin et al., 2018) regressor to be an engagingness evaluator and achieve state-of-the-art correlation with human engagingness judgments on two dialogue datasets. *Weakly Supervised Engagingness Evaluator* (WeSEE) can not only output real numbers that reflect fine-grained engagingness levels, but it can also use single-turn text data to make predictions, thus making it broadly applicable.

In our experiments, we calculate the Pearson and Spearman correlations of WeSEE predictions and human annotations. WeSEE achieves Pearson and Spearman coefficients of 0.36 and 0.38, respectively, on the Fine-grained Evaluation of Dialog (FED) dataset (Mehri and Eskénazi, 2020a), and 0.58 and 0.62 on the DailyDialog-Human dataset (Ghazarian et al., 2020), which is the new state-of-the-art performance on both datasets.

Main contributions. The main contributions of this paper are: (1) We propose to use RD as weak labels for turn-levsel engagingness, which avoids the need for explicit human annotations. (2) We formulate engagingness prediction as a regression task, therefore, the predicted scores can distinguish different magnitudes of engagingness. (3) We show that a BERT-based model can already have decent predictions with only single dialogue turns, while using more turns can correlate better with human annotation. (4) We share our source code, datasets used, implemented baselines and trained parameters at https://anonymous.4open.science/r/WeSEE.

2 Related Work

We start by providing a summary of the state-of-the-art in automatic dialogue evaluation. After that, we outline the main limitations related to measuring dialogue engagingness that motivate our work.

Dialogue quality is a multi-faceted phenomenon and cannot be evaluated along a single dimension (See et al., 2019; Phy et al., 2020; Yeh et al., 2021). However, most evaluation approaches proposed to date evaluate either the overall dialogue quality or the response quality on the turn-by-turn level (Yi et al., 2019; Pang et al., 2020; Li et al., 2021; Sinha et al., 2020; Mehri and Eskénazi, 2020b,a; Zhang et al., 2021; Phy et al., 2020; Gao et al., 2020). Being versatile also means sacrificing performance as well as interpretability with respect to the individual aspects of the dialogue quality, such as dialogue engagingness (Yeh et al., 2021). Our experiments show that such general-purpose quality evaluators do not achieve a high correlation with manually-labeled engagingness scores.

Engagingness evaluation is studied less than overall dialogue quality evaluation. The few approaches that exist have several drawbacks. First, training supervised models that predict engagingness requires manual labels, which are difficult to obtain (Yi et al., 2019; Ghazarian et al., 2020). Second, defining annotation guidelines for measuring dialogue engagingness has proved to be a hard task. For example, Yi et al. (2019) resorted to binary labels (engaging/non-engaging) that are easier to acquire but are not very descriptive. Ghazarian et al. (2020) 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.

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

To the best of our knowledge, the only other approach to engagingness prediction that does not require human engagingness annotations is due to Mehri and Eskénazi (2020a). They use the log-likelihood of a curated pool of the follow-up utterances produced by DialoGPT (Zhang et al., 2020) as their engagingness scores. Log-likelihood is not bounded and changes with utterance length. In contrast, the normalised WeSEE scores fall in the range $[0, 1]$ and allow one to compare the engagingness of candidate responses of different lengths.

3 Our Approach: Engagingness Evaluator Trained on Weak Labels

We use $D_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,n}]$ to represent the i-th dialogue session in the dataset that has up to n turns, with one turn denoting the message from one speaker at a time. Consecutive messages from the same speaker are merged into 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 j may consist of up to m tokens: $X_{i,j} = [x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,m}]$.

The *remaining depth* (RD) of $X_{i,j}$ normalised to $[0, 1]$ is calculated as:

$$\mathrm{RD}_{i,j} = \frac{n - j}{n - 1}, \quad (1)$$

which we subsequently use in place of the ground-truth engagingness label (that is, as a weak supervision signal) when formulating the RD prediction problem as a regression task. Thereby, each pair $(X_{i,j}, \mathrm{RD}_{i,j})$ is treated as a single data point for training the prediction model.

Our WeSEE model is based on BERT as illustrated in Fig. 2. The dialogue turns are embedded with BERT and then averaged for making the predictions. More concretely, we first use the pre-trained BERT model (Devlin et al., 2018) to get a vector representation of the turn $X_{i,j}$. To use the context available from the dialogue history, we also embed up to $k \geq 0$ turns that occurred before the j-th turn in the same i-th dialogue:

$$h_{i,j} = \mathrm{Mean}(\,\mathrm{BERT}(X_{i,j}), \mathrm{BERT}(X_{i,j-1}), \ldots, \mathrm{BERT}(X_{i,j-k})), \quad (2)$$

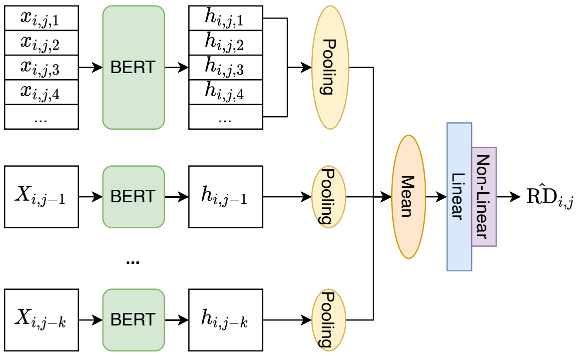

Figure 2: WeSEE model architecture.

where Mean denotes mean pooling and $h_{i,j} \in \mathbb{R}^{hid\_sz}$ is a hid_sz-dimensional contextualised vector representation for turn $X_{i,j}$. Thus, hid_sz is a hyper-parameter that determines the hidden size of our BERT-based turn embeddings. The representation for each turn $\mathrm{BERT}(X_{i,j})$ is a vector obtained by pooling the BERT positional outputs. We evaluate 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 $h_{i,j}$ to a scalar as the predicted engagingness level and use a simple cut-off to normalise it to $[0, 1]$ range:

$$\hat{\mathrm{RD}}_{i,j} = \min(\max(\mathrm{Linear}(h_{i,j}), 0), 1). \quad (3)$$

WeSEE is then trained by minimising the Mean Squared Error (MSE):

$$\mathcal{L}_{i,j} = (\mathrm{RD}_{i,j} - \hat{\mathrm{RD}}_{i,j})^2. \quad (4)$$

Up to now WeSEE is just trained to predict RD labels, which is not sufficient to predict turn-level engagingness (see Section 5.3). To make sure that our model predicts engagingness rather than remaining depth, we use a small set of dialogues annotated with engagingness labels only at the validation phase. We save only the model parameters that peak on the Pearson correlation with engagingness labels. Thereby, our model can use relatively few turn-level engagingness labels (that are expensive to obtain) only for validation and test, 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): How do the predictions produced by WeSEE, when trained on the weak RD labels, correlate with human engagingness scores? (RQ3): How does each component, such as training on RD labels, regression formulation, different numbers of historical turns, pooling method, contribute to the performance of WeSEE? (RQ4): What can we learn by checking WeSEE's predictions?

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 open-domain dialogue datasets in English that meet this requirement: DailyDialog (DD, Li et al., 2017), PersonaChat (PC, Zhang et al., 2018), Empathetic Dialogues (ED, Rashkin et al., 2019), Wizard of Wikipedia (WoW, Dinan et al., 2018), and BlendedSkillTalk (BST, Smith et al., 2020). We use only the dialogue text without any additional attributes, such as persona descriptions in PC. Since these datasets are relatively small (see Appendix A.2 for statistics of the datasets), and are different in style and average dialogue length, we combine them for training WeSEE to better generalize across different dialogues.

For ground-truth engagingness labels, we use FED (Mehri and Eskénazi, 2020a) and DailyDialog-Human (DD-H, Ghazarian et al., 2020), the only publicly available datasets that contain turn-level engagingness labels produced by human annotators. 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.

Baselines. For checking the predictability of RD labels, we compare WeSEE with the following methods: (1) Random baseline that randomly predicts a score between 0 and 1; (2) Average baseline that uses the average dialogue length in stead of n in Eq. 1 for making predictions; (3) WeSEE-U model with the linear layer **u**ntrained; and (4) 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:[1] FED-metric (Mehri and Eskénazi, 2020a) and PredictiveEngagement (PredEnga) (Ghazarian et al., 2020). There are some models that were *not* proposed for explicit engagingness evaluation but were reported to have a good correlation with human engagingness judgements (Yeh et al., 2021), such as DialogRPT (Gao et al., 2020), USL-H (Phy et al., 2020) and DynaEval (Zhang et al., 2021), which we also adopt as baselines.

[1]The approach proposed in (Yi et al., 2019) was excluded from the evaluation due to the difficulties in reproducing their results. Neither their implementation nor their trained checkpoints are available at the time of writing.

	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

Table 1: 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.

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 baseline and evaluate the model performance on the target task of turn-level engagingness prediction, we report the Pearson and Spearman correlations between the models' predictions and human annotations for FED and DD-H.

5 Results and Analysis

5.1 RQ1: Predictability of Remaining Depth

The MSE results and correlation with RD labels for WeSEE are shown in Table 1 and Table 2, respectively. Below are our observations. Unsurprisingly, Random and WeSEE-U both perform badly on both MSE and correlating with RD labels. Although WeSEE-S trained on shuffled RD labels manages to reduce MSE, it shows almost no improvement on correlation coefficients. After training on normal RD labels, WeSEE achieved much lower MSE and high correlation coefficients on most datasets. These comparisons indicate that there are underlying patterns between textual content and the RD labels, which can be captured by WeSEE. The Average baseline achieves much lower MSE and higher correlation coefficients than WeSEE. This is due to the fact that Average does not consider the actual content of dialogue turns, but instead makes prediction only using the progress of a given dialogue and the expected total number of turns. As

	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

Table 2: 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 the same as for Table 1.

we will soon discuss in §5.2, accurately 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. E.g., 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 (See et al., 2019), 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 omit some of the noise. Since WeSEE is trained to employ textual content to make predictions, and the generic responses are likely to be followed by fewer dialogue turns, we observed that WeSEE learns to assign lower values to them. There are presumably other types of noise; they prevent the correlation coefficients of WeSEE in Table 2 from being exact 1.

Among the datasets reported in Table 1 and 2, 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. The BST dataset consists of human-machine dialogues (Smith et al., 2020); machine generated messages are prone to be generic (See et al., 2019), 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 dataset mixture. For our experiments below, we train WeSEE by mixing the DD, PC, ED and WoW datasets together, to achieve better generalisation.

	FED		DD-H	
	P	S	P	S
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

Table 3: Correlation between model predictions and human engagingness annotations. P: Pearson; S: Spearman. All correlation results that are not statistically significant (with $p\text{-}value < 0.05$) are *italicised*. Higher is better. Best results in each column are **bold faced**. WeSEE uses DD-H as the validation set.

5.2 RQ2: Predictability of dialogue engagingness

The correlation of WeSEE and baseline models with human engagingness annotations is reported in Table 3. 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 §5.3, but in this subsection, we select WeSEE model weights with the highest correlation on 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.

Utilising heuristics to accurately predict RD labels, as done by the Average baseline, does not yield a good correlation with human engagingness scores; see Table 3. 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 correlations. 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. This is also reported by other works for other evaluation metrics, which typically see a correlation around 0.2-0.5 (Mehri and Eskénazi, 2020a; Ghazarian et al., 2020; Gupta et al., 2019; Lowe et al., 2017).

Although the FED-metric relies entirely on the pretrained DialoGPT, which cleverly 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 (Mehri and Eskénazi, 2020a), but consistent with later work (Yeh et al., 2021). 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 (e.g., the width and up-/down-votes of user comments), none of which are common in real dialogue data, DialogRPT only achieves moderate 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 §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 utilises a BERT-MLM score. Training as a classification task loses much fine-grained information such as the subtle differences between RD labels, which restricts their ability for engagingness prediction. Although WeSEE is also based on BERT and shares a similar model architecture as PredEnga, we train WeSEE as a regression model, allowing it to capture subtle differences of RD labels. Our ablation study (§5.3) shows that this regression formulation is more suitable than classification with RD labels.

	FED		DD-H	
	P	S	P	S
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**

Table 4: Model performances when using only a single dialogue turn. P: Pearson; S: Spearman. All correlation results that are not statistically significant (with $p\text{-}value < 0.05$) are *italicised*. Higher is better. Best results in each column are **bold faced**. WeSEE uses DD-H as the validation set.

DynaEval outperforms other baseline models on FED. DynaEval is trained on dialogue datasets (i.e., ED, ConvAI2 (Dinan et al., 2019) 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 containin 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 is that it was not trained on engagingness labels. Acquiring enough high-quality engagingness (class) labels is itself a difficult problem, while WeSEE circumvents this problem with weak supervision.

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 4. 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 hypothesise that this is because DialogRPT uses the transformer

	FED		DD-H	
	P	S	P	S
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

Table 5: Ablation study results. P: Pearson; S: Spearman. Correlation results that are not statistically significant ($p\text{-}value < 0.05$) are *italicised*. Higher is better.

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 predictions deteriorate. WeSEE does not suffer from this problem, as we use mean pooling of all tokens excluding padding tokens as the turn representation.

5.3 RQ3: Ablation study

We ablate the core components of WeSEE to better understand their impact on the overall performance; see Table 5. These components are: (1) training on RD labels; (2) regression formulation instead of classification; (3) history size; and (4) 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).

Table 2 shows that WeSEE-S trained with shuffled RD labels performs poorly. In the -Shuffle row of Table 5, we confirm this using correlation with human annotations. Thus, although RD labels are used as noisy engagingness labels, there is useful information for training a engagingness evaluator. Due to the noisy nature of RD labels, we cannot rely totally on them for training WeSEE. As can be seen from the -ValLoss row, if we select WeSEE's model weights according to the lowest validation MSE loss 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 not statistically significant. This does not 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 ConvAI (Logacheva et al., 2018) dataset (CA); 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 (Ghazarian et al., 2020). During training/fine-tuning WeSEE on the CA dataset, we also used DD-H as the validation set. As shown in Table 5, WeSEE trained on CA with 1 (-CA1) or 3 (-CA3) turns performs worse than their counterparts trained only on RD labels. Thus, weak RD labels are more useful than low-quality human engagingness labels for training WeSEE.

Next, to see the importance of our regression formulation, we modify WeSEE to be a classifier, and map the RD labels to (1) binary labels $\{0, 1\}$ using a threshold 0.5, (2) 5 class labels using thresholds of $\{0.2, 0.4, 0.6, 0.8\}$, and (3) 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 show that, although this classification formulation shows some positive correlation especially with a finer-grained label buckets, the correlation is much weaker than the WeSEE regression model. RD labels are already weak, noisy labels; mapping them to discrete class labels introduces another more noise, limiting the performance of the trained classifiers.

By training and testing WeSEE with more than one historical turn (-H* rows), 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. Since WeSEE does mean pooling for the representation of all participating dialogue turns, it loses the speaker information of each turn. To see how this design influences the prediction, we also consider 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 are shown in the -Flat-H* rows. Using flat history performs consistently worse; the difference between is bigger for using more dialogue turns as can be seen from the FED results on -Flat-H3 and -H3. Thus, speaker information acts as a distracting factor for predicting engagingness, and therefore, we adopt the order-invariant design of dialogue turns in Fig. 2, similar to PredEnga.

The last three rows in Table 5 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 FED except that max pooling shows no noticeable difference.

5.4 RQ4: Result analysis

Appendix C provides more details and examples drawn from case studies we conducted to analyse our results. The main insights gained from these case studies are: (1) WeSEE can distinguish conversation starters and endings by assigning higher scores to the former and lower scores to the latter. 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; the middle of a dialogue is usually the most content-rich and dynamic section. (2) When an utterance contains a question, starts a new topic, or being more detailed, WeSEE usually assigns a higher score, which concurs with the identified factors facilitating engagingness (See et al., 2019; Roller et al., 2021). (3) WeSEE struggles to predict correct labels for short and uninformative responses, and questions that terminate the conversation (e.g., "Anything else I can do?").

6 Conclusion

We studied the problem of predicting turn-level dialogue engagingness and proposed a novel approach that sets the new state-of-the-art results across several dialogue datasets. 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 calculated automatically. We can use any multi-turn dialogue dataset for training our model. When trained on a mixture of four popular dialogue datasets, the proposed *Weakly Supervised Engagingness Evaluator* (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 hypothesise that this is due to DD-H's having only two turns for each data point, which is too short for WeSEE-H3. The WeSEE model developed in this work can be applied to evaluate engagingness of dialogue 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 optimise for open-domain dialogue 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.

7 Ethical Considerations

All the training/validation/test data used in this work is publicly available. As far as we know, the creators of these datasets have taken ethical issues into consideration when creating the datasets. 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

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*.

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.

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*.

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*, pages 386–395. Association for Computational Linguistics.

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*, pages 7789–7796. AAAI Press.

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*, pages 379–391. Association for Computational Linguistics.

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):40–55.

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*, pages 986–995. Asian Federation of Natural Language Processing.

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*.

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*, pages 47–57. Springer.

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*, pages 1116–1126. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskénazi. 2020a. 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*, pages 225–235. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskénazi. 2020b. USR: an unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 681–707. 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, ACL 2020, Online, July 5-10, 2020*, pages 3619–3629. 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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4164–4178. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

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*, pages 300–325. Association for Computational Linguistics.

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)*, pages 1702–1723. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2430–2441. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2021–2030. Association for Computational Linguistics.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv:1801.03625*.

Falcon William and The PyTorch Lightning team. 2019. Pytorch lightning.

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*, pages 38–45. Association for Computational Linguistics.

Omry Yadan. 2019. Hydra - a framework for elegantly configuring complex applications. Github.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706.

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*, pages 65–75. Association for Computational Linguistics.

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*, pages 5676–5689. Association for Computational Linguistics.

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*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

APPENDICES

We provide additional details on our experimental results, both to aid the reproducibility of the results in this paper (Appendix A) and to provide further insights into the results produced by WeSEE (Appendix C).

A Reproducibility

A.1 Link to source code

https://anonymous.4open.science/r/WeSEE. Our implementation is based on Hugging Face Transformers (Wolf et al., 2020), PyTorch Lightning (William and team, 2019), and Hydra (Yadan, 2019). The data downloading and preprocessing are automatically taken care of in our training scripts, parameter settings included. Reproducing the best-performed model requires only one line of code. Please refer to the README in the above link.

A.2 Dataset statistics

Statistics for the datasets we use to train WeSEE are shown in Table 6. In our experiments, we train WeSEE on the mixture of DD, PC, ED and WoW. The reason for this is to add more diversity and generalisability to the trained model. These datasets all have different styles, average dialogue lengths, and together they show more general scenarios of open-domain dialogues. We note that although these datasets are created in a lab environment, there are still noticeable patterns of using engaging/not engaging responses as desired in the dialogue sessions. E.g., dialogue participants tend to speak greetings, starting topics, asking questions in the beginning of a dialogue, and express farewells, use more generic responses in the end of a dialogue. CA dataset is only used for comparison in §5.3 and not in our final model.

A.3 Parameter settings

We chose the BERT base uncased model (Devlin et al., 2018) as implemented in the Transformers library[2] as our turn encoder. The parameters for the linear projection layer of WeSEE are randomly initialised. The WeSEE model contains 109M trainable parameters (weights), in total. We select hyper-parameters using two different criteria, as described in the end of §3. We

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	–	–

Table 6: Statistics for the datasets used to train WeSEE.

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

also evaluated four alternative pooling methods, two activation functions mentioned in §3 and $k \in \{1, 2, 3, 4, 5\}$ for deciding upon the most suitable configuration. In our preliminary experiments, we trained the WeSEE model using an SGD optimiser with a learning rate (LR) chosen from the set $\{5e{-}2, 5e{-}3, 5e{-}4, 5e{-}5, 5e{-}6\}$, and found out that $5e{-}2$ worked 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. 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.

B WeSEE Correlations for F&L k Turns

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 visualises 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 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.

C Results Analysis

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

In Figure 4 are 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, our 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: Our model 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, our WeSEE predicts longer conversations will happen. There may be other 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.

However, there are also some tricky cases that our 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 conver-

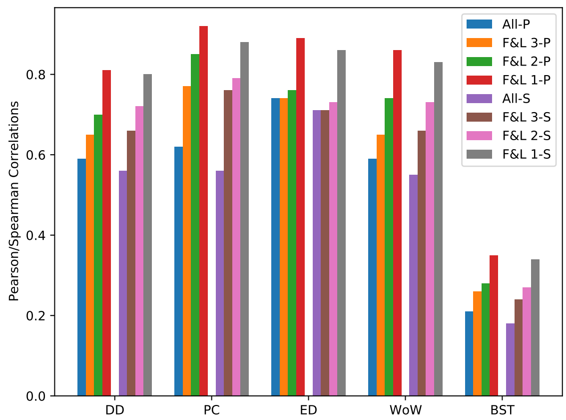

Figure 3: WeSEE correlations with RD for all turns and first & last k (F&L k) turns only. -P: Pearson, -S: Spearman.

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

Figure 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.

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

Figure 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	H1
everything is going extremely well. how are you?	0.90	0.89
what is the meeting about?	0.80	0.76
try me. what is your problem?	1.00	0.61
not that much more, no.	0.40	0.27
i did not want to hear that now	0.80	0.33

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

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

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

sation 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 -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. E.g., a "Thank you." together with a leaving body-language clearly shows that the conversation is ending. In the pure textual setting, this is sometimes impossible to accurately predict. There is another tendency that our 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.

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 (normalised to $[0, 1]$), and there is also some cases that our model 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.

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.