Measuring and Improving Semantic Diversity of Dialogue Generation

Anonymous ACL submission

Abstract

Response diversity has become an important criterion for evaluating the quality of open-domain dialogue generation models. However, current evaluation metrics for response diversity do not capture semantic diversity of generated responses, as they only consider lexical aspects of the responses. In this paper, we introduce a new automatic evaluation metric to measure the semantic diversity of generated responses. Through human evaluation, we demonstrate that our proposed metric highly correlates to human judgments on response diversity than existing lexical-level diversity metrics. Furthermore, motivated by the analysis of an existing dialogue dataset, we propose a simple yet effective learning method that improves the semantic diversity of generated responses through response re-weighting based on the semantic distribution of the training dataset. Through automatic and human evaluation, we show that our proposed learning method better improves both response diversity and coherency compared to other baseline methods.

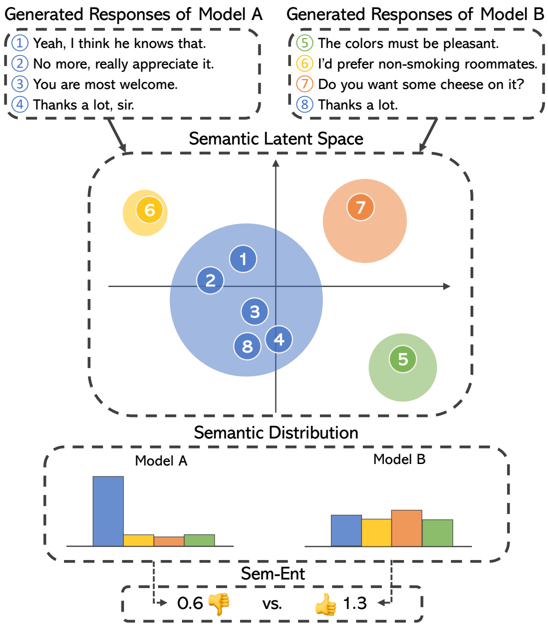

Figure 1: An illustration of measuring semantic diversity of generated responses. Although both Model A and Model B generate lexically diverse responses, we argue that the responses of Model B seem more varied in human perception because they are semantically diverse. Our proposed Sem-Ent measures semantic diversity based on the semantic distribution of generated responses.

1 Introduction

Open-domain dialogue generation (Sordoni et al., 2015; Bordes et al., 2017) has greatly progressed with the development of large-scale pretrained language models (Radford et al.; Roller et al., 2021) in the last decade. However, although dialogue generation models can produce fluent responses for a given context, they are also known for frequently generating dull and uninformative generic responses (e.g., "*I don't know*"), degrading the interestingness of responses (Serban et al., 2016; Li et al., 2016a). To alleviate this problem, many studies (Zhao et al., 2017; Li et al., 2017a; Zhang et al., 2018) have been conducted to enhance the *diversity* of generated responses, and response diversity has become an important criterion for evaluating the quality of generated responses.[1]

The current evaluation protocol employs lexical-level evaluation metrics such as *distinct-n* (Dist-*n*) (Li et al., 2016a) and *entropy-n* (Ent-*n*) (Serban et al., 2017; Zhang et al., 2018) to measure the diversity of generated responses. However, it is unclear whether lexical-level evaluation metrics can successfully capture the human judgment on response diversity. For instance, in Figure 1, responses generated by model A and model B both show high lexical diversity, but humans intuitively recognize that the responses of model B

[1] According to recent survey papers (Ni et al., 2021; Liang and Li, 2021), more than thirty studies within five years have assessed dialogue generation models from the diversity perspective.

are more diverse. We argue that considering a *semantic diversity* of the generated responses is more important for capturing human judgment on response diversity. However, the lexical-level metrics cannot directly capture the semantic diversity since responses including similar words can have very different semantics, and responses with different words can have similar semantics (Yarats and Lewis, 2018). Nevertheless, most studies have conducted an evaluation with only the lexical-level evaluation metrics to measure the diversity of generated responses because there is no alternative metric to measure the semantic diversity.

To this end, we propose ***Sem-Ent*** (**Sem**antic-**Ent**ropy), which is a new automatic evaluation metric for measuring the semantic diversity of generated responses. Sem-Ent first maps generated responses into a semantic latent space using a pre-trained language model (e.g., DialoGPT (Zhang et al., 2020) and BERT (Devlin et al., 2019)). Then, the metric measures the semantic diversity of generated responses by measuring how the responses are evenly distributed in the semantic latent space based on semantic clusters, as shown in Figure 1. Through human evaluation, we demonstrate that Sem-Ent is more highly correlated with human judgments on response diversity than existing lexical-level evaluation metrics. The human evaluation further shows that Sem-Ent highly correlates with human judgments about how they feel generated responses are interesting.

Furthermore, we observe that the semantic distribution of responses in the dialogue dataset is highly imbalanced. This imbalance leads the model to produce semantically less diverse responses. To address this problem, we propose a simple yet effective learning method of dialogue generation models. Our proposed method, ***DRESS*** (Diversifying RESponses Semantically), induces dialogue generation models to learn more about responses with rare semantics and learn less about responses with frequent semantics. From this, dialogue generation models could produce more semantically diverse responses. Experiments on two benchmark datasets demonstrate that DRESS shows substantially better semantic diversity compared to state-of-the-art baseline methods, along with the gain in response coherency. Interestingly, DRESS achieves better performance in evaluation metrics for lexical-level diversity than baselines even though it focuses on improving the semantic diversity of generated responses. Moreover, human evaluation results also affirm the effectiveness of DRESS, where DRESS outperforms all baseline methods in terms of appropriateness and informativeness of generated responses.

Our Contributions: (1) A new automatic evaluation metric for measuring semantic diversity (Sem-Ent), which is highly correlated with human judgment on response diversity. (2) A simple yet effective learning method of dialogue generation models (DRESS) for improving the semantic diversity of generated responses. (3) Experiments on two benchmark datasets, showing that DRESS outperforms the baseline methods in both semantic diversity and lexical-level diversity. (4) A Python library[2] of Sem-Ent, contributing to the community of open-domain dialogue generation.

2 Related Work

2.1 Open-domain Dialogue Models for Enhancing Response Diversity

Since generating dull and uninformative responses is a well-known and essential problem in open-domain dialogue (Vinyals and Le, 2015; Li et al., 2016a), numerous lines of works have been proposed to address this issue. Li et al. (2016a) replace the standard maximum likelihood objective into maximum mutual information objective to penalize generic responses. This new objective function has been continuously adopted in subsequence works to increase the specificity and diversity of generated responses (Li et al., 2016c; Zhang et al., 2018, 2020). Another line of work improves diversity by modeling the one-to-many relationship of open-domain dialogue using latent variables to generate multiple and diverse responses (Serban et al., 2017; Zhao et al., 2017; Bao et al., 2020a,b; Chen et al., 2019; Zhang et al., 2019; Gao et al., 2019). Some methods selectively penalize frequent responses by removing them from the training dataset (Csáky et al., 2019) or applying negative training to frequent responses (He and Glass, 2020). Using different decoding algorithms can improve the response diversity; Li et al. (2016b) and Vijayakumar et al. (2018) directly modify the beam search algorithm to promote the response diversity. Sampling-based decoding algorithms such as top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2019) are known to improve the diversity of generated responses. Wang et al. (2021) diversify re-

[2]Link will be released after publication.

sponses by adaptively modifying the target token distribution with a lightweight decoder to prevent the model from being over-confident.

2.2 Metrics for Capturing Response Diversity

Response diversity metrics for open-domain dialogue generation models can mainly be categorized into two groups. Referenced metrics (Zhao et al., 2017; Gao et al., 2019) use the reference responses provided by human annotators to capture the response diversity by computing a recall value based on various similarity metrics such as BLEU and embedding similarity. On the other hand, unreferenced metrics measure the response diversity without the use of reference responses generated by human annotators. Therefore, unreferenced metrics are more widely adopted than referenced metrics because they can measure response diversity even in the absence of reference responses. Dist-n (Li et al., 2016a) measures the response diversity with the fraction of distinct n-grams over possible n-grams in all generated responses. Ent-n metric (Serban et al., 2017; Zhang et al., 2018) is suggested to improve the Dist-n metric by taking the frequency difference of n-grams into account. LF (Li et al., 2019) calculates the frequency of low-frequency words in generated responses as the response diversity. Our work focuses on introducing a semantic diversity metric that alleviates the limitation of the aforementioned unreferenced diversity metrics of considering only lexical aspect of generated responses.

3 Measuring Semantic Diversity

3.1 Sem-Ent

Let $\mathcal{D} = \{(c_1, r_1), (c_2, r_2), \cdots (c_m, r_m)\}$ denote a training dataset consisting of m dialogues where c_i and r_i denote the context and its response of the i-th dialogue, respectively. Dialogue generation is to generate a response r for a given context c.

We are motivated by recent empirical observations that responses can be clustered by the semantic similarity between the responses (Ko et al., 2020; Gao et al., 2020). By following Csáky et al. (2019); Pillutla et al. (2021), we cluster responses in $\mathcal{D}$ by utilizing a pretrained language model. Here, we select DialoGPT (Zhang et al., 2020) as the language model. Each response r_i in $\mathcal{D}$ is turned into a semantic representation $e(r_i)$ by the language model, and then k semantic clusters are formed from the semantic representations by the k-means algorithm (Lloyd, 1982). Let $\mathcal{C}$ denote a set of the obtained k semantic clusters.

Consider a test dataset $\tilde{\mathcal{D}} = \{(\tilde{c}_1, \tilde{r}_1), \cdots, (\tilde{c}_n, \tilde{r}_n)\}$ consisting of n dialogues. During evaluation, a dialogue generation model M generates responses $\mathcal{R}^M = \{r_1^M, \cdots, r_n^M\}$ for the contexts $\{\tilde{c}_1, \cdots, \tilde{c}_n\}$ in $\tilde{\mathcal{D}}$, respectively. Sem-Ent measures the semantic diversity of $\mathcal{R}^M$ generated by the model M. To compute Sem-Ent, we require a semantic distribution $P(\mathcal{R}^M)$, but there is no direct way to obtain the exact distribution. Thus, we approximate the semantic distribution $P(\mathcal{R}^M)$ using a distribution $\tilde{P}(\mathcal{C}) = \left[\tilde{p}(1); \cdots; \tilde{p}(k)\right]$ of the semantic clusters $\mathcal{C}$ as follows:

$$\tilde{p}(j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big(\phi_{\mathcal{C}}(e(r_i^M)) = j\big), \quad (1)$$

where $\phi_{\mathcal{C}}(x) \in \{1, \cdots, k\}$ is a cluster mapping function that returns the cluster id of x from $\mathcal{C}$. $\tilde{p}(j)$ is the probability of the j-th cluster, indicating how many generated responses are assigned to the j-th semantic cluster.

Sem-Ent is an entropy of $\tilde{P}(\mathcal{C})$, which is calculated with $\tilde{P}(\mathcal{C})$ approximating the semantic distribution of $\mathcal{R}^M$ as follows:

$$\text{Sem-Ent}(\mathcal{R}^M) = -\sum_{j=1}^{k} \tilde{p}(j) \cdot \log \tilde{p}(j). \quad (2)$$

Interpretation of Sem-Ent is quite straightforward: Sem-Ent gets lower when the semantic distribution gets more imbalanced, i.e., when models generate responses belonging to only several specific semantic clusters. Conversely, Sem-Ent gets the highest value of $\log k$ when generated responses are uniformly distributed to each semantic cluster.

3.2 Correlation with Human judgment

We conduct a human evaluation to demonstrate that Sem-Ent successfully captures human judgments on response diversity.

Experimental Setup. We use a similar experimental setup to that of Pillutla et al. (2021) for analyzing the correlation between response diversity metrics and human judgment. We prepare eight inference settings from two generation models (Blender-90M (Roller et al., 2021) and BART-large (Lewis et al., 2020)) and four decoding algorithms (greedy, beam, top-k sampling, and nucleus sampling). The generation models are fine-tuned on DailyDialog (Li et al., 2017b) dataset that consists of daily conversations about various topics.

Metric	Correlation	Dist-3	Ent-3	LF	MAUVE	Sem-Ent
Diversity/BT	Pearson	0.348 (0.399)	0.702 (0.052)	-0.232 (0.580)	0.134 (0.750)	**0.810 (0.015)**
	Spearman	0.381 (0.352)	0.667 (0.071)	0.000 (1.000)	0.547 (0.160)	**0.762 (0.028)**
Interesting/BT	Pearson	0.261 (0.533)	0.671 (0.068)	-0.260 (0.533)	0.098 (0.817)	**0.789 (0.020)**
	Spearman	0.381 (0.352)	**0.714 (0.047)**	0.048 (0.911)	0.523 (0.182)	0.667 (0.020)

Table 1: Correlation of various diversity measures with human judgments. "BT" denotes the Bradley-Terry score for a pair-wise human evaluation and the value inside the parenthesis indicates p-value.

Then, each inference setting is paired with other settings, which gives a total of 28 (${}_8\mathrm{C}_2$) pairs of settings.

For every round, a human annotator is assigned to a set that contains ten independent contexts $c_1, \cdot, c_{10}$ and two sides of responses $r_{1a}, \cdot, r_{10a}$ and $r_{1b}, \cdot, r_{10b}$ generated with two different settings. The annotator is asked to select which side of responses is better in two criteria; whether (1) shows more diversity and (2) shows more interesting and creative responses, using a 5-point Likert scale. We obtain 25 preference ratings for each pair of inference settings. These annotation results are converted into each setting's score by using the Bradley-Terry model (Marden, 1996) fitted by pairwise annotations. We measure the correlation between the Bradley-Terry score and diversity metrics to check how each metric correlates with the human judgment on each criterion. More details about human evaluation are included in Appendix.

Baseline Metrics. We compare Sem-Ent with existing lexical-level response diversity metrics: Dist-n (Li et al., 2016a), Ent-n (Serban et al., 2017; Zhang et al., 2018) and LF (Li et al., 2019). We also include recently proposed MAUVE (Pillutla et al., 2021) as a baseline metric. MAUVE shares some properties with Sem-Ent such that it evaluates the distributional property of generated responses with semantic latent representations. However, it is designed to measure the divergence of generated responses from human responses, not for directly measuring response diversity. We compare Sem-Ent to MAUVE to verify that our Sem-Ent is more suitable for measuring the response diversity in open-domain dialogue generation.

Results. Table 1 shows the correlation between the human judgments and the different diversity metrics in terms of Pearson and Spearman rank correlation. Our Sem-Ent shows the highest correlation (on both Pearson correlation and Spearman correlation) with human judgment on response diversity compared to other evaluation metrics with a significant margin. Especially, Dist-n, the most commonly used metric for response diversity, shows a much lower correlation (0.348) compared to Sem-Ent (0.810). These results support that Sem-Ent is a good surrogate for measuring human judgment on response diversity and strongly suggest that analyzing the semantic diversity of generated responses is crucial for capturing human perception of response diversity. Moreover, MAUVE shows a lower correlation with human judgment on response diversity. This result implies that a closer gap between human responses and generated responses does not always indicate that generated responses are diverse since human responses contain many dull responses frequently (also studied in Section 4.1 and by Csáky et al. (2019)).

We also observe that Sem-Ent shows a high correlation with human judgment on interestingness; Sem-Ent has a similar correlation to Ent-n and shows a substantially higher correlation than Dist-n, LF, and MAUVE. We believe that the strong correlation of Sem-Ent with human judgment on response diversity leads to a high correlation with a closely related model property, interestingness.

In Section 6, we further justify that Sem-Ent is robust to a choice of configurations used for the metric such as a choice of the language model for extracting semantic representations of responses and a number of clusters k.

4 DRESS: Diversifying RESponses Semantically

4.1 Diagnosing the Semantic Distribution of Dialogue Dataset

As shown in Section 3.2, semantic distribution of responses provides a crucial clue for understanding the diversity of the responses. Therefore, we analyze the semantic distribution of the responses in the training dataset. Figure 2 depicts the semantic distribution $\tilde{P}(\mathcal{R})$ of the responses $\mathcal{R} = \{r_1, r_2, \cdots, r_m\}$ in the training data of DailyDialog dataset. As shown in the figure, the semantic distribution of the training dataset $\mathcal{D}$ is highly skewed – almost half of the responses fall into the

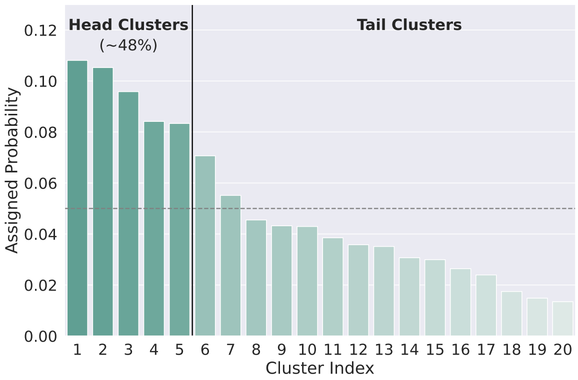

Figure 2: Semantic distribution of the responses in the train split of DailyDialog. Clusters are sorted in the descending order of the assigned probabilities. The dashed line indicates the uniformly distributed probability, 0.05.

Index	Responses
2	· Yeah . I know . · Thank you . · You are most welcome . · No more , thank you very much . · Not yet .
13	· that sounds great . Do you know if there are any vegetable dishes that are spicy ? · Do you want cheese on it ? · I agree . The colors must be soft and pleasant . You should feel comfortable when you cook our dinners.
18	· I bought a new mattress and some fresh bed-clothes . I also bought a new dressing table and a new bedside table . · I ' d prefer non-smoking roommates , but I guess I ' ll have to take what I can get ! · A single room with a front view is 100 dollars per night , one with a rear is 80 dollars .

Table 2: Response examples of the semantic clusters. *Index* column indicates the Cluster Index in Figure 2.

top five frequent clusters (head clusters). Moreover, the frequent clusters tend to contain more generic and dull responses compared to infrequent clusters (tail clusters), as illustrated in Table 2. Contrarily, responses in the infrequent clusters have a wider variety of topics, intents, and diverse vocabularies. Since the training data is skewed towards semantically generic and dull responses, naively training with this data will lead to a low semantic diversity of generated responses.

4.2 Improving Semantic Diversity with DRESS

We introduce a simple yet effective learning method of generation models for improving semantic diversity, DRESS, which addresses the problem of the imbalanced semantic distribution by re-weighting the instances in the training dataset. The purpose of DRESS is simple: inducing generation models to learn more about responses in the infrequent semantic clusters and contrarily learn less about responses in the frequent semantic clusters. To this end, DRESS modifies the learning objective into the weighted loss function and applies Negative Training (He and Glass, 2020; Li et al., 2020) to the modified objective.

A conventional dialogue generation model is trained by optimizing an NLL (negative log-likelihood) objective as follows:

$$L_{NLL}(D) = -\sum_{i=1}^{m} \log p_\theta(r_i|c_i), \quad (3)$$

where θ indicates parameters of dialogue generation models. Instead of using vanilla NLL objective, we propose to utilize weighted NLL objective in DRESS using weight of responses $w(r_i)$:

$$L_{DRESS}(D) = -\sum_{i=1}^{m} w(r_i) \cdot \log p_\theta(r_i|c_i). \quad (4)$$

The goal of weighted NLL objective is to assign smaller weights to the responses in frequent semantic clusters and assign bigger weights to responses in infrequent semantic clusters to balance the semantic distribution. To meet this condition, the weighting function $w(r)$ should satisfy the constraint: if $\tilde{p}(\phi_c(e(r_i))) \leq \tilde{p}(\phi_c(e(r_j)))$, then $w(r_i) \geq w(r_j)$. Inspired by focal loss (Lin et al., 2017) which is used in the long-tail classification problem (Liu et al., 2019b; Hong et al., 2021), we calculate w as follows:

$$w(r) = \left(1 - \tilde{p}(\phi_c(e(r)))\right)^\gamma, \quad (5)$$

where γ is a hyperparameter for controlling a degree of re-weighting (higher γ means more intense re-weighting).

Moreover, to penalize responses in frequent semantic clusters intensively, we utilize Negative Training (He and Glass, 2020; Li et al., 2020) jointly with the weighted objective function. For every epoch, the model generates responses to each given context. If generated responses are included in head clusters (here, the assigned probability of clusters is bigger than 0.1), then those generated responses are assumed as negative examples, i.e., assigning $w(r) = -1$.

5 Experiments

5.1 Experimental Setup

We conduct experiments to demonstrate that the proposed learning method successfully improves response diversity.

Dataset. We conduct experiments on two English open-domain dialogue datasets: DailyDialog and OpenSubtitles (Lison and Tiedemann, 2016). DailyDialog consists of 13K dialogues which includes 87K context-response pairs, and we split the dialogues into train/valid/test sets in 8:1:1. The test set of DailyDialog contains 6.7K context-response pairs. OpenSubtitles is a large corpus containing movie scripts, and we use the version released in 2018 with 100K context-response pairs for the training and validation set each. We get rid of context-response pairs whose response is shorter than five words from the original test set and randomly sample 10K pairs as test data.

Automated Metrics. As the goal of diversity-promoting dialogue generation models is to generate diverse responses without hurting the coherency of responses, we focus on two criteria: response diversity and coherency. For measuring response diversity, we use both lexical-level diversity metrics (Dist-n, Ent-n, and LF) and a semantic diversity metric (Sem-Ent, $k = 20$). For measuring response coherency, we employ MaUdE (Sinha et al., 2020), an unreferenced dialogue response evaluation metric that shows a high correlation with human judgments on the fluency of responses.

Human Evaluation. We further conduct a pairwise comparison through the human evaluation for evaluating generated responses since automatic evaluations are sometimes not trustworthy. We use Amazon Mechanical Turk to collect the annotations. Each annotator evaluates which model is better in terms of *Appropriateness* for measuring response coherency and *Informativeness* for evaluating whether the given response has meaningful information relevant to its given context. We collect annotations for 50 test cases per each model pair, and three annotators rate each test case to improve the robustness of the evaluation result. More details about evaluation protocol (e.g., interface for collecting annotation) are shown in Appendix.

5.2 Baseline Methods

MMI (Li et al., 2016a) increases response diversity by maximizing the mutual information between context and response rather than maximizing the likelihood as in conventional dialogue models. We utilize the MMI-antiLM as our MMI baseline.

CVAE (Zhao et al., 2017) is a representative model among dialogue generation models that utilize latent variables to increase response diversity. CVAE builds the response generation process as a conditional variational auto-encoder of a response with dialogue context as a condition.

EDF (Entropy-based Data Filtering) (Csáky et al., 2019) enhances response diversity by filtering out context-response pairs that increase one-to-many or many-to-one problems in the training dataset. We use target side entropy to filter the pairs.

NT (Negative Training) (He and Glass, 2020) directly penalizes the generation of generic responses by applying reverse direction gradient for the losses of the generic responses, leading to maximizing the loss rather than minimizing it.

AdaLabel (Wang et al., 2021) alleviates the overconfidence problem of generation models to improve response diversity by dynamically smoothing the target token distribution with an auxiliary lightweight decoder.

5.3 Implementation Details

We take two Transformer-based sequence-to-sequence models: Blender-90M (Roller et al., 2021) and BART-large (Lewis et al., 2020) as the underlying generation models to demonstrate that our method widely works well on different architectures. For DRESS, we set $\gamma = 30$ and the number of clusters $k = 20$ in our whole experiments unless otherwise specified. All models use greedy decoding strategy, and we utilize both blocking repeated n-grams (Paulus et al., 2017) ($n = 3$) within the generated response and the input sequence to prevent models from repeating subsequences. Moreover, we release our implementation code[3] publicly to help researchers reproduce the result.

6 Results and Analysis

6.1 Evaluation Results

Table 3 shows the automatic evaluation results. Overall, DRESS achieves the best performance in both semantic and lexical-level response diversity while showing high response fluency for most of the experimental setups. To be more specific, as shown in the table, DRESS shows a substantially higher semantic diversity (Sem-Ent) than all other baseline models in every experimental setup.

[3]Link will be released after publication.

Backbone	Method	Dist-1	Dist-2	Dist-3	Ent-1	Ent-2	Ent-3	LF	MaUdE	**Sem-Ent**
Blender-90M (DailyDialog)	Vanilla	0.0453	0.2103	0.3881	7.1322	10.7502	12.3950	0.2234	0.8489	2.5486
	MMI	0.0349	0.1677	0.3069	7.0730	10.3806	11.9808	0.2155	0.8208	2.5784
	CVAE	0.0471	0.2389	0.4459	7.4074	11.2797	12.9969	0.2449	0.8552	2.6261
	EDF	0.0473	0.2271	0.4226	7.2888	11.0283	12.7132	0.2402	0.8593	2.5872
	NT	**0.0475**	0.2351	0.4422	7.3994	11.2561	13.0111	0.2467	**0.8597**	2.6434
	AdaLabel	0.0377	0.1982	0.3915	7.1546	10.8772	12.6829	0.2158	0.8443	2.6038
	DRESS(-NT)	0.0445	0.2295	0.4360	7.4560	11.3273	13.1028	0.2474	0.8460	2.7576
	DRESS	0.0460	**0.2404**	**0.4571**	**7.5468**	**11.5094**	**13.3060**	**0.2576**	0.8575	**2.7819**
BART-large (DailyDialog)	Vanilla	0.0462	0.2168	0.4056	7.3913	11.2075	12.8648	0.2593	0.8854	2.4251
	MMI	0.0497	0.2329	0.4355	7.4748	11.4060	13.0898	0.2623	0.8787	2.4646
	CVAE	0.0429	0.2416	0.5117	7.2728	11.2968	13.1643	0.2558	0.8744	2.4215
	EDF	**0.0597**	**0.2926**	0.5355	7.9606	12.1776	13.8786	0.3036	0.8918	2.5842
	NT	0.0571	0.2919	0.5424	8.0267	12.3098	14.0577	0.3070	0.9024	2.6690
	AdaLabel	0.0482	0.2573	0.5136	7.9152	12.0968	13.9496	0.2936	0.8947	2.6336
	DRESS(-NT)	0.0554	0.2909	0.5448	8.1722	12.5195	14.3244	**0.3079**	**0.9192**	2.8444
	DRESS	0.0547	0.2906	**0.5504**	**8.1821**	**12.5533**	**14.3890**	0.3052	0.9153	**2.8548**
Blender-90M (OpenSubtitles)	Vanilla	0.0373	0.1550	0.2698	6.5882	9.5097	10.7983	0.1758	0.8459	2.4702
	MMI	0.0426	0.1660	0.2755	6.4854	9.2276	10.3364	0.2005	0.8721	2.4469
	CVAE	0.0393	0.1804	0.3398	7.0092	10.5135	11.8959	0.2073	**0.9214**	2.5726
	EDF	0.0476	0.2019	0.3536	7.0189	10.3899	11.8036	0.2161	0.8777	2.5738
	NT	0.0504	0.2216	0.3969	7.3734	11.0928	12.6594	0.2480	0.8944	2.7049
	AdaLabel	0.0431	0.1913	0.3573	7.0306	10.5280	12.0680	0.2063	0.8708	2.6407
	DRESS(-NT)	0.0499	0.2178	0.3817	7.3316	10.8422	12.2530	0.2308	0.8927	2.7114
	DRESS	**0.0524**	**0.2351**	**0.4180**	**7.5113**	**11.2355**	**12.7612**	**0.2612**	0.9041	**2.7654**
BART-large (OpenSubtitles)	Vanilla	0.0262	0.1028	0.1806	5.8507	8.2064	9.2760	0.1532	0.7803	2.2043
	MMI	0.0275	0.1094	0.1923	6.0557	8.5303	9.6961	0.1595	0.8067	2.1626
	CVAE	0.0226	0.1460	0.3495	6.2232	9.7304	11.4593	0.1507	0.8600	2.3005
	EDF	**0.0474**	0.2056	0.3572	7.0338	10.5464	11.9977	0.2209	0.8558	2.5346
	NT	0.0228	0.0948	0.1594	5.5542	8.2025	9.6915	0.1165	0.8298	2.6368
	AdaLabel	0.0381	0.1772	0.3316	7.0306	10.5667	12.0747	0.2030	0.8647	2.5652
	DRESS(-NT)	0.0456	0.2006	0.3509	7.1669	10.6915	12.1509	0.2220	0.8618	2.6620
	DRESS	0.0472	**0.2178**	**0.3890**	**7.4656**	**11.2761**	**12.8601**	**0.2322**	**0.8873**	**2.7406**

Table 3: Automatic evaluation results in terms of various diversity metrics (Dist-n, Ent-n, LF, and Sem-Ent) and coherency metric (an average MaUdE of generated responses). **Bolded** value indicates the best result and underlined value indicates the runner-up among the results. DRESS(-NT) indicates the variant version of DRESS that only utilizes the weighted NLL without Negative Training.

Figure 3 illustrates the detailed semantic distribution of the generated responses. While the Vanilla model shows a high probability on the head semantic clusters (e.g., Cluster 1, 2, 4) and low probability on the tail semantic clusters (e.g., Cluster 13∼20), DRESS effectively reduces the probabilities of the head semantic cluster and boosts probabilities of the tail clusters. It is quite intriguing that DRESS also achieves better performance in lexical-level response diversity (Dist-n, Ent-n, and LF). Furthermore, MaUdE results indicate that DRESS preserves better response coherency compared to other baseline methods.

Apart from automatic evaluation, we further compare DRESS with baseline methods in pairwise human evaluation to verify the effectiveness of DRESS. Table 4 shows the evaluation results, showing clear improvements in terms of appropriateness and informativeness from using DRESS.

6.2 Analysing DRESS

Changing Hyperparameters of DRESS. We examine how the automatic results change when we vary the hyperparameters of DRESS: γ in Equation 4.2 and the number of clusters k. Table 5 shows the results about the effect of the hyperparameters. We find that increasing γ induces models to produce more diverse responses, which can be shown by improvement in Dist-3, Ent-3, and Sem-Ent. We also observe that decreasing k induces the models to generate more diverse responses. However, MaUdE gets degraded while response diversity improves, which implies a trade-off between response diversity and coherence.

Ablation Study. To verify the effect of our weighted NLL, we conduct an ablation study. In

Comparison (A vs. B)	Appropriateness			Informativeness		
	A wins (%)	B wins (%)	Tie (%)	A wins (%)	B wins (%)	Tie (%)
Ours vs Vanilla	**35.3**	24.7	40.0	**36.0**	28.0	36.0
Ours vs MMI	**40.0**	34.7	25.3	**40.7**	36.0	23.3
Ours vs CVAE	**44.7**	30.0	25.3	**36.7**	36.0	27.3
Ours vs EDF	**35.3**	24.7	40.0	**32.7**	23.3	44.0
Ours vs NT	**28.0**	25.3	46.7	**37.3**	26.0	36.7
Ours vs AdaLabel	**28.7**	24.0	47.3	**32.7**	31.3	36.0

Table 4: Human pair-wise comparison results in terms of appropriateness and informativeness of generated responses. The evaluation is conducted on the test set of DailyDialog with Blender-90M using greedy decoding.

Config	Dist-3	Ent-3	MaUdE	Sem-Ent
$\gamma = 1.0$	0.4333	12.8968	0.8570	2.6233
$\gamma = 5.0$	0.4400	12.9989	0.8593	2.6551
$\gamma = 10.0$	0.4410	13.0670	0.8583	2.6959
$\gamma = 30.0$	0.4571	13.3060	0.8575	2.7819
$\gamma = 100.0$	0.4625	13.5839	0.8436	2.8444
$k = 10$	0.4748	13.7596	0.8390	2.8451
$k = 20$	0.4571	13.3060	0.8575	2.7819
$k = 50$	0.4318	13.0001	0.8513	2.7009
$k = 100$	0.4311	12.8857	0.8637	2.6258

Table 5: Analysing the effect of hyperparameters, γ and k. When changing γ, we fix k to 20. When changing k, we fix γ to 30.0.

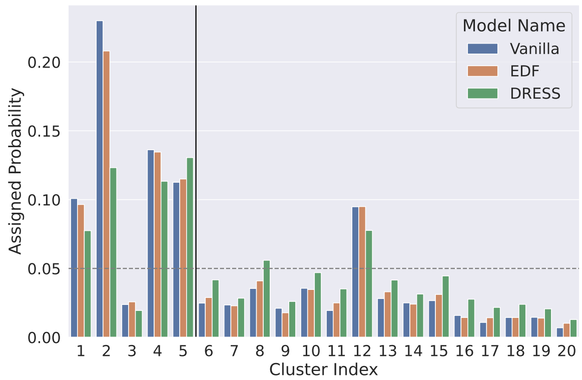

Figure 3: Probability distribution of the responses generated by Vanilla, EDF and DRESS. The dashed line indicates the uniformly distributed probability, 0.05.

Table 3, DRESS(-NT) indicates the variant of DRESS without Negative Training and only utilizes weighted NLL. DRESS(-NT) shows a slight degradation in Sem-Ent compared to DRESS. Nonetheless, DRESS(-NT) achieves better performance in Sem-Ent than other baseline methods excluding DRESS. Moreover, DRESS(-NT) also shows a higher lexical-level diversity than other baseline methods, along with high MaUdE scores.

6.3 Robustness of Sem-Ent on the Choice of Configurations

In this section, we examine the robustness of Sem-Ent changing the configurations used for calculating the metric. Several configurations can be changed in Sem-Ent, including the choice of language models for mapping responses r into a semantic representation $e(r)$ and the number of clusters k for the k-means algorithm. Varying the configurations, we compute Sem-Ent on responses generated by Blender-90M for the test set of DailyDialog with all methods (in Table 3). We then measure the Spearman correlation between the computed Sem-Ent of different configurations.

For the choice of language models, we compare three variants: DialoGPT, RoBERTa (Liu et al., 2019a), and GPT2-large (Radford et al.). The average Spearman correlation between the pairs of these three variants (3 pairs) is 0.8809. For the number of clusters, we vary the number k with values in $\{10, 20, 50, 100\}$ and compare the Sem-Ent rankings. The average Spearman correlation between these configurations (6 pairs) is 0.9821. High correlations show that Sem-Ent produces similar rankings of different models regardless of different configurations, indicating that Sem-Ent is a robust metric for calculating response diversity.

7 Conclusion

In this work, we argue that semantic diversity is overlooked while measuring response diversity of dialogue generation; thus, we present a new automatic evaluation metric, Sem-Ent, which can measure the semantic diversity of generated responses. Sem-Ent correlates with human judgments on response diversity more than other automatic diversity metrics and also shows a high correlation with human judgments in interestingness. Moreover, we introduce a new learning method, DRESS, to improve the semantic diversity of dialogue generation. Evaluation results show that DRESS improves both the semantic diversity and lexical-level diversity of dialogue generation, along with the gain in response coherency.

Ethical Considerations

Dialogue generation models can reveal some biases and toxicities from their responses since these models leverage large-scale web-crawled data for pretraining. This is a common consideration for works related to dialogue generation. Moreover, while our paper focuses on diversifying responses in semantic viewpoint, the model may unintentionally learn about offensive words while diversifying responses. We believe it will be meaningful to reduce potential harmful responses considering semantics in future work.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020a. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020b. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. *arXiv preprint arXiv:1906.01781*.

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of NAACL-HLT*, pages 1229–1238.

Yifan Gao, Piji Li, Wei Bi, Xiaojiang Liu, Michael Lyu, and Irwin King. 2020. Dialogue generation on infrequent sentence functions via structured meta-learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 431–440.

Seungju Han, Beomsu Kim, Seokjun Seo, Enkhbayar Erdenee, and Buru Chang. 2021. Understanding and improving the exemplar-based generation for open-domain conversation. *arXiv preprint arXiv:2112.06723*.

Tianxing He and James Glass. 2020. Negative training for neural dialogue response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2044–2058.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. 2021. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636.

Beomsu Kim, Seokjun Seo, Seungju Han, Enkhbayar Erdenee, and Buru Chang. 2021. Distilling the knowledge of large-scale generative models into retrieval models for efficient open-domain conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3357–3373.

Wei-Jen Ko, Avik Ray, Yilin Shen, and Hongxia Jin. 2020. Generating dialogue responses from a semantic latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4339–4349.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2019. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Hongru Liang and Huaqing Li. 2021. Towards standard criteria for human evaluation of chatbots: A survey. *arXiv preprint arXiv:2105.11197*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019b. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

John I Marden. 1996. *Analyzing and Modeling Rank Data*. CRC Press.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 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*, pages 300–325.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and William B Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Oriol Vinyals and Quoc V Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying dialog generation via adaptive label smoothing. *arXiv e-prints*, pages arXiv–2105.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability–an empirical approach to interpreting inter-rater reliability. *arXiv preprint arXiv:2106.07393*.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pages 5591–5599. PMLR.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31:1810–1820.

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

Yuchi Zhang, Yongliang Wang, Liping Zhang, Zhiqiang Zhang, and Kun Gai. 2019. Improve diverse text generation by self labeling conditional variational auto encoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2767–2771. IEEE.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

A Appendix

A.1 Descriptive Statistics about Results

Confidence Interval of MaUdE In Table 3 and Table 5 from the main paper, we report the average MaUdE score of responses generated by each method. To provide descriptive statistics of evaluation, here we provide a 95% confidence interval of MaUdE in Table 6 and Table 7. Note that we only report confidence intervals of MaUdE since other diversity metrics (Dist-n, Ent-n, LF, Sem-Ent) return a single value from a set of responses, thus can not calculate the confidence interval.

Inter-Rater Reliability of Pairwise Human Evaluation We calculate a Fleiss' Kappa for pairwise human evaluation results to measure the annotation variance. We find that Fleiss' Kappas are 0.09 and 0.04 for appropriateness and informativeness, respectively. Although these values are not high, as Kulikov et al. (2019) and Wong et al. (2021) show on their paper, inter-rater reliability of annotation results using crowd-sourced annotators (such as our case, using Amazon Mechanical Turk) can be low since annotators show high cultural and training variances, especially when the task is subjective as our case. Note that 64 annotators participated in our human evaluation. Also, we limited the number of maximum annotations that a single annotator can be assigned to reduce the bias, which might have increased inter-rater diversity.

A.2 Human Evaluation Protocol

Evaluation for Comparing Metrics We use Amazon Mechanical Turk for collecting assessments, and Figure 4 shows the instructions and the interface for the human evaluation. We mitigate the bias from the annotator by setting a maximum number of annotations per worker as 20 and randomly shuffling the order of the model and the corresponding response. Since our task does not require particular expertise in linguistics, we open the evaluations to non-experts. Nonetheless, to control the annotation quality, we only allow the annotators who satisfy the following requirements: (1) HITs approval rate greater than 95%, (2) Location is one of Australia, Canada, New Zealand, United Kingdom, and the United States, (3) Lifetime number of HITs approved greater than 1000, following Kim et al. (2021); Han et al. (2021). We estimated that each HITs takes around 1.5 minutes on average (87 seconds per each HIT estimated by the 85th percentile of response times) and set the payment to USD 16 per hour. Therefore, annotators are paid USD 0.40 per HITs.

Evaluation for Comparing Methods As we described above, we also use Amazon Mechanical Turk, and we use the same setting to mitigate the bias and control the annotation quality. Figure 5 shows the instructions and the interface for the human evaluation. Here, annotators are paid USD 0.25 per HITs as we estimated that each HITs takes around 1.4 minutes on average (84 seconds per HITs estimated by the 85th percentile of response times) and set the payment to USD 10.7 per hour since the difficulty of the task is easier than above.

A.3 Evaluation Details

Bradley-Terry Model We use the Bradley-Terry model from pairwise human evaluation results to obtain the ranking of the models. Given parameters $\theta_1, \cdot, \theta_n$, for two items i and j, the probability of the outcome $i \succ j$ is $p(i \succ j) = e^{\theta_i}/(e^{\theta_i} + e^{\theta_j})$. For more details about the Bradley-Terry model, please refer to choix manual.

Calculating Dist-n, Ent-n, LF We use *NLTK* (Loper and Bird, 2002) package while calculating Dist-n, Ent-n, and LF, particularly for tokenizing sentence and preparing n-grams. When calculating Low-Frequency Token Ratio (LF), we choose words with an occurrence count less than 100 in each dataset.

Number of Experiments We run an experiment only once since our evaluation requires a human evaluation which requires an extra annotation budget.

A.4 Additional Examples of the Semantic Clusters

We provide additional response examples of the semantic clusters in DailyDialog dataset in Table 8.

A.5 Analysis of the Distribution of Generated Responses

Figure 6 illustrates the cumulative semantic probability distributions of the generated responses. DRESS clearly shows the most similar cumulative distribution to that of uniform distribution, which is a distribution that achieves the highest Sem-Ent value. Moreover, DRESS dramatically reduces the distribution of head clusters containing generic responses compared to other baseline methods and conversely enlarges the distribution of tail clusters.

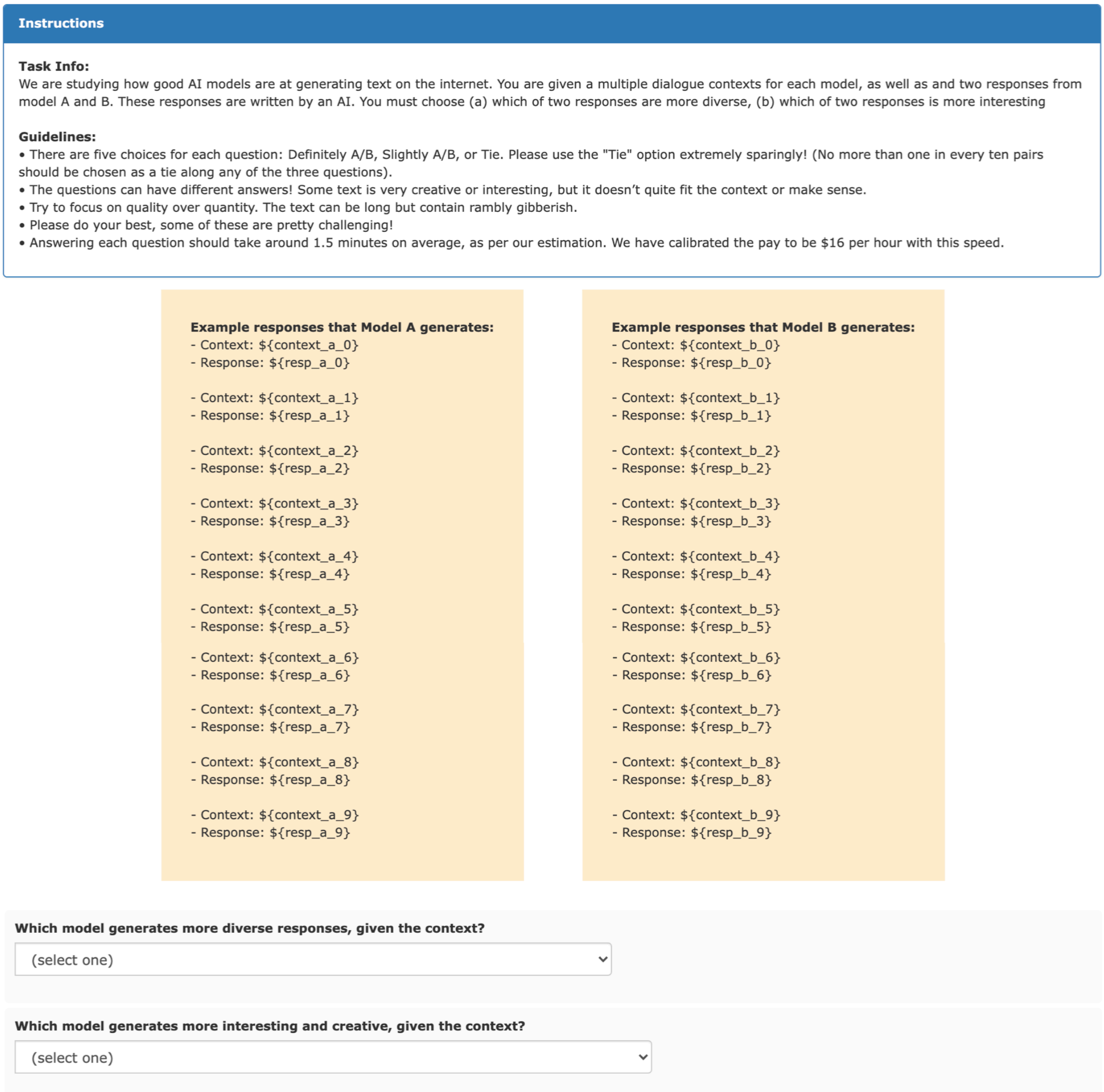

Figure 4: The interface of human evaluation for assessing how responses are (a) diverse, (b) more interesting and creative.

A.6 Limitations of our Work

In this section, we discuss the potential limitations of our methods and the experimental procedure. To start with, our proposed diversity metric Sem-Ent requires a pre-trained language model to calculate the result. This indicates that it requires relatively heavier computational resources to calculate Sem-Ent compared to other lexical-based diversity metrics such as Dist-n and Ent-n. Moreover, extending Sem-Ent to other languages or other domains could be problematic if no high-quality pre-trained language model is available on that language or domain.

In terms of the experimental procedure, we performed the experiment once rather than running it multiple times with different seeds. Since our evaluation process incorporates a human annotation, which requires a payment to human annotators, we were not able to perform multiple sets of experiments due to the limitation on budget. In the same perspective, we were not able to obtain a sufficient number of annotations to acquire statistically significant results for every pairwise comparison. We run an experiment only once since our evaluation requires a human evaluation which requires an extra annotation budget. Also, we only experimented with the English dialogue dataset (DailyDialog and English portion of the OpenSubtitles). Therefore our results do not necessarily guarantee the same

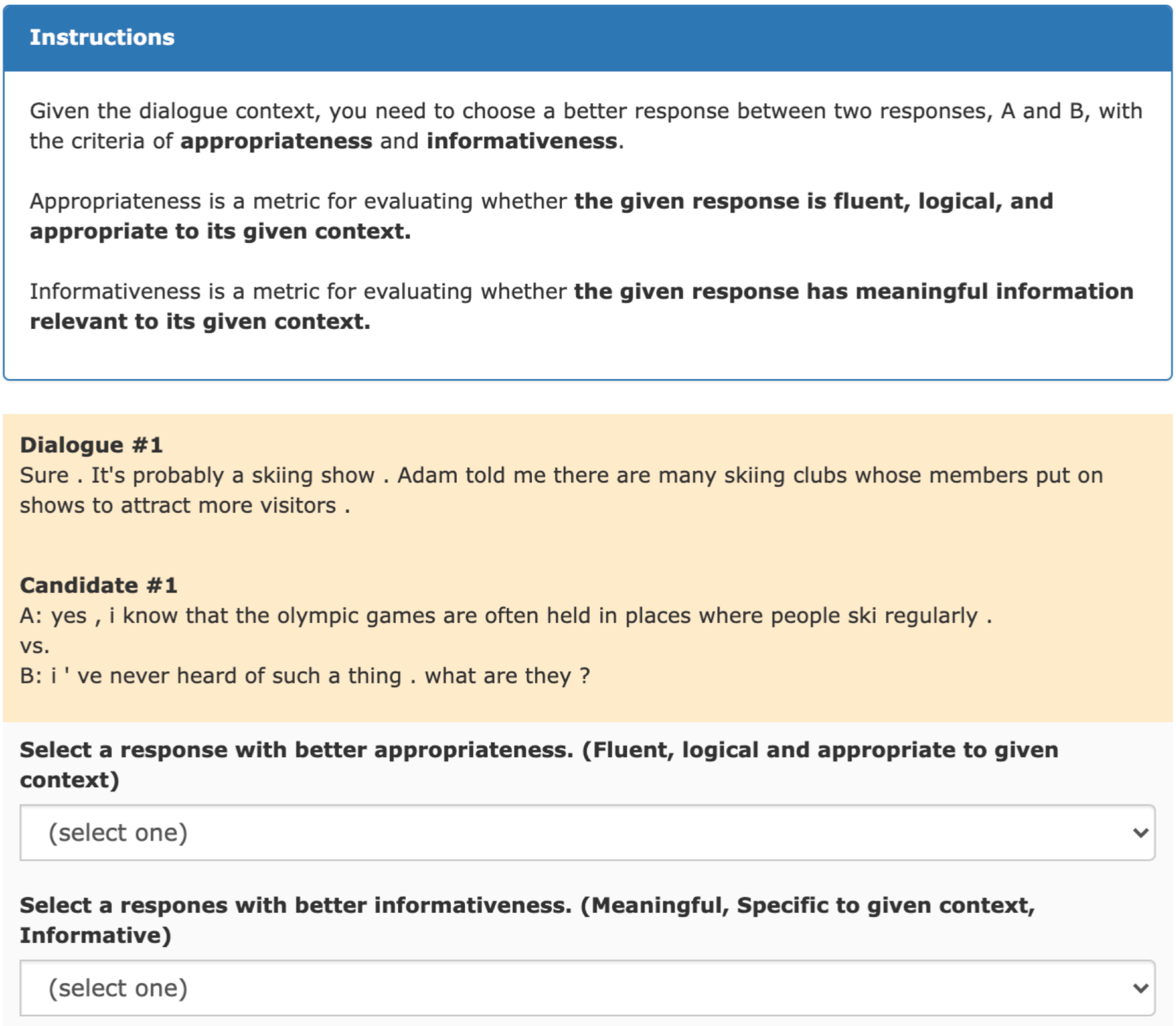

Figure 5: The interface of pairwise human evaluation for appropriateness and informativeness.

result in other languages rather than English.

Also, we'd like to clarify that our proposed metric, Sem-Ent, only focuses on measuring generated responses' diversity and does not consider the response coherency. Although this is our intention since we aimed to build an unreferenced diversity metric, this limitation yields a drawback that Sem-Ent should always be jointly used with another metric that measures the response coherency (e.g., MaUdE). Expanding Sem-Ent to consider the coherency with an input context will be an intriguing future direction for our research.

A.7 Further Implementation Details

Training Models All of our experiments are done using the ParlAI (Miller et al., 2017) framework. We leverage model weights of Blender-90M and BART-large from ParlAI. Blender-90M is pre-trained on Reddit corpus, and BART-large is pre-trained jointly on Wikipedia and Toronto Books. Note that Blender-90M has 90M parameters, and BART-large consists of 400M parameters. All baselines and DRESS use the initial learning rate of $7e-6$ with Adam optimizer, except CVAE for Blender-90M trained on DailyDialog using $2e-5$, MMI for Blender-90M trained on OpenSubtitles using $1e-6$, and CVAE for Blender-90M trained on OpenSubtitles using $1e-5$. We search for the appropriate learning rate for those exceptions since those exceptions are not stable enough to train the model. We use a learning rate scheduler that reduces its learning rate by multiplying 0.5 when the loss has stopped decreasing. All Blender-90M models and all BART-large models are trained using batch size of 32 and 16 on single A100 GPU, respectively. Training a single model takes less than a day with these configurations.

Language Model for Calculating Sem-Ent In

Backbone	Method	MaUdE (± 95% CI)
Blender-90M (DailyDialog)	Vanilla	0.8489 ± 0.005
	MMI	0.8208 ± 0.005
	CVAE	0.8552 ± 0.005
	EDF	<u>0.8593</u> ± 0.005
	NT	**0.8597** ± 0.005
	AdaLabel	0.8443 ± 0.005
	DRESS(-NT)	0.8460 ± 0.005
	DRESS	0.8575 ± 0.005
BART-large (DailyDialog)	Vanilla	0.8854 ± 0.005
	MMI	0.8787 ± 0.005
	CVAE	0.8744 ± 0.005
	EDF	0.8918 ± 0.004
	NT	0.9024 ± 0.004
	AdaLabel	0.8947 ± 0.004
	DRESS(-NT)	**0.9192** ± 0.003
	DRESS	<u>0.9153</u> ± 0.003
Blender-90M (OpenSubtitles)	Vanilla	0.8459 ± 0.004
	MMI	0.8721 ± 0.004
	CVAE	**0.9214** ± 0.003
	EDF	0.8777 ± 0.004
	NT	0.8944 ± 0.003
	AdaLabel	0.8708 ± 0.004
	DRESS(-NT)	0.8927 ± 0.003
	DRESS	<u>0.9041</u> ± 0.003
BART-large (OpenSubtitles)	Vanilla	0.7803 ± 0.005
	MMI	0.8067 ± 0.005
	CVAE	0.8600 ± 0.004
	EDF	0.8558 ± 0.004
	NT	0.8298 ± 0.005
	AdaLabel	<u>0.8647</u> ± 0.004
	DRESS(-NT)	0.8618 ± 0.004
	DRESS	**0.8873** ± 0.003

Table 6: MaUdE with a 95% confidence interval when automatically evaluating various methods.

this work, we test three language models to obtain embeddings from the response: DialoGPT, RoBERTa, and GPT2-large. For reproducibility, we utilize model weights which are publicly opened on HuggingFace Transformers (Wolf et al., 2020): `microsoft/DialoGPT-large`, `roberta-base`, and `gpt2-large` for DialoGPT, RoBERTa, and GPT2-large, respectively.

Software and Hardware We use Python 3.8, PyTorch 1.9.0 (py3.8_cuda11.1_cudnn8.0.5_0), HuggingFace Transformers 4.6.1, and ParlAI 1.3.0. All the experiments are done using NVIDIA A100-40GB GPUs, along with AMD EPYC 7742 64-Core Processors.

License The DailyDialog dataset has CC-BY-NC-SA 4.0 license. OpenSubtitles dataset does not specify the license on the dataset. For the pre-trained models, DialoGPT, RoBERTa, and GPT-2 large is all released with the MIT license. Since CC-BY-NC-SA 4.0 and MIT license both allow the utilization of the resource for research purposes, the use of these scientific artifacts in this work is valid.

Config	MaUdE (± 95% CI)
γ = 1.0	0.8570 ± 0.004
γ = 5.0	0.8593 ± 0.004
γ = 10.0	0.8583 ± 0.004
γ = 30.0	0.8575 ± 0.004
γ = 100.0	0.8436 ± 0.004
k = 10	0.8390 ± 0.004
k = 20	0.8575 ± 0.004
k = 50	0.8513 ± 0.004
k = 100	0.8637 ± 0.004

Table 7: MaUdE with a 95% confidence interval when analysing the effect of hyperparameters, γ and k.

Index	**Responses**
1	· I'm going to the store . · Oh , yes . Hi , how are you ? · All right . Hop in , please . · I am , sir . · No problem . I 'll wait for your call .
2	· Yeah . I know . · Thank you . · You are most welcome . · No more , thank you very much . · Not yet .
18	· I bought a new mattress and some fresh bedclothes . I also bought a new dressing table and a new bedside table . · I ' d prefer non·smoking roommates , but I guess I ' ll have to take what I can get ! · A single room with a front view is 100 dollars per night , one with a rear is 80 dollars .
19	· Yes . Will you also make copies and file them using both methods ? · you should probably call the IT department and have them check your computer for virus . · I see . Well , can I have a look at your phone ? Unfortunately , this phone can ' t be used in the US . it ' s not compatible with our 3G network .
20	· A driver 's license or something showing that you live in this city . · I want to change a new car . I like Honda best , especially the red one . But it is too expensive . · We use a vacuum cleaner that removes all the dirt , and we throw away all of the trash that we can find .

Table 8: Additional response examples of the semantic clusters of DailyDialog dataset. *Index* column indicates the Cluster Index in Figure 2.

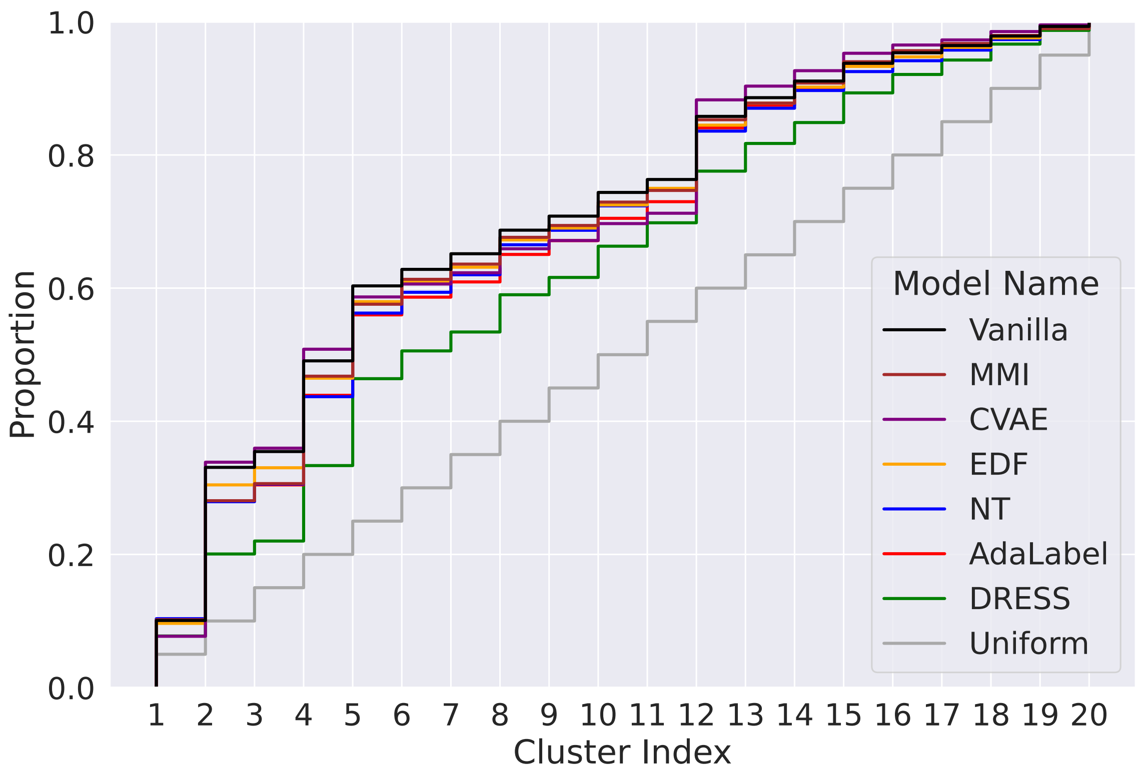

Figure 6: Cumulative probability distribution of the responses generated by different methods. *Uniform* illustrates the case of uniform cluster distribution.