

Measuring and Improving Semantic Diversity of Dialogue Generation

Seungju Han Beomsu Kim Buru Chang*

Hyperconnect

wade3han@snu.ac.kr, {beomsu.kim, buru.chang}@hpcnt.com

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 often fail to capture the semantic diversity of generated responses, as they mainly consider lexical aspects of the generated 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 captures human judgments on response diversity better than existing lexical-level diversity metrics. Furthermore, motivated by analyzing an existing dialogue dataset, we propose a simple yet effective learning method that improves the semantic diversity of generated responses. Our learning method weights training samples based on the semantic distribution of the training set. We show that our learning method improves response diversity and coherency better than other baseline methods through automatic and human evaluation.

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., 2019; Roller et al., 2021) in the last decade. However, although dialogue generation models can produce fluent responses, they are also known for frequently generating dull and uninformative generic responses (e.g., "I don't know"), degrading their engagingness (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.

*Corresponding author

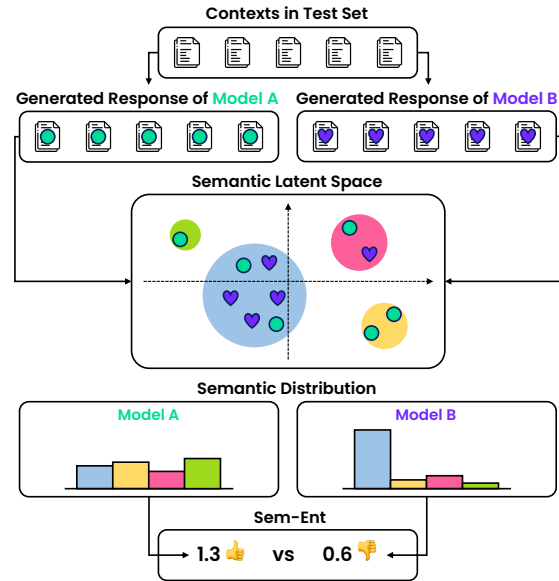


Figure 1: An illustration of our proposed Sem-Ent that measures semantic diversity based on the semantic distribution of generated responses.

The current evaluation protocol has relied on lexical-level evaluation metrics such as Distinct- n (Dist- n) (Li et al., 2016a) and Entropy- n (Ent- n) (Serban et al., 2017) to measure the diversity of generated responses. However, according to recent studies (Tevet and Berant, 2021; Stasaski and Hearst, 2022), these lexical-level evaluation metrics often fail to capture *semantic diversity* since responses including similar words can have different semantics and responses with different words can have similar semantics (Yarats and Lewis, 2018).

In this paper, we propose **Sem-Ent** (**Semantic-Entropy**), 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 pretrained language model (e.g., DialoGPT (Zhang et al., 2020) and BERT (Devlin et al., 2019)). Then, the evaluation metric measures the semantic diversity of generated responses by calculating how the responses are evenly distributed in the semantic

latent space based on entropy, 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.

Furthermore, we propose a simple yet effective learning method of dialogue generation models to improve semantic diversity of generated responses. We observe that the semantic distribution of responses in a dialogue dataset is highly imbalanced, leading dialogue generation models to produce semantically less diverse responses. To address this problem, our proposed method, **DRESS** (**D**iversifying **R**ESponses **S**emantically), learns 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 better semantic diversity compared to state-of-the-art baseline methods, along with the gain in response coherency. Interestingly, DRESS also achieves better performance in lexical-level diversity metrics than baselines, even though it focuses only on improving the semantic diversity. Moreover, human evaluation results show the effectiveness of DRESS, where DRESS outperforms all baseline methods in 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 judgments 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) An implementation of Sem-Ent will be released, contributing to the community of open-domain dialogue generation.

2 Related Work

2.1 Enhancing Response Diversity

Since generating dull and uninformative responses is a well-known and important problem in open-domain dialogue (Vinyals and Le, 2015; Li et al., 2016a), numerous methods have been proposed to address this issue. The maximum mutual information objective function is utilized to penalize generic responses and improve the diversity of gen-

erated responses (Li et al., 2016a,c; 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 set (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 also known to improve the diversity of generated responses. Wang et al. (2021) diversify responses 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 using reference responses generated by human annotators. 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. Low-Frequency (LF) (Li et al., 2019) calculates the frequency of low-frequency words in generated responses as the response diversity.

Semantic diversity. Recently, several studies have focused on the semantic diversity of generated responses. Tevet and Berant (2021) release the McDiv benchmark to evaluate the semantic diversity

metrics and [Stasaski and Hearst \(2022\)](#) propose a new semantic diversity metric, natural language inference (NLI) diversity, leveraging pretrained NLI models ([Bowman et al., 2015](#)).

The major difference between our Sem-Ent and NLI diversity is that NLI diversity can only capture the semantic diversity of generated responses for a single context, while Sem-Ent measures the overall semantic diversity of generated responses for multiple contexts of the test set. This is an important distinction since the latter provides insight into how well generated responses vary depending on which context is provided as an input while the former cannot. To see the shortcoming of NLI diversity more clearly, take the following an example: suppose that given a context c_a as an input, a dialogue generation model generated responses $\{r_{a,1}, r_{a,2}, \dots\}$ that are "semantically diverse" according to NLI diversity. Now, further suppose that given another context c_b , the model generates responses $\{r_{b,1}, r_{b,2}, \dots\}$ that are also "semantically diverse" among themselves but appear similar to the responses $\{r_{a,1}, r_{a,2}, \dots\}$ produced for the context c_a . In such a case, NLI diversity cannot capture the fact that the generated responses $\{r_{a,1}, \dots\}$ and $\{r_{b,1}, \dots\}$ for the contexts c_a and c_b are semantically similar despite the two contexts being different contexts; Sem-Ent can because it measures the semantic diversity of generated responses for a set of different contexts of the test set. In the next Section 3, we will describe our proposed semantic diversity metric, Sem-Ent, in detail.

3 Measuring Semantic Diversity

3.1 Sem-Ent

Let $\mathcal{D} = \{(c_i, r_i)\}_{i=1}^m$ denote a training set 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 set $\tilde{\mathcal{D}} = \{(\tilde{c}_i, \tilde{r}_i)\}_{i=1}^n$ consisting of n dialogues. During evaluation, a dialogue generation model M generates responses $\mathcal{R}^M = \{r_i^M\}_{i=1}^n$ for the contexts $\{\tilde{c}_i\}_{i=1}^n \in \tilde{\mathcal{D}}$, respectively. To compute semantic diversity, Sem-Ent requires a semantic distribution $P(\mathcal{R}^M)$, but there is no direct way to obtain the exact distribution. Thus, we approximate the semantic distribution $P(\mathcal{R}^M)$ as $\tilde{P}(\mathcal{R}^M) = [\tilde{p}(1); \dots; \tilde{p}(k)]$ using the the semantic clusters \mathcal{C} as follows:

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

where $\phi_{\mathcal{C}}(r) \in \{1, \dots, k\}$ is a cluster mapping function that returns the cluster index of a response r 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 $\tilde{P}(\mathcal{R}^M)$ that approximates the semantic distribution $P(\mathcal{R}^M)$ 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 borrow a pairwise experimental setup of [Pillutla et al. \(2021\)](#) for analyzing the correlation between diversity metrics and human judgments. Our evaluation is based on the observation that the degree of response diversity varies depending on the types of generative models and decoding algorithms ([Holtzman et al., 2019](#); [Tevet and Berant, 2021](#)). From this, we first prepare eight different response generation 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

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 diversity matrices with human judgments on response diversity. BT denotes the Bradley-Terry score for a pairwise human evaluation and the value inside the parenthesis indicates p-value. We set the number of semantic clusters as $k=20$. Evaluation results with different n for Dist- n and Ent- n are reported in Appendix A.3.

sampling, and nucleus sampling). We then obtain 28 pairs of generation settings from the eight response generation settings.

For each pair, we randomly choose ten contexts from the test set of a DailyDialog dataset (Li et al., 2017b) and generate two response sets using the two generation settings, respectively, for the ten contexts. Human annotators are asked to select which response set is better in two criteria, diversity and interestingness, using a 5-point Likert scale. We obtain 25 pairwise annotations for each pair of response generation settings. These annotation results are converted into each response generation setting’s score by using the Bradley-Terry model (Marden, 1996). By Bradley-Terry model, the probability of the outcome $i \succ j$ is calculated as $p(i \succ j) = e^{\theta_i} / (e^{\theta_i} + e^{\theta_j})$ when parameters $\theta_1, \dots, \theta_n$, for two items i and j are given. For more details about the Bradley-Terry model, please refer to choix manual¹.

We measure the correlation between the Bradley-Terry score and diversity metrics to check how each metric correlates with the human judgments on each criterion. More details about human evaluation are included in Appendix A.

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 the 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 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. Note that we do not set NLI diversity as a baseline because it

is incompatible with our human evaluation which measures the overall semantic diversity.

Results. Table 1 shows the correlation between the human judgments and diversity metrics in terms of Pearson and Spearman rank correlation. Our Sem-Ent shows the highest Pearson and Spearman rank correlation with human judgments 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 estimating human judgments 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 judgments on response diversity. This result implies that generated responses that have similar representations to human responses (i.e., high MAUVE scores) are not always semantically diverse since human responses are also often generic (Csáky et al., 2019) (further analyzed in Section 4.1).

We also observe that Sem-Ent shows a high correlation with human judgment on interestingness (See et al., 2019); Ask annotators about how interesting or boring did they find about the generated response. Sem-Ent has a similar correlation to Ent- n and shows a substantially higher correlation than Dist- n , LF, and MAUVE. We believe that semantically diverse responses could improve the interestingness of dialogue generation models and Sem-Ent could somewhat capture human judgments on this response interestingness.

Robustness of Sem-Ent to the Choice of Configuration. Sem-Ent could be affected by the changes in the configurations used for calculating the score, such as the types of language models and the number of semantic clusters k . We conduct an additional evaluation to examine the robustness of Sem-Ent to the choice of the configurations, By chang-

¹<https://github.com/lucasmaystre/choix>

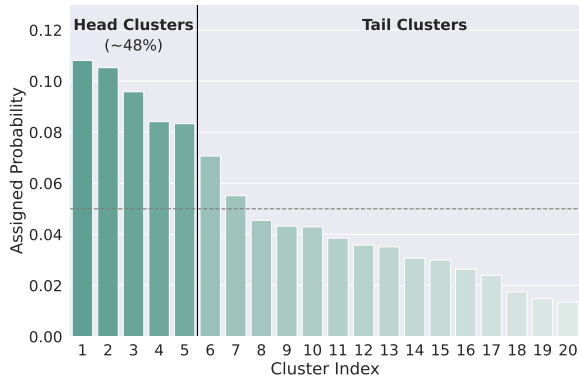


Figure 2: Semantic distribution of responses in the training set. Semantic clusters are sorted in the descending order of the assigned probabilities. The dashed line indicates the uniformly distributed probability, 0.05.

ing the configurations, we obtain multiple Sem-Ent scores of generated responses and then analyze the correlation between the scores. The more details of the additional evaluation and results are described in Appendix B. We observe that Sem-Ent shows a high correlation among the obtained scores, verifying the robustness of Sem-Ent against the choice of configurations.

4 Improving Semantic Diversity

4.1 Diagnosing the Semantic Distribution of Dialogue Dataset

As described in Section 3, semantic distribution of responses provides a crucial clue for understanding the diversity of the responses. Therefore, we analyze the semantic distribution of the responses \mathcal{R} in the training set of the DailyDialog dataset. Figure 2 shows that the semantic distribution of \mathcal{R} is highly skewed – almost half of the responses fall into the 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 shown in Table 2. Conversely, responses in the infrequent clusters have a wider variety of topics, intents, and diverse vocabularies. Since the training set is skewed towards semantically generic and dull responses, naively training with this data will lead to the low semantic diversity of generated responses.

4.2 DRESS

We introduce a simple yet effective learning method of dialogue generation models for improving semantic diversity, DRESS, which addresses

Index	Responses
1	<ul style="list-style-type: none"> • Tell me . • You're welcome . • Bye . • What's up ?
2	<ul style="list-style-type: none"> • Yeah . I know . • Thank you . • That's cool . • Not yet .
19	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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 2: Response examples of the semantic clusters. *Index* column indicates the index of semantic cluster in Figure 2.

the problem of the imbalanced semantic distribution by weighting the responses in the training set. The purpose of DRESS is simple: inducing generation models to learn more about responses in the infrequent semantic clusters and 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:

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

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

Moreover, to penalize responses in frequent semantic clusters intensively, we jointly utilize Negative Training (NT) (He and Glass, 2020; Li et al., 2020) with the weighted objective function. For every epoch, the model generates responses to each given context. If generated responses are included in head clusters, 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 DRESS successfully improves response diversity.

Dataset. We utilize two open-domain dialogue datasets: DailyDialog (Li et al., 2017b) 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 sets. 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). 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 generated responses.

Human Evaluation. We further conduct a pair-wise comparison through the human evaluation for evaluating the general conversation ability of generation models since automatic evaluations are sometimes not trustworthy. We use Amazon Mechanical Turk to collect the annotations. Following many previous studies (Gao et al., 2019; He and Glass, 2020; Han et al., 2022), each annotator evaluates which response 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. Note that *Diversity* cannot be assessed in this pair-wise evaluation setup. More details about evaluation protocol (e.g., interface for collecting annotation) are described in Appendix C.

5.2 Baseline Methods

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

CVAE (Zhao et al., 2017) builds the response generation process as a conditional variational auto-encoder of a response with dialogue context to increase response diversity.

EDF (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 source side entropy to filter the pairs.

NT (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 over-confidence 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

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	<u>0.2389</u>	<u>0.4459</u>	7.4074	11.2797	12.9969	0.2449	0.8552	2.6261
	EDF	<u>0.0473</u>	0.2271	0.4226	7.2888	11.0283	12.7132	0.2402	<u>0.8593</u>	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	<u>11.3273</u>	<u>13.1028</u>	<u>0.2474</u>	0.8460	<u>2.7576</u>
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	<u>0.0571</u>	<u>0.2919</u>	0.5424	8.0267	12.3098	14.0577	<u>0.3070</u>	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	<u>0.5448</u>	8.1722	<u>12.5195</u>	<u>14.3244</u>	0.3079	0.9192	<u>2.8444</u>
DRESS	0.0547	0.2906	0.5504	8.1821	12.5533	14.3890	0.3052	<u>0.9153</u>	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	<u>0.0504</u>	<u>0.2216</u>	<u>0.3969</u>	<u>7.3734</u>	<u>11.0928</u>	<u>12.6594</u>	<u>0.2480</u>	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	<u>2.7114</u>
DRESS	0.0524	0.2351	0.4180	7.5113	11.2355	12.7612	0.2612	<u>0.9041</u>	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	<u>0.2056</u>	<u>0.3572</u>	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	<u>0.8647</u>	2.5652
	DRESS(-NT)	0.0456	0.2006	0.3509	<u>7.1669</u>	<u>10.6915</u>	<u>12.1509</u>	<u>0.2220</u>	0.8618	<u>2.6620</u>
DRESS	<u>0.0472</u>	0.2178	0.3890	7.4656	11.2761	12.8601	0.2322	0.8873	2.7406	

Table 3: Automatic evaluation results in various diversity metrics (Dist- n , Ent- n , LF, and Sem-Ent) and coherency metric (an average MaUdE of generated responses). **Bold** and underline indicate the best and runner-up results, respectively. DRESS(-NT) indicates the variant version of DRESS that only utilizes the weighted NLL without NT.

the underlying generation models to demonstrate that our method widely works well on different architectures. For DRESS, we find the optimal hyperparameters for the number of semantic clusters k and the weighting factor γ through the grid search from $k \in \{10, 20, 50, 100\}$ and $\gamma \in \{1.0, 5.0, 10.0, 30.0, 100.0\}$, and we set them as $k=20$ and $\gamma=30$ 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 code² to help researchers reproduce the result.

²<https://github.com/hyperconnect/sem-ent>

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 diversities while showing high response coherency for most of the experimental setups. To be more specific, DRESS shows a substantially higher semantic diversity (Sem-Ent) than all other baseline models in every experimental setup. It is quite intriguing that DRESS also achieves better performance in lexical-level response diversity (Dist- n , Ent- n , and LF). This observation demonstrates that improving semantic diversity entails the improvement of lexical-level diversity. Furthermore, MaUdE results indicate that DRESS preserves better response coherency compared to other baseline methods.

Table 4 summarizes the pairwise human evalua-

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 pairwise 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.

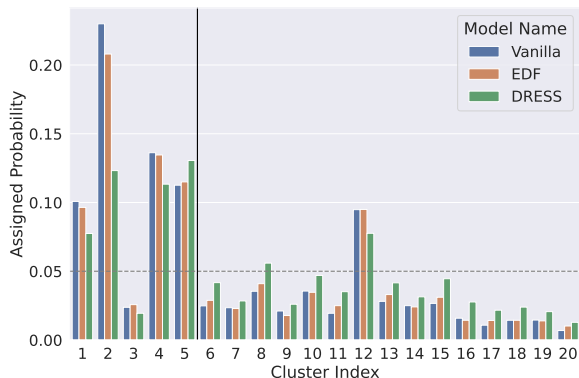


Figure 3: Probability distribution of the responses generated by Vanilla, EDF and DRESS. The dashed line indicates the uniformly distributed probability, 0.05.

tion results. DRESS shows better conversation abilities in appropriateness and informativeness than other baseline methods while improving response diversity (as shown in the automatic evaluation).

6.2 Analysis of Experimental Results

Semantic Distribution of Generated Responses.

To analyze how DRESS increase the semantic diversity, we compare the semantic distribution of responses generated by DRESS with that of Vanilla and EDF. Figure 3 illustrates the detailed semantic distribution of the generated responses of Blender-90M on DailyDialog. 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 except Cluster 5 and boosts probabilities of the tail clusters except Cluster 12, where the only one has a higher assigned probability than 0.05 in the tail clusters.

Changing Hyperparameters of DRESS. As described in Section 5.3, we found the optimal hyperparameter pair ($\gamma=30$, $k=20$) for Blender-90M on DailyDialog. To examine the effect of each hyperparameter, we observe the change of evaluation

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: Analyzing the effect of γ and k .

results by varying the hyperparameters. Note that we fix γ to 30.0 and k to 20 when varying k and γ , respectively. Table 5 shows the results about the effect of the hyperparameters. We find that increasing γ induces dialogue generation 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 dialogue generation models to generate more diverse responses. However, MaUDE gets degraded while response diversity improves implying a trade-off between response diversity and coherence.

Ablation Study. To verify the effect of our weighted NLL, we conduct an ablation study. In Table 3, DRESS(-NT) indicates the variant of DRESS without NT 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 baselines excluding DRESS. Moreover, DRESS(-NT) also shows a higher lexical-level diversity than other baselines, along with high MaUDE scores. From these observations, we verify the effectiveness of our proposed weighted NLL that semantically diversifies generated responses.

7 Conclusion

In this work, 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 shows a high correlation with human judgments on interestingness. Moreover, we introduce a new learning method, DRESS, to mitigate the problem of the imbalanced semantic distribution of dialogue datasets. Evaluation results show that DRESS improves both the semantic and lexical-level diversities of generated responses, along with the gain in response coherency.

Limitations of This 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 in 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 could not perform multiple sets of experiments due to the limited budget. We could not obtain a sufficient number of annotations to acquire statistically significant results for every pairwise comparison. In the same perspective, we conducted the human evaluation in only criteria *Appropriateness* and *Informativeness*. We cannot include further criteria, such as *Diversity* and *Interestingness*, since these criteria require further evaluation setups requiring a considerable annotation cost. We run an experiment only once since our evaluation requires a human evaluation which requires an extra annotation budget. Furthermore, we only experimented with the English dialogue dataset (DailyDialog and the English portion of the OpenSubtitles). Therefore our results do not necessarily guarantee the same result in other languages rather than English.

At last, we would like to clarify that our proposed metric, Sem-Ent, only focuses on measuring the response diversity and does not consider the response coherency. Although this is our intention

since we aim 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.

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 from the 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.
- Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.
- Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. *Generating multiple diverse responses with multi-mapping and posterior mapping selection*. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4918–4924. International Joint Conferences on Artificial Intelligence Organization.
- 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. 2022. [Understanding and improving the exemplar-based generation for open-domain conversation](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 218–230, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Iliia 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, Iliia 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*.

- 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.
- 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. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- 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.
- Katherine Stasaski and Marti A Hearst. 2022. Semantic diversity in dialogue with natural language inference. *arXiv preprint arXiv:2205.01497*.
- Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346.
- 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.

Appendix

A Details of Human Evaluation for Sem-Ent (Section 3.2)

A.1 Human Annotation

To collect human annotations, we use Amazon Mechanical Turk, and Figure 4 shows the instructions and the interface for the human annotators. To mitigate the bias from the annotators, we set a maximum number of annotations per annotator as 20 and randomly shuffle the order of the response generation settings and the corresponding response. Since our task does not require particular expertise in linguistics, we open the annotations 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, as following Kim et al. (2021); Han et al. (2022). We estimate 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. As a result, annotators are paid USD 0.40 per HITs.

A.2 Baseline Diversity Metrics

To calculate Dist- n , Ent- n , and LF, we use *NLTK* package (Loper and Bird, 2002) for tokenizing responses and preparing n -grams. When calculating LF, we choose words with an occurrence count of less than 100 in each dataset.

A.3 Further Evaluation Results on Different n

We further report the experimental results with different n for Dist- n and Ent- n (Table 6). Like the results in Table 1, our proposed Sem-Ent also shows the higher correlation with human judgements on response diversity than Dist- n and Ent- n .

A.4 Why We Did Not Compare Sem-Ent with NLI Diversity (Stasaski and Hearst, 2022)

As we propose a new diversity metric (Sem-Ent), we also understand that it is required to compare Sem-Ent with a recently proposed diversity metric (NLI Diversity) to verify its effectiveness. However, we would like to emphasize that the two diversity metrics target the different evaluation settings.

NLI Diversity targets the sample-wise semantic diversity setting where the metric measures whether semantically diverse responses are generated for a single test sample. Thus, Stasaski and Hearst (2022) conducted experiments on the McDiv benchmark (Tevet and Berant, 2021), which evaluates the semantic diversity metrics in the sample-wise semantic diversity setting to show the effectiveness of NLI Diversity. On the other hand, our Sem-Ent targets the overall semantic diversity setting that assesses the diversity of generated responses throughout the test set. A higher Sem-Ent score indicates that the generation model covers a wide range of semantic topics throughout the test set. Therefore, our human evaluation simulates the overall semantic diversity setting, unlike the McDiv benchmark, as shown in Figure 4. Note that NLI Diversity is incompatible with this evaluation setting because dialogue generation models should produce multiple responses for each test sample. Moreover, there is no guarantee that generated responses throughout the test set will be semantically diverse even if dialogue generation models show high diversity for each test sample. We believe that our proposed Sem-Ent and NLI Diversity can be used complementary to each other in evaluating the response diversity without conflict.

B Robustness of Sem-Ent (Section 3.2)

We examine the robustness of Sem-Ent by changing the configurations used for calculating the metric. Several configurations can be changed in Sem-Ent, including the types 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 (Roller et al., 2021) for the test set of DailyDialog (Li et al., 2017b) with all methods (in Table 3). We then measure the Spearman rank correlation between the computed Sem-Ent scores of the different configurations.

For the choice of language models, we compare three variants: DialoGPT (Zhang et al., 2020), RoBERTa (Liu et al., 2019a), and GPT2-large (Radford et al., 2019). The average Spearman rank 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 scores ranked by Sem-Ent. The average Spearman rank correlation between these

Metric	Correlation	Dist-1	Dist-2	Dist-4	Ent-1	Ent-2	Ent-4	Sem-Ent
Diversity/BT	Pearson	-0.043 (0.918)	0.258 (0.536)	0.292 (0.481)	0.518 (0.187)	0.454 (0.257)	0.394 (0.333)	0.810 (0.015)
	Spearman	0.190 (0.651)	0.380 (0.351)	0.380 (0.351)	0.595 (0.119)	0.476 (0.232)	0.380 (0.351)	0.762 (0.028)
Interesting/BT	Pearson	-0.143 (0.734)	0.161 (0.702)	0.198 (0.637)	0.469 (0.240)	0.392 (0.335)	0.323 (0.434)	0.789 (0.020)
	Spearman	0.142 (0.735)	0.380 (0.351)	0.380 (0.351)	0.642 (0.085)	0.523 (0.182)	0.380 (0.351)	0.667 (0.020)

Table 6: Further evaluation results on different n for Dist- n and Ent- n .

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 diversity metric against the choice of configurations.

C Details of Human Evaluation for DRESS (Section 5)

C.1 Human Annotation

To collect human annotations for verifying the effectiveness of our proposed DRESS, we also use Amazon Mechanical Turk and the same setting in Section A.1 to mitigate the bias and control the quality of the human annotations. Figure 5 shows the instructions and the interface for the human evaluation. Here, human annotators are paid USD 0.25 per HITs as we estimate 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 this annotation is easier than the human annotations for Sem-Ent in Section A.1.

C.2 Why We Could Not Set *Diversity* as a Human Evaluation Criterion

Unlike *Appropriateness* and *Informativeness*, annotators require multiple conversations to annotate *Diversity* of generation models because it is hard to capture the response diversity from a single conversation. Additional Amazon Mechanical Turk sessions should be conducted besides our pairwise human evaluation to obtain annotations on *Diversity*. However, as described in Limitation Section, it was challenging under our limited human evaluation budget. Therefore, we conducted a pairwise human evaluation on only *Appropriateness* and *Informativeness* following the most common evaluation setup, as shown in Figure 5. We hope that this experimental result will be used as a reference to help understand the general conversational ability of our proposed DRESS, even though DRESS is proposed to increase semantic diversity.

D Implementation Details (Section 3.2 and Section 5)

D.1 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 provided in the ParlAI framework. Blender-90M is pretrained on Reddit corpus, and BART-large is pretrained 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 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 a batch size of 32 and 16 on a single A100 GPU, respectively. Training a single model takes less than a day with these configurations.

D.2 Language Models for Calculating Sem-Ent

In this work, we test three language models to obtain semantic representations of responses: DialoGPT, RoBERTa, and GPT2-large. For reproducibility, we utilize model weights, which are publicly provided in HuggingFace Transformers (Wolf et al., 2020): `microsoft/DialoGPT-large`, `roberta-base`, and `gpt2-large`, for DialoGPT, RoBERTa, and GPT2-large, respectively.

D.3 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

Instructions

Task Info:
 We are studying how good AI models are at generating text on the internet. You are given a multiple dialogue contexts for each model, as well as two responses from model A and B. These responses are written by an AI. You must choose (a) which of two responses are more diverse, (b) which of two responses is more interesting

Guidelines:

- There are five choices for each question: Definitely A/B, Slightly A/B, or Tie. Please use the "Tie" option extremely sparingly! (No more than one in every ten pairs should be chosen as a tie along any of the three questions).
- The questions can have different answers! Some text is very creative or interesting, but it doesn't quite fit the context or make sense.
- Try to focus on quality over quantity. The text can be long but contain rambly gibberish.
- Please do your best, some of these are pretty challenging!
- Answering each question should take around 1.5 minutes on average, as per our estimation. We have calibrated the pay to be \$16 per hour with this speed.

Example responses that Model A generates:

- Context: \${context_a_0}
- Response: \${resp_a_0}

- Context: \${context_a_1}
- Response: \${resp_a_1}

- Context: \${context_a_2}
- Response: \${resp_a_2}

- Context: \${context_a_3}
- Response: \${resp_a_3}

- Context: \${context_a_4}
- Response: \${resp_a_4}

- Context: \${context_a_5}
- Response: \${resp_a_5}

- Context: \${context_a_6}
- Response: \${resp_a_6}

- Context: \${context_a_7}
- Response: \${resp_a_7}

- Context: \${context_a_8}
- Response: \${resp_a_8}

- Context: \${context_a_9}
- Response: \${resp_a_9}

Example responses that Model B generates:

- Context: \${context_b_0}
- Response: \${resp_b_0}

- Context: \${context_b_1}
- Response: \${resp_b_1}

- Context: \${context_b_2}
- Response: \${resp_b_2}

- Context: \${context_b_3}
- Response: \${resp_b_3}

- Context: \${context_b_4}
- Response: \${resp_b_4}

- Context: \${context_b_5}
- Response: \${resp_b_5}

- Context: \${context_b_6}
- Response: \${resp_b_6}

- Context: \${context_b_7}
- Response: \${resp_b_7}

- Context: \${context_b_8}
- Response: \${resp_b_8}

- Context: \${context_b_9}
- Response: \${resp_b_9}

Which model generates more diverse responses, given the context?

(select one)

Which model generates more interesting and creative, given the context?

(select one)

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

Processors.

D.4 License

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

E Further Analysis (Section 6)

E.1 Confidence Interval of MaUdE Scores

In Table 3 and Table 5, 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 7 and Table 8. 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, we can not calculate the confidence interval of the diversity metrics.

E.2 Inter-Annotator 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)

Instructions

Given the dialogue context, you need to choose a better response between two responses, A and B, with the criteria of **appropriateness** and **informativeness**.

Appropriateness is a metric for evaluating whether the given response is fluent, logical, and appropriate to its given context.

Informativeness is a metric for evaluating whether the given response has meaningful information relevant to its given context.

Dialogue #1

Sure . It's probably a skiing show . Adam told me there are many skiing clubs whose members put on shows to attract more visitors .

Candidate #1

A: yes , i know that the olympic games are often held in places where people ski regularly .
vs.
B: i ' ve never heard of such a thing . what are they ?

Select a response with better appropriateness. (Fluent, logical and appropriate to given context)

(select one)

Select a responses with better informativeness. (Meaningful, Specific to given context, Informative)

(select one)

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

show that inter-annotator reliability of annotation results using crowd-sourced annotators (such as in 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, and we limited the number of maximum annotations that a single annotator can be assigned to reduce the bias, which might have increased inter-annotator diversity.

E.3 Additional Response Examples in Figure 3 (Section 6.2)

We observe the unexpected evaluation results in Figure 3. Cluster 3 shows a lower assigned probability than 0.05. On the other hand, Cluster 12 shows a higher assigned probability than 0.05. Unexpectedly, DRESS increases the assigned probability of Cluster 5 in the head clusters. Therefore, we provide response examples in the corresponding clusters, in Table 9. Cluster 3 includes the many generic responses still and Clusters 5 & 12 include the responses in specific topics such as countries (Cluster 5) and purchase (Cluster 12). Unfortunately, we fail to find the reasons for the unexpected experimental results although we examine the generated responses. We conjecture that there is a gap between the semantic distributions of contexts in the training set and the test set in the DailyDialog dataset where the test set includes many contexts belonging to Cluster 5 & 12 than the training set.

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

Table 7: MaUdE with a 95% confidence interval when automatically evaluating various methods.

E.4 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 the head clusters containing generic responses compared to other baseline methods and conversely enlarges the distribution of the tail clusters.

Config	MaUdE (\pm 95% CI)
$\gamma = 1.0$	0.8570 ± 0.004
$\gamma = 5.0$	0.8593 ± 0.004
$\gamma = 10.0$	0.8583 ± 0.004
$\gamma = 30.0$	0.8575 ± 0.004
$\gamma = 100.0$	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 8: MaUdE with a 95% confidence interval when analysing the effect of hyperparameters, γ and k .

Index	Responses
3	<ul style="list-style-type: none"> • That is wonderful . • That is a wonderful choice . • Sounds painful ! • Thanks . That ' s terrific ! • Thanks . This is fun !
5	<ul style="list-style-type: none"> • OK, It's here , one of the best makes in China . • Well , my whole family is in the United States now , but we're from Costa Rica originally . • I surely do . They must have had advanced machines in ancient China to do that .
12	<ul style="list-style-type: none"> • Here is your change and your receipt . Do you have goods unpaid on you , sir . • We found your samples very attractive . We ' re interested in buying your garments if your prices are reasonable . • A postcard costs you five yuan . A dozen post-cards cost you 60 yuan .

Table 9: Additional response examples of the semantic clusters of the test set in the DailyDialog dataset. *Index* column indicates the Cluster Index in Figure 2.

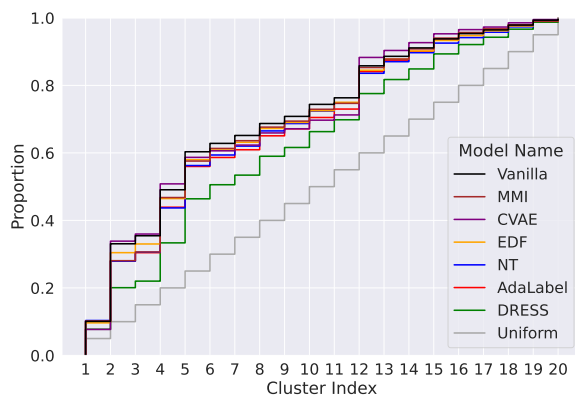


Figure 6: Cumulative probability distribution of the responses generated by different methods. *Uniform* illustrates the case of uniform cluster distribution.