DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances

Xiaodong Gu[1], Kang Min Yoo[2] and Jung-Woo Ha[2]

[1]School of Software, Shanghai Jiao Tong University, China
[2]NAVER AI Lab, Korea
xiaodong.gu@sjtu.edu.cn, {kangmin.yoo, jungwoo.ha}@navercorp.com

Abstract

Recent advances in pre-trained language models have significantly improved neural response generation. However, existing methods usually view the dialogue context as a linear sequence of tokens and learn to generate the next word through token-level self-attention. Such token-level encoding hinders the exploration of discourse-level coherence among utterances. This paper presents DialogBERT, a novel conversational response generation model that enhances previous PLM-based dialogue models. DialogBERT employs a hierarchical Transformer architecture. To efficiently capture the discourse-level coherence among utterances, we propose two training objectives, including masked utterance regression and distributed utterance order ranking in analogy to the original BERT training. Experiments on three multi-turn conversation datasets show that our approach remarkably outperforms the baselines, such as BART and DialoGPT, in terms of quantitative evaluation. The human evaluation suggests that DialogBERT generates more coherent, informative, and human-like responses than the baselines with significant margins.

Introduction

Multi-turn open-domain dialogue modeling is an active research topic in the field of natural language processing. However, generating a coherent and informative response for a given dialogue context remains a challenge. A critical challenge is the learning of rich and robust context representations of dialogue utterances (Ortega and Vu 2017; Pragst et al. 2018), namely the challenge of encoding a dialogue context into a vector that adequately captures the semantics (e.g., topic, intention).

Large-scale pre-training language models using Transformer-based architectures have recently achieved remarkable successes in a variety of NLP tasks (Devlin et al. 2018; Sun et al. 2019; Zhang, Wei, and Zhou 2019; Zheng et al. 2020). As such, there are increasingly work that aims to use pre-training language models for conversation modeling (Mehri et al. 2019; Zhang et al. 2019; Rothe, Narayan, and Severyn 2019). For example, DialoGPT (Zhang et al. 2019) extends the GPT-2 (Radford et al. 2019) to generate conversation responses on large-scale dialogue corpus. Meena (Adiwardana et al. 2020) trains

a sequence-to-sequence model (Cho et al. 2014) with the Evolved Transformer (So, Liang, and Le 2019) on large-scale multi-turn conversations. Blender, developed by Facebook, provides recipes for building open-domain chatbots that perform well in human evaluations (Roller et al. 2020).

However, existing pre-training conversation models usually view the dialogue context as a linear sequence of tokens and learns to generate the next word through token-level self-attention. One issue of this approach is that the high-level relationships between utterances are harder to capture using word-level semantics. For example, the discourse-level relationship between the utterances *"coffee please"* and *"here you are"* (Figure 1) is apparent, but word-level comparisons, such as $<coffee, you>$ and $<please, are>$, obscures the high-level relationship. Furthermore, this full pairwise attention is inefficient since it requires each word in the context and the decoder to interact with all other words regardless of their distances and semantic units.

To alleviate the issues above, we present DialogBERT, a novel conversational response generation model. DialogBERT employs a hierarchical Transformer architecture to represent the dialogue context. It first encodes dialogue utterances through a Transformer encoder and then encodes the resulting utterance vectors using a discourse-level Transformer to obtain a representation of the entire dialogue context. To efficiently capture discourse-level coherence among utterances, we propose two training objectives in analogy to the original BERT training: 1) **masked context regression**, which masks a randomly-selected utterance and predicts the encoding vector for the masked utterance directly; and 2) **distributed utterance order ranking**, which organizes randomly shuffled utterances of a conversation into a coherent dialogue context through a *Learning-to-Rank* (Cao et al. 2007) neural network.

We evaluate DialogBERT on popular multi-turn conversation datasets, namely Weibo, MultiWOZ and DailyDialog. Results show that DialogBERT outperforms baselines in terms of perplexity, BLEU, and NIST. Human evaluation supports the superiority of our approach in capturing discourse-level semantics and generating more plausible dialogue responses.

Related Work

This work is closely related to (1) pre-trained language models, (2) pre-trained models for conversations, and (3) adopting auxiliary multi-task objectives for improving pre-trained language models.

Pre-trained Language Models. The current paradigm has gradually evolved from word vectors (Pennington, Socher, and Manning 2014) and contextualized word embedding models (Peters et al. 2018). Recent works have explored various architecture choices and training objectives for large-scale pre-trained language models (Devlin et al. 2018; Radford et al. 2019; Yang et al. 2019) based on Transformers (Vaswani et al. 2017). A recent work proposed using the denoising autoencoder framework with composite corruption schemes (Lewis et al. 2019).

Pre-trained Models for Dialogue Generation. Recent advances in pre-trained language models have spurred success in dialogue response generation. Specifically, Budzianowski and Vulić (2019) explored the use of pre-trained language Transformers for task-oriented dialogues. Wolf et al. (2019b) also proposed adopting auxiliary unsupervised objectives for pre-training dialogue language models. Extracting language representations from pre-trained transformers for dialogue tasks has been explored in Henderson et al. (2019).

Another important line of work pertains to designing specific Transformer-based architectures that captures dialogue structures and directly pre-training these architectures on dialogue corpora. Mehri et al. (2019) proposed a Transformer-based hierarchical model and various unsupervised objectives for pre-training contextual semantics of dialogue utterances. DialogBERT differs from the methods proposed by Mehri et al. (2019) in both the architecture and objectives. DialogBERT contains a context encoder that models the discourse coherence, while Mehri et al. (2019) proposed optimizing the utterance encoder directly.

Lastly, an emerging trend in dialogue generation explores the feasibility of directly pre-training Transformer-based language modeling architectures on large-scale dialogue corpora. Recent works, such as DialoGPT (Zhang et al. 2019), Meena (Adiwardana et al. 2020), and Blender (Roller et al. 2020), demonstrate strong generation performances attainable from training Transformer-based language generators on open-domain discourses.

Multi-task Learning for Pre-training. Our work is also profoundly related to auxiliary multi-task learning, which augments the pre-training of language models. The common theme is to guide the language modeling Transformers with complementing objectives. One way is to augment language modeling with annotation rationales (Melamud, Bornea, and Barker 2019). ERNIE2 (Sun et al. 2019) is a continual multi-task learning framework for language understanding that combines unsupervised and supervised objectives. HIBERT (Zhang, Wei, and Zhou 2019) has been proposed to model documents using a hierarchical BERT architecture trained with the masked sentence decoding scheme, where the goal is to predict the entire erased sentence. Our work differs from HIBERT in that (1) we directly match the context sensitive sentence representations with the real utterance encoding and (2) we propose a novel objective that trains
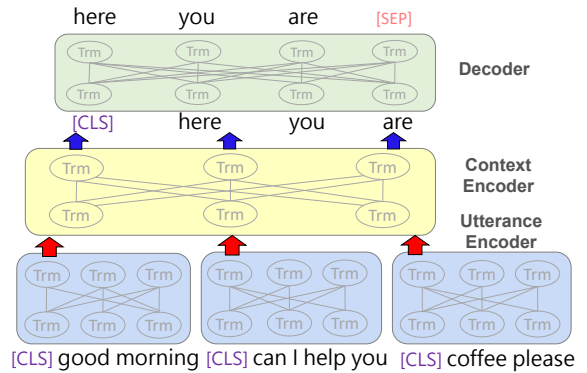

Figure 1: The hierarchical Transformer encoder-decoder architecture. We omit the [SEP] token in the end of each utterance for simplicity.

the model to predict utterance orders. ELECTRA (Clark et al. 2020) explores combining discriminative and generative objectives for learning language modeling.

Approach

Let $\mathcal{D} = (u_1, u_2, \ldots, u_T)$ denote a dialogue, where $\mathcal{C} = (u_1, u_2, \ldots, u_{T-1})$ is the dialogue context (history) and u_T is the response. Each $u_i = (w_1^i, w_2^i, \ldots, w_{|u_i|}^i)$ in $\mathcal{C}$ is an utterance and w_j^i is the j-th word in u_i. Given the context $\mathcal{C}$, our goal is to generate the next utterance (response) u_T that is coherent to $\mathcal{C}$. Consequently, there are two objectives we want to achieve: 1) learning to represent the dialogue context $\mathcal{C}$ and 2) learning the conditional probability of generating u_T given $\mathcal{C}$.

Hierarchical Transformer Encoder

To obtain a better representation of $\mathcal{C}$, we employ a hierarchical transformer encoder architecture. As shown in Figure 1, two transformer encoders are hierarchically nested: an utterance encoder $f_\theta(\cdot)$ to transform each utterance in $\mathcal{C}$ to a vector and a context encoder $g_\phi(\cdot)$ to learn utterance representations given their surrounding utterances in the context. Both encoders are based on the Transformer encoder described in (Vaswani et al. 2017).

For each $u_i = (w_1^i, w_2^i, \ldots, w_{|u_i|}^i)$, inspired by the language modeling pre-training strategy of BERT (Devlin et al. 2018), we add [CLS] and [SEP] tokens at the first and the last positions, respectively. Hence, w_1^i=[CLS] and $w_{|u_i|}^i$=[SEP]. Then, an embedding layer maps u_i onto a continuous space:

$$\mathbf{e}_i = (\mathbf{w}_1^i + \mathbf{p}_1, \mathbf{w}_2^i + \mathbf{p}_2, \ldots, \mathbf{w}_{|u_i|}^i + \mathbf{p}_{|u_i|}) \quad (1)$$

where $\mathbf{w}_j^i$ and $\mathbf{p}_j$ are the word and positional embeddings of w_j^i, respectively. Then, the utterance encoder $f_\theta(\cdot)$ transforms $\mathbf{e}_i$ into a list of hidden representations $(\mathbf{u}_1^i, \mathbf{u}_2^i, \ldots, \mathbf{u}_{|u_i|}^i)$:

$$\mathbf{u}_1^i, \ldots, \mathbf{u}_{|u_i|}^i = f_\theta(\mathbf{e}_i). \quad (2)$$

We take the first hidden representation $\mathbf{u}_1^i$ (i.e., the representation at the [CLS] token) as the representation of utterance

u_i. Similar to the representation of each word in u_i, we also take the positions of utterances into account. The final representation of u_i is $\mathbf{u}_i = \mathbf{u}_1^i + \mathbf{p}_i$.

In analogy to the utterance encoder, the context encoder $g_\phi(\cdot)$ is another Transformer encoder that is applied on the utterance level. As shown in Figure 1, after encoding the utterances, the context encoder transforms the sequence of utterance representations $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{|\mathcal{C}|})$ into context sensitive utterance representations $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{|\mathcal{C}|})$:

$$\mathbf{H} = \mathbf{h}_1, \ldots, \mathbf{h}_{|\mathcal{C}|} = g_\phi(\mathbf{u}_1, \ldots, \mathbf{u}_{|\mathcal{C}|}) \tag{3}$$

Equation 3 is the final product of encoding the dialogue context with a hierarchical bidirectional transformer encoder.

The hierarchical Transformer architecture above is reminiscent of HIBERT in document generation (Zhang, Wei, and Zhou 2019). Although both methods encode texts in a hierarchical manner, our objectives and training methods are significantly different. HIBERT is designed for document summary and is only trained through word-by-word decoding of masked sentences in documents. By contrast, DialogBERT has two new loss terms specifically designed for dialog coherence modeling, i.e., the masked utterance regression which is a discourse level regression of masked utterances and the distributed utterance order ranking that reorders randomly-shuffled utterances.

In the next section we will introduce these unsupervised objectives for training DialogBERT.

Training Objectives

In order to capture the discourse-level coherence in dialog contexts, we propose two novel objectives inspired by BERT in addition to the conventional objective of response generation, i.e., maximizing the log probabilities of decoded words. Following three subsections describe the objectives in turn.

Next Utterance Generation (NUG) As the primary goal of response generation, our first training objective is to generate the next utterance (response) given the dialog context. As is shown in Figure 1, we first apply the hierarchical encoder to the context $\mathcal{C}$ and obtain its context sensitive utterance representations $(\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|\mathcal{C}|})$. Then, we generate the next utterance $u_T = (w_1^T, \ldots, w_N^T)$ using a Transformer decoder $p_\psi(\cdot)$ (w_1^T is also a [CLS] token to be consistent to the utterance encoder). The decoder predicts each word w_j^T conditioned on $w_1^T, \ldots, w_{j-1}^T$ and $\mathbf{h}_1, \ldots, \mathbf{h}_{|\mathcal{C}|}$ by estimating the following probability distribution:

$$p(u_T|\mathcal{C}, \theta, \phi, \psi) = \Pi_{j=1}^N p_\psi(w_j^T|w_{<j}^T, \mathbf{H}), \tag{4}$$

where N represents the maximum sequence length for decoding, while θ, ϕ, and ψ denote the model parameters of the utterance encoder, the context encoder, and the decoder, respectively.

Finally, the NUG task aims to minimize the cross entropy loss in the decoder:

$$\mathcal{L}_{\text{dec}}(\theta, \phi, \psi|w_1^T, \ldots, w_N^T, \mathcal{C}) = -\sum_{i=1}^N \log p_\psi(w_i^T|w_{<i}^T, \mathbf{H})$$
$$\tag{5}$$

where N denotes the maximum sequence length of the generated response.
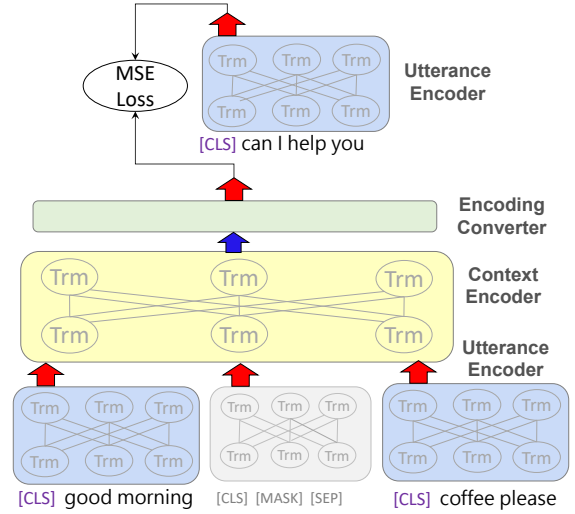

Figure 2: Illustration of masked utterance regression. The utterance "*can I help you*" is masked. We estimate its encoding using an encoding converter given its context-sensitive representation and compare the estimated encoding with the real one produced by the utterance encoder.

Masked Utterance Regression (MUR) In analogy to the masked LM (MLM) in BERT, we design the masked utterance regression as an auxiliary task for enhancing context representation learning (Figure 2). Given a dialogue context $\mathcal{C} = (u_1, u_2, \ldots, u_{T-1})$, we randomly select one utterance in $\mathcal{C}$ and the selected utterance is 1) 80% of time, replaced with a mask utterance [CLS, MASK, SEP], or 2) 10% of time unchanged in order to simulate the input context during test time (with no masked utterance), or 3) 10% of time replaced with a random utterance from the training set. We then try to reconstruct the vector of the masked utterance.

After obtaining the masked context $\tilde{\mathcal{C}} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{|\mathcal{C}|})$, we predict the original utterance vectors from $\tilde{\mathcal{C}}$. We first apply the hierarchical encoder to $\tilde{\mathcal{C}}$ and obtain its context sensitive utterance representations $(\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \ldots, \tilde{\mathbf{h}}_{|\mathcal{C}|})$. Then, we transform these representations back to the original utterance vectors using a fully connected neural network:

$$\hat{\mathbf{u}}_i = \mathbf{W}\tilde{\mathbf{h}}_i + \mathbf{b} \tag{6}$$

where $\hat{\mathbf{u}}_i$ denotes the predicted original utterance vector; $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters.

Finally, this objective aims to minimize the mean squared error (MSE) between the estimated representations of masked utterances and their original vectors:

$$\mathcal{L}_{\text{mur}}(\theta, \phi, \mathbf{W}, \mathbf{b}|\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_{|\mathcal{C}|}, \mathcal{C}, \tilde{\mathcal{C}},)$$
$$= \frac{1}{|\tilde{\mathcal{C}} \setminus \mathcal{C}|} \sum_{u_i \in \tilde{\mathcal{C}} \setminus \mathcal{C}} ||\hat{\mathbf{u}}_i - \mathbf{u}_i||_2^2 \tag{7}$$

where $\tilde{\mathcal{C}} \setminus \mathcal{C}$ denotes the set of masked utterances; θ, ϕ, $\mathbf{W}$, and $\mathbf{b}$ are training parameters.
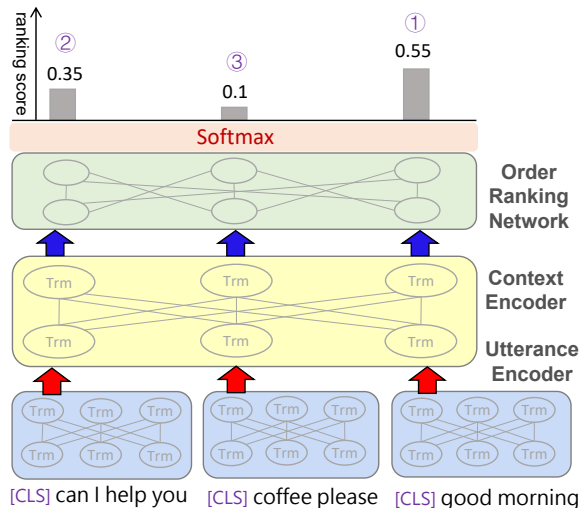
Figure 3: Illustration of distributed utterance order ranking.

Distributed Utterance Order Ranking (DUOR) Coherence is an essential aspect of conversation modeling. In a coherent discourse, utterances should respect specific orders of relations and logic. The ordering of utterances in a context determines the semantic of the conversation. Therefore, we hypothesize that learning to order a set of disordered utterances in such a way that maximizes the discourse coherence will have a critical impact in learning the representation of dialogue contexts.

The goal of the utterance re-ordering task is to organize randomly shuffled utterances of a conversation into a coherent dialogue context (Kumar et al. 2019). Formally, given a context $\mathcal{C} = [u_{o_1}, u_{o_2}, \ldots, u_{o_{|\mathcal{C}|}}]$ with order $o = [o_1, o_2, \ldots, o_{|\mathcal{C}|}]$, we want an ordered context $\mathcal{C}^* = [u_{o_1^*}, u_{o_2^*}, ..., u_{o_{|\mathcal{C}|}^*}]$ where $o^* = [o_1^*, o_2^*, \ldots, o_{|\mathcal{C}|}^*]$ is the most coherent permutation of utterances. For example, the correct order for the utterances in Figure 3 is $o^* = [3, 1, 2]$.

Previous work (Sun et al. 2019) models the sentence order prediction as a task of permutation index classification, namely, categorizing the index of the original permutation among k classes where $k = \sum_{n=1}^{|u|} n!$. However, the explosive permutation space for sets of utterances is often huge and largely overlapping (e.g., a slight modification to a permutation could lead to a far different class), leading to poor performance in current machine learning-based classifiers (Luo et al. 2019; Agrawal et al. 2013).

In this paper, we design a distributed order ranking network (DORN) on top of the context encoder. Instead of categorizing permutation indices conditioned on the overall encoding of the context, DORN predicts the order index of each utterance in a distributed manner. As Figure 3 shows, DORN takes as input the hidden status of the shuffled utterances from the context encoder and produces a score for each individual utterance. These scores are then used for re-ordering these utterances (i.e., sorting these scores provides the correct ordering in the context). Inspired by the self-attention

Dataset	Weibo	MultiWOZ	DailyDialog
conversations	5,040,000	8,438	13,118
total turns	20,521,891	113,556	102,979
train samples	15,481,891	106,794	76,052
valid samples	89,994	12,902	7,069
test samples	84,052	12,914	6,740

Table 1: Overview of the datasets

mechanism, the order prediction network computes the pairwise inner products between hidden states and then calculates the score s_i for each utterance u_i by averaging all its inner products to other utterances:

$$s_i = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \mathbf{W} \mathbf{h}_i^T \mathbf{h}_j \quad (8)$$

where $\mathbf{W}$ denotes the parameters of DORN.

In the training phase, we employ the *learning to rank* (Cao et al. 2007) framework. Specifically, we view the predicted scores as the extent to which each utterance is ranked in the first place in a context. Given these scores, we estimate the "rank-1" probability for each utterance with softmax:

$$\hat{P}(u_i) = \frac{exp(s_i)}{\sum_{j=1}^{|\mathcal{C}|} exp(s_j)} \quad (9)$$

Accordingly, we assign a gold target value to each utterance to indicate the ground truth order. Then, the "rank-1" probability estimated by the gold target values is given as:

$$P(u_i) = \frac{exp(y_i)}{\sum_{j=1}^{|\mathcal{C}|} exp(y_j)} \quad (10)$$

where $y_i \in [0, 1]$ is the gold score for utterance u_i. We set $y_i = i/|\mathcal{C}|$ in our implementation. Our goal is to minimize the KL divergence between the two distributions:

$$\mathcal{L}_{\mathrm{duor}}(\theta, \phi, \mathbf{W}|P, \mathcal{C}) = \mathrm{KL}(\hat{P}(u)||P(u))$$
$$= \sum_{k=1}^{|\mathcal{C}|} \hat{P}(u_k) \log(\frac{\hat{P}(u_k)}{P(u_k)}) \quad (11)$$

where θ, ϕ and $\mathbf{W}$ denote training parameters.

Overall, our objective is defined as the weighted sum of each loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{dec} + \lambda_0 \times \mathcal{L}_{mur} + \lambda_1 \times \mathcal{L}_{duor}, \quad (12)$$

where λ_0 and λ_1 denote the coefficients of each loss. We empirically set both λ_0 and λ_1 to 1 in our experiments.

Experimental Setup

Dataset

We evaluate DialogBERT using the following datasets (statistics shown in Table 1):

Weibo is a large-scale multi-turn conversation benchmark, introduced in the NLPCC2018 task5[1]. The dataset originates

---

[1]http://tcci.ccf.org.cn/conference/2018/dldoc/taskgline05.pdf

from the Sina Weibo (microblog), a well-known Chinese Twitter-equivalent service, a popular data source for training conversation models (Zeng et al. 2019).

MultiWOZ[2] is a dataset that contains human-to-human written conversations spanning over multiple domains and topics (Budzianowski et al. 2018). The dataset contains 10k dialogues. Although designed for task-oriented conversations, its daily-chatting style makes it suitable for evaluating open-domain conversations (Mehri et al. 2019).

DailyDialog[3] is a popular dataset for evaluating open-domain dialogue generation models (Shen et al. 2018; Gu et al. 2019). The dataset contains multi-turn daily English dialogues suitable for English learners. Compared to Weibo and Multi-WOZ, DailyDialog includes more chit-chat style utterances.

Implementation & Reproducibility

We implemented our approach on top of the Huggingface Transformer repository (Wolf et al. 2019a). We trained two different sizes of the model to accommodate to the datasets. For the Weibo dataset, we trained a base-sized model in which both encoders (including the utterance encoder and the context encoder) and the decoder use the hyper-parameter settings of '*bert-base-chinese*' (L=12, H=768, A=12). Since the base-size model is vulnerable to over-fitting in small datasets, we also trained small-sized models on MultiWOZ and DailyDialog. The small-size model reduces the '*bert-base-uncased*' configuration to 6 transformer layers, has a hidden size of 256, and contains 2 attention heads (L=6, H=256, A=2). We limit the number of utterances in each context to 7 (Adiwardana et al. 2020) and the utterance length to 30 words. All of the experiments use the default BERT tokenizer (e.g., *bert-base-uncased* for English datasets). All models were optimized with AdamW (Loshchilov and Hutter 2018) optimizer using an initial learning rate of $5e^{-5}$. We used the adaptive learning-rate scheduler with 5,000 warmup steps. We implemented all the models with the PyTorch library. Experiments took place on a machine with Ubuntu 16.04 and an NVidia Tesla P40 GPU.

We compute the validation loss every 2,000 iterations and choose the best set of parameters. The test scores from the best checkpoint represent the final results. We repeat the experiments five times and report the average results. We search for hyperparameters using NAVER Smart Machine Learning (NSML) (Sung et al. 2017; Park et al. 2019).

During response generation, we perform top-1 sampling according to the probabilities estimated by Equation 4.

Baseline Models

We compare our approach with previous Transformer-based response generation methods. (i) BART (Lewis et al. 2019) is a pre-trained Transformer encoder-decoder model for conditional text generation. BART can be seen as a generalized BERT (due to the bidirectional encoder) and GPT (with the left-to-right decoder). As such, BART is prevalent in conversational systems (Adiwardana et al. 2020). (ii) DialoGPT (Zhang et al. 2019) is a pre-training language model

for conversation generation based on GPT-2 (Radford et al. 2019). (iii) ContextPretrain (Mehri et al. 2019) refers to a set of pre-training methods for dialogue context representation learning. The paper proposes masked context retrieval and inconsistent utterance identification. To control the expressive power, we use the Transformer-based decoder in place of the RNN decoder. To ensure fairness, we used the same hyperparameter settings (e.g., the number of transformer layers and attention heads) for all baseline models.

Evaluation Metrics

Following (Adiwardana et al. 2020), we use perplexity to measure the performance of response generation. Studies (Adiwardana et al. 2020) show that this widely-used metric correlates with human judgment scores significantly. We further adopt two commonly used metrics in related works including BLEU (Papineni et al. 2002) and NIST (Doddington 2002). BLEU[4] measures how many n-grams in a generated response overlap with those of the reference. We calculate BLEU-4 scores in our experiments. NIST (Doddington 2002) is a variant of BLEU that penalizes uninformative n-grams by assigning weights to n-grams according to their information gain[5].

Evaluation Results

Automatic Evaluation

Table 2 shows the performance of each method on the automatic metrics. To demonstrate the effect of the proposed training objectives, we present the results of DialogBERT by ablating different combinations of the objectives. Broadly, DialogBERT (pre-trained with all proposed training objectives) achieves the best performance on most automatic metrics, especially for the perplexity. Compared to models with flat context encoding such as BART and DialoGPT, DialogBERT outperforms baseline models on all metrics with a large margin. Such improvements are consistent across all three datasets, affirming the superiority of the hierarchical Transformer architecture of DialogBERT.

The results show that both the proposed masked utterance regression and distributed utterance order prediction achieve substantial improvements over a simple hierarchical Transformer. Furthermore, combining both objectives further enhances the performance. Interestingly, the improvement on the Weibo dataset is relatively more significant. We conjecture that this is due to the richness of the data, allowing more room for fitting using the auxiliary objectives.

In all three datasets, DialoGPT performs considerably worse than the other methods regarding the perplexity. Presumably, this is due to the the GPT-2 language model's autoregressive nature and the single-turn setting (Zhang et al. 2019). We further discuss this issue in the case study.

To investigate the efficacy of the context encoder, we conduct an ablation study on the encoder size. We train Dialog-BRET with various numbers of Transformer layers in the

---

Model	Weibo			DailyDialog			MultiWOZ		
	PPL	BLEU	NIST	PPL	BLEU	NIST	PPL	BLEU	NIST
BART	34.15	6.5	6.6	22.92	7.59	9.78	5.28	11.83	19.69
DialoGPT	87.14	7.24	8.28	46.42	13.82	14.08	9.44	18.47	34.15
ContextPretrain-default	66.13	7.98	9.41	35.02	10.52	15.60	8.48	18.12	38.67
ContextPretrain-transformer	36.92	8.10	9.17	22.31	10.14	13.55	5.52	17.95	37.16
NUG only	31.73	8.25	9.45	22.16	14.41	22.33	5.06	19.50	49.83
NUG+MUR	28.08	8.49	10.12	21.14	13.81	21.50	4.96	19.89	52.74
NUG+DUOR	**27.29**	8.46	10.04	20.82	14.42	22.65	5.00	19.87	53.29
NUG+MUR+DUOR	27.30	**8.54**	**10.21**	**20.49**	**14.61**	**23.34**	**4.92**	**19.91**	**55.03**

Table 2: Comparison between DialogBERT (including 4 variants in the bottom 4 rows) and baseline models on three datasets. We trained a base-size model in Weibo and small-size models in DailyDialog and MultiWOZ respectively. (NUG= Hierarchical BERT with NUG loss only, MUR=Masked Utterance Regression, DUOR=Distributed Utterance Order Ranking)

Size	MultiWOZ			DailyDialog		
	PPL	BLEU	NIST	PPL	BLEU	NIST
L = 12	4.90	19.64	54.84	20.78	14.46	22.70
L = 9	4.91	19.19	54.74	20.67	14.55	23.17
L = 6	4.92	19.91	55.03	20.49	14.61	23.34
L = 3	4.94	19.19	52.48	20.85	13.89	21.57

Table 3: Performance of context encoder with difference sizes.

Model	MultiWOZ			DailyDialog		
	PPL	BLEU	NIST	PPL	BLEU	NIST
MURetr	4.98	19.69	51.25	20.05	13.98	21.20
MURegr	**4.96**	**19.89**	**52.74**	**21.14**	**13.81**	**21.50**
UOC	5.15	18.68	48.07	23.85	10.58	14.90
DUOR	**5.00**	**19.89**	**53.29**	**20.82**	**14.42**	**22.65**

Table 4: Efficacy of our objectives compared to conventional ones (MURetr = Masked Utterance Retrieval, MURegr = Masked Utterance Regression, UOC= Utterance Order Classification, DUOR=Distributed Utterance Order Ranking). Bold fonts represent our proposed objectives.

context encoder. As shown in Table 3, the number of Transformer layers in the context encoder has a modest effect on the performance, which indicates the ease of extracting context representations. We conjecture that this may be because the discourse-level interaction is relatively easier to capture than the word-level semantics.

We also investigated the effects of two proposed training objectives, which we compare with previously-proposed objectives for paragraphs, namely the masked utterance retrieval (Mehri et al. 2019) and the utterance order classification (Sun et al. 2019). As Table 4 shows, the proposed MURegr performs competitively with the *masked utterance retrieval* objective, supporting the idea that the MURegr is a simple yet effective substitute for dialog context representation. One explanation is that MURegr offers more fine-grained representation learning than retrieval-based objectives (negative utterance classification) thanks to regression

(exact vector matching). Thus, for missing utterance reconstruction, the goal of exact matching (regression) becomes more impactful than the goal of classification. Meanwhile, the proposed DUOR objective performs significantly better than the existing *utterance order classification* objective.

We proceed to analyze the computational efficiency. Table 5 shows the model size, training and test time for each model. Although DialogBERT has a larger model size due to an extra "context encoder", it does not incur significant overhead since the length (the number of utterances in each dialogue) of the context encoder input is usually short.

Human Evaluation

We conduct a human evaluation using the Amazon Mechanical Turk platform. We used the DailyDialog as the evaluation corpus since its properties (daily chit-chats) make it easier for annotators to judge. We randomly sampled 200 dialogues to be judged by the workers. For each, we presented (i) the entire dialogue context, (ii) the response generated by our model, and (iii) the response generated by the competing models without disclosing the source models. For each pair of sample responses, we asked three different annotators to blindly evaluate the quality regarding the three criteria (coherence, informativeness and human-likeness) and express their preference using a 3-point Likert scale: "win" (ours is better), "loss" (the other is better) and "tie" (equally good or bad). We screened the workers by qualifications, and we also filtered out low-quality answers using spam detection methods. Table 6 shows the overall distribution of the answers. Overall, the workers preferred our model over other baselines regarding the relevance to the dialogue context, informativeness, and human-likeness.

Case Study

We provide two generated dialogues in Table 7 from the MultiWOZ and DailyDialog datasets. The two samples illustrate that DialogBERT generates more coherent responses than the baseline models, consistent with our automatic and human evaluation results. Interestingly, DialogBERT exhibits *meticulousness* that makes its response more human-like. For example, the first sample shows that DialogBERT produces

Model	Weibo			DailyDialog			MultiWOZ		
	Size	**Train**	**Test**	**Size**	**Train**	**Test**	**Size**	**Train**	**Test**
BART	139.4M	58h	11h	24.2M	2.8h	0.9h	24.2M	6.6h	1.8h
DialoGPT	102.1M	27.5h	9.8h	12.7M	2.6h	0.85h	12.7M	8.9h	1.8h
ContextPretrain	20.5M	14.6h	21h	23.3M	16.8h	1.4h	23.3M	17.9h	3.2h
DialogBERT	337.6M	27.9h	23.3h	40.2M	9.0h	1.3h	40.2M	9.8h	1.9h

Table 5: Comparison of model size (i.e., number of parameters), training and inference time between DialogBERT and baseline models in terms of three datasets (M: million, h: hours).

Comparison	Coherence			Informativeness			Human-likeness		
	Win	**Tie**	**Loss**	**Win**	**Tie**	**Loss**	**Win**	**Tie**	**Loss**
Ours *vs.* BART	41%	34%	25%	38%	43%	19%	39%	45%	15%
Ours *vs.* DialoGPT	39%	45%	16%	40%	48%	12%	49%	43%	8%
Ours *vs.* ContextPretrain	44%	34%	22%	40%	42%	18%	39%	42%	19%

Table 6: Human evaluation on the test set of DailyDialog. The NUG+MUR+DUOP in Table 2 is selected for comparison. All results have statistical significance of $p < 0.001$.

Context	A: I'm looking for a college in town. B: Is there a particular area you're looking for a college in? christ's college is in the centre area, it is located on saint andrew. A: Could you give me the address for it?
BART:	Sure.
DialoGPT:	Yes will, give thecode thecode. post is21,21 and entrance is. address saint'street pose cbbu and admission free
ContextPretrain:	Yes, it is moderately priced and do you want free parking and wifi?
DialogBERT:	**Their address is 98 king street and their post code is cb11ln. Can I help you with anything else?**
Human:	The address is saint andrew's street and the postcode is cb23bu. Is there anything else I can help you with?
Context	A: Then how much is that by airmail? B: Let me see. it's 57 dollars, 20 cents, including tax. A: I guess I have to send it by airmail. B: Do you want to ensure the contents, miss? A: Yes, please.
BART:	I would your e mail.
DialoGPT:	Thank. only one much.
ContextPretrain:	Do you need any discounts?
DialogBERT:	**Would you please write the number and tell me the best?**
Human:	Please fill out this form, also please write the value of the items in this space.

Table 7: Sample conversations from multiple models with human reference

very detailed responses such as the "*98 king street*" and the "*post code cb1ln*" instead of vague ones observed in samples from other models. This observation is in agreement with our human study on informativeness and human-likeness.

Furthermore, in both examples, DialoGPT produces less relevant results than other methods. Again, this is related to the single-turn setting (Zhang et al. 2019). Specifically, DialoGPT treats response generation as a pure auto-regressive language model. The entire context is taken as input to generate subsequent words step by step, making it difficult to handle multiple turns, especially in relatively small conversation datasets.

Conclusion

In this paper, we proposed a neural response generation model named DialogBERT. Instead of encoding the dialogue context as a linear sequence of tokens, DialogBERT employs a hierarchical Transformer encoder architecture. As a natural extension of the original BERT training, we proposed two training objectives: masked utterance regression and distributed utterance re-ordering. We showed that the proposed objectives enable the conversation model to capture multi-level (discourse-level and utterance-level) coherences. Additionally, we showed that DialogBERT notably outperforms baseline models on the response generation tasks.

Acknowledgments

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* .

Agrawal, R.; Gupta, A.; Prabhu, Y.; and Varma, M. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, 13–24.

Budzianowski, P.; and Vulić, I. 2019. Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 15–22.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278* .

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, 129–136.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations (ICLR 2020)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145.

Gu, X.; Cho, K.; Ha, J.-W.; and Kim, S. 2019. DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.

Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.-H.; Vulić, I.; et al. 2019. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *arXiv preprint arXiv:1911.03688* .

Kumar, P.; Brahma, D.; Karnick, H.; and Rai, P. 2019. Deep Attentive Ranking Networks for Learning to Order Sentences. *arXiv preprint arXiv:2001.00056* .

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* .

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam. *open review for ICLR 2018* .

Luo, Y.; Wong, Y.; Kankanhalli, M.; and Zhao, Q. 2019. G-Softmax: Improving Intraclass Compactness and Interclass Separability of Features. *IEEE transactions on neural networks and learning systems* .

Mehri, S.; Razumovskaia, E.; Zhao, T.; and Eskenazi, M. 2019. Pretraining Methods for Dialog Context Representation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3836–3845.

Melamud, O.; Bornea, M.; and Barker, K. 2019. Combining Unsupervised Pre-training and Annotator Rationales to Improve Low-shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3875–3884.

Ortega, D.; and Vu, N. T. 2017. Neural-based context representation learning for dialog act classification. *arXiv preprint arXiv:1708.02561* .

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Park, H.; Kim, J.; Kim, M.; Kim, J.-H.; Choo, J.; Ha, J.-W.; and Sung, N. 2019. VisualHyperTuner: Visual analytics for user-driven hyperparameter tuning of deep neural networks. In *Demo at SysML Conference*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

Pragst, L.; Rach, N.; Minker, W.; and Ultes, S. 2018. On the vector representation of utterances in dialogue context. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *Open AI Blog* 1: 9.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* .

Rothe, S.; Narayan, S.; and Severyn, A. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv:1907.12461* .

Shen, X.; Su, H.; Niu, S.; and Demberg, V. 2018. Improving variational encoder-decoders in dialogue generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

So, D. R.; Liang, C.; and Le, Q. V. 2019. The evolved transformer. *arXiv preprint arXiv:1901.11117* .

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2019. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412* .

Sung, N.; Kim, M.; Jo, H.; Yang, Y.; Kim, J.; Lausen, L.; Kim, Y.; Lee, G.; Kwak, D.; Ha, J.-W.; et al. 2017. Nsml: A machine learning platform that enables you to focus on your models. *arXiv preprint arXiv:1712.05902* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019a. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149* .

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5754–5764.

Zeng, J.; Li, J.; He, Y.; Gao, C.; Lyu, M. R.; and King, I. 2019. What you say and how you say it: Joint modeling of topics and discourse in microblog conversations. *Transactions of the Association for Computational Linguistics* 7: 267–281.

Zhang, X.; Wei, F.; and Zhou, M. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566* .

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv preprint arXiv:1911.00536* .

Zheng, Y.; Zhang, R.; Huang, M.; and Mao, X. 2020. A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data. In *AAAI*, 9693–9700.