DIALOG-POST: Multi-Level Self-Supervised Objectives and Hierarchical Model for Dialogue Post-Training

Zhenyu Zhang, Lei Shen, Yuming Zhao, Meng Chen,[*] Xiaodong He
JD AI Research, Beijing, China
{zhangzhenyu47,shenlei20,zhaoyuming3}@jd.com
{chenmeng20,xiaodong.he}@jd.com

Abstract

Dialogue representation and understanding aim to convert conversational inputs into embeddings and fulfill discriminative tasks. Compared with free-form text, dialogue has two important characteristics, hierarchical semantic structure and multi-facet attributes. Therefore, directly applying the pretrained language models (PLMs) might result in unsatisfactory performance. Recently, several work focused on the dialogue-adaptive post-training (DialPost) that further trains PLMs to fit dialogues. To model dialogues more comprehensively, we propose a DialPost method, DIALOG-POST, with multi-level self-supervised objectives and a hierarchical model. These objectives leverage dialogue-specific attributes and use self-supervised signals to fully facilitate the representation and understanding of dialogues. The novel model is a hierarchical segment-wise self-attention network, which contains inner-segment and inter-segment self-attention sublayers followed by an aggregation and updating module. To evaluate the effectiveness of our methods, we first apply two public datasets for the verification of representation ability. Then we conduct experiments on a newly-labelled dataset that is annotated with 4 dialogue understanding tasks. Experimental results show that our method outperforms existing SOTA models and achieves a 3.3% improvement on average.

1 Introduction

As an indispensable way of communication, dialogue is related to many research and application scenarios in academia and industry. Better dialogue representation and understanding serve for several tasks, including intent classification, emotion recognition, and response selection, thus how to represent and model dialogues is an essential topic. Compared with free-form text, dialogue modeling has to pay more attention to the following characteristics: (1) hierarchical semantic structure (Serban et al., 2016; Xing et al., 2018; Zhang et al., 2019), i.e., dialogue → utterance → token, and (2) multi-facet attributes (See et al., 2019; Shen et al., 2021a), such as speaker-shift, content-relatedness, fact-awareness, and coherence. Therefore, directly applying pre-trained language models (PLMs) to the dialogue understanding tasks is inappropriate.

To better utilize PLMs for dialogue representation and understanding, researchers use data samples from dialogue corpora to conduct a second phase pre-training of PLMs, i.e, dialogue-adaptive post-training (DialPost). At first, the training objectives were just those for general language modeling (Masked Language Modeling and Next Sentence Prediction) (Whang et al., 2020, 2021; Xu et al., 2021). After that, researchers tried to design some novel objectives that fit dialogue characteristics more. For example, Wu et al. (2021) utilized Span Boundary Objective and Perturbation Masking Objective in post-training to capture the dialogue semantics in span and token levels. Liu et al. (2021) and Wu et al. (2020) constructed positive and negative samples for context-response pairs, and continued training PLMs with contrastive learning to better maintain the dialogue coherence.

Existing DialPost methods either focus on token-level or utterance-level semantics, which only consider a limited subset of dialogue attributes, e.g., speaker-shift (Xu and Zhao, 2021), coherence (Li et al., 2020a), and response-similarity (Wu et al., 2020). However, the comprehensive modeling of multi-facet attributes with multi-level training objectives is not well explored. Moreover, previous DialPost methods handle the whole dialogue as a linear sequence of successive tokens and feed it to PLMs that obtain the token representations indiscriminately with flat self-attention mechanisms. Such a way of modeling is sub-optimal to capture the hierarchical semantic relations of dialogues (Zhang and Zhao, 2021).

[*]Corresponding author.

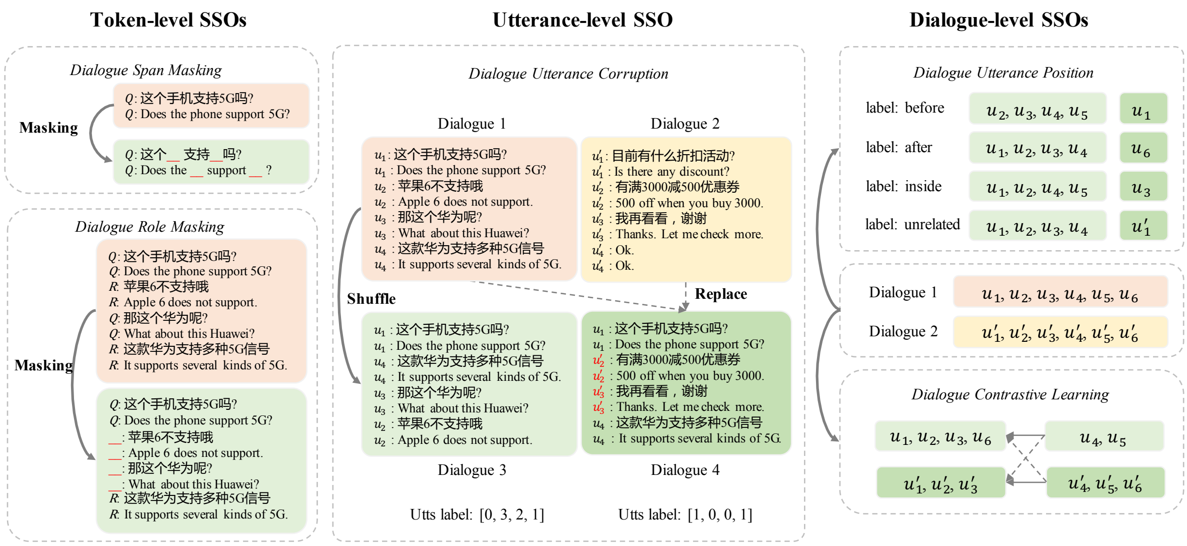

Figure 1: Illustration of multi-level SSOs in DIALOG-POST. Q and R represent speaker roles. u_i represents utterance. The utterance/dialogue in green color represents the corrupted utterance/dialogue.

To tackle the above issues, we propose a post-training method for dialogues, namely DIALOG-POST, which consists of five **S**elf-**S**upervised **O**bjectives (SSOs) and a hierarchical model. The former is designed to capture the multi-facet attributes of dialogues, while the latter is used to model the hierarchical relations in dialogues. Specifically, SSOs correspond to two token-level, one utterance-level, and two dialogue-level self-supervised learning tasks. For the token-level objectives, we use different sampling approaches to mask spans and roles, which capture fact-awareness and speaker-shift, respectively. For the utterance-level objective, we corrupt a dialogue via two operations on utterances, and then train the model to maintain coherence by either detecting the corrupted utterances or recovering the utterance order. For the dialogue-level objectives, we model the content-relatedness of both utterance-context pairs and context-context pairs by utterance position prediction and dialogue-based contrastive learning. The model is a **H**ierarchical **S**egment-wise **S**elf-**A**ttention network (HSSA) that contains inner-segment and inter-segment self-attention layers along with an aggregation and updating module.

To evaluate the proposed method, we conduct experiments in two aspects, i.e., dialogue representation and understanding. We first verify the representation ability of DIALOG-POST with dialogue-based semantic textual similarity (D-STS) and semantic retrieval (SR) tasks on two public datasets, JDDC and ECD. DIALOG-POST outperforms baselines by 1.7% in D-STS task of JDDC. Then, we annotate a dataset with four dialogue understanding tasks and conduct experiments on them. Experimental results show that our method consistently outperforms baselines and achieves a 87.5% average score (+3.3%) for dialogue understanding.

Our contribution can be summarized as follows: (1) We propose a post-training method (DIALOG-POST) for dialogue representation and understanding, which consists of five multi-level SSOs and a hierarchical model. (2) We conduct extensive experiments to evaluate DIALOG-POST with two public and one newly-labelled dataset. (3) We analyse the effectiveness of each component of DIALOG-POST, and conduct ablation study to demonstrate the necessity of objectives in different levels.

2 Approach

In this section, we introduce the multi-level self-supervised objectives (SSOs) and HSSA model.

2.1 Multi-Level SSOs

As illustrated in Figure 1, we design five multi-level SSOs to post-train the dialogue encoder, which consist of two token-level SSOs ($\mathcal{L}_{DSM}$ and $\mathcal{L}_{DRM}$), one utterance-level SSO ($\mathcal{L}_{DUC}$), and two dialogue-level SSOs ($\mathcal{L}_{DUP}$ and $\mathcal{L}_{DCL}$).

Token-Level SSOs. A good conversation should avoid presenting contradictory contents about facts (Zhang and Zhao, 2021). Therefore, the ability of realizing important words and phrases, denoted as fact-awareness, is a fundamental attribute and helps keep the factual consistency. Here, we design a Dialogue Span Masking (**DSM**) objective,

$\mathcal{L}_{DSM}$, to capture the fact-awareness. First, we sample 50% utterances from a dialogue. Then, we perform the span masking (Joshi et al., 2020) for each selected utterance, and the model needs to recover those masked spans. By this means, the facts in each utterance and their dependency within or cross utterances can be learned.

Speaker-shift is a distinctive attribute of dialogues (Gu et al., 2020). In a real scenario, two speakers carry out a conversation in an interactive way, and one speaker may continuously shoot multiple utterances (Xu and Zhao, 2021). We propose the Dialogue Role Masking (**DRM**) objective , $\mathcal{L}_{DRM}$, which aims to predict the masked role tokens. Before that, 80% of role tokens are randomly masked in a dialogue. In Figure 1, the masked tokens and speaker roles are marked with "_".

Utterance-Level SSO. An utterance is the most basic semantic unit in dialogues (Jiao et al., 2019; Zhu et al., 2020; Li et al., 2020b; Henderson et al., 2020), and utterance corruptions could break the entire coherence. To better maintain the coherence by mimicking possible corruptions, we propose a Dialogue Utterance Corruption (**DUC**) objective, $\mathcal{L}_{DUC}$. Given a dialogue $\mathcal{D}$ containing m utterances, i.e., $\mathcal{D} = \{u_1, u_2, ..., u_m\}$, $n_c = \lceil 0.3 * m \rceil$, we could corrupt a dialogue via 2 operations:

- Replace: We sample n_c utterances from other dialogues $\mathcal{D}^{'}$ with each utterance $u_j^{'} \in \mathcal{D}^{'}$, $j \in [1, n_c]$. Then, we replace n_c randomly selected utterances in $\mathcal{D}$ with the sampled ones, and assign each utterance a label $\mathcal{Y} = \{y_1, y_2, ..., y_m\}$, where y_t is 0 for the replaced utterance; otherwise 1 for the original utterance. The goal is to predict 0 or 1 for y_t.

- Shuffle: We sample n_c utterances from $\mathcal{D}$ and then shuffle them to change their order. The goal is to predict orders of the n_c shuffled utterance, and the size of label set $\mathcal{Y}$ equals to n_c with each $y_t \in [1, n_c]$.

In practice, we randomly apply one operation, and use different classification heads to predict $\mathcal{Y}$ for either "Replace" or "Shuffle". Two examples are given in Figure 1 for better understanding.

Dialogue-Level SSOs. A conversation usually contains topic changes and redundant messages regarding a utterance or partial context. Therefore, we need to detect relevant information via exploring the relationship of utterances and contexts. Previous works mainly focus on the response selection task (Liu et al., 2021; Wu et al., 2020) that measures the similarity of each context-response pair. To consider utterances in different positions, not only the last one (i.e., response), we model the content-relatedness of utterance-context pairs, and propose a Dialogue Utterance Position (**DUP**) objective, $\mathcal{L}_{DUP}$. We first regard an utterance as *query*, and a list of consecutive utterances as *context*, then their relationship can be defined as follows: (1) Before: query u_b is before the context $\{u_k, u_{k+1}, ..., u_m\}$, i.e., $1{\le}b{<}k$; (2) After: query u_a is after the context $\{u_1, u_2, ..., u_j\}$, i.e., $j{<}a{\le}m$; (3) Inside: query u_i is inside the context $\{u_1, ..., u_{i-1}, u_{i+1}, ..., u_m\}$, i.e., $1{<}i{<}$m; (4) Unrelated: the context is $\{u_1, u_2, ..., u_m\}$, while query $u^{'}$ is sampled from another dialogue. Finally, we feed the *context* and *query* into a dialogue encoder under the sequence-pair classification setting (Devlin et al., 2019).

In addition, we extend an utterance to consecutive utterances, and capture the content-relatedness of context-context pairs with a Dialogue Contrastive Learning (**DCL**) objective, $\mathcal{L}_{DCL}$. Specifically, we randomly sample $n_c = \lceil 0.3 * m \rceil$ consecutive utterances $\mathcal{D}_p = \{u_1, u_2, ..., u_{n_c}\}$ from a dialogue $\mathcal{D}$. Then we replace each utterance in $\mathcal{D}_p$ with a special token "[UMASK]", and construct an incomplete dialogue $\mathcal{D}_r$=$\mathcal{D}$/$\mathcal{D}_p$. Given a batch of $\mathcal{D}_r$-$\mathcal{D}_p$ pairs, we apply the in-batch contrastive learning loss (Wang and Isola, 2020; Gao et al., 2021) to compute $\mathcal{L}_{DCL}$:

$$\mathcal{L}_{DCL} = -\frac{1}{N}\sum_{i}^{N}\log\frac{e^{\mathrm{sim}(f(\mathcal{D}_{r_i}), f(\mathcal{D}_{p_i}))/\tau}}{\sum_{j\neq i} e^{\mathrm{sim}(f(\mathcal{D}_{r_i}), f(\mathcal{D}_{p_j}))/\tau}}.$$

For a given $\mathcal{D}_{r_i}$, we calculate the cosine similarity with the corresponding $\mathcal{D}_{p_i}$ against the other partial context $\mathcal{D}_{p_j}$. We use the average output of the encoder $f(\cdot)$ and set temperature τ to 0.1.

Continuous Multi-Task Learning. Inspired by Sun et al. (2020), we apply the popular continuous multi-task learning (CMTL) framework for model training. CMTL can pre-train models with multi-task objectives efficiently and prevent knowledge forgetting of previous tasks when training with the current task objective(s). Since our method consists of several tasks, CMTL is extremely proper for our experiments. The final objective is calculated as:

$$\mathcal{L} = \mathcal{L}_{DSM} + \mathcal{L}_{DRM} + \mathcal{L}_{DUC} + \mathcal{L}_{DUP} + \mathcal{L}_{DCL}.$$

Table 1 illustrates the details of training process. For each stage (denoted as S_i), we train the model with multiple tasks and each task with used for given steps, i.e., for S_2, we train the model using DRM for 5K steps and DSM for 30K steps.

SSO	S_1	S_2	S_3	S_4	S_5
DRM	20K	5K	5K	5K	5K
DSM	0	30K	10K	5K	5K
DUC	0	0	40K	5K	5K
DUP	0	0	0	40K	10K
DCL	0	0	0	0	50K

Table 1: The illustration of CMTL.

2.2 HSSA Model

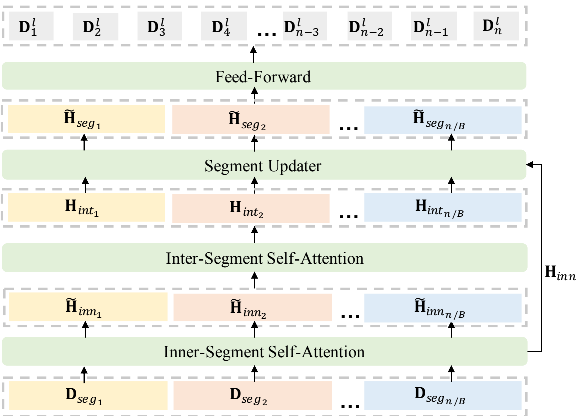

Figure 2: Overview of a HSSA layer.

As shown in Figure 2, the proposed hierarchical segment-wise self-attention (HSSA) model contains several layers, and each layer is a block consisting of inner-segment self-attention, inter-segment self-attention, segment updater, and feed-forward sub-layers.

For the l-th layer, we split the dialogue hidden states $\mathbf{D}^{l-1} \in \mathbb{R}^{n \times d}$ from the previous layer into $\frac{n}{B}$ segments, where n is length of input sequence, and each segment $\mathbf{D}_{seg_i}$ contains B hidden states. For each $\mathbf{D}_{seg_i}$, we first apply the self-attention mechanism $\mathrm{SA}(\cdot)$ (Vaswani et al., 2017) to obtain an inner-segment representation $\mathbf{H}_{inn_i} = \mathrm{SA}(\mathbf{D}_{seg_i}) \in \mathbb{R}^{B \times d}$. Then, we aggregate $\mathbf{H}_{inn_i}$, and compute the attention scores between the aggregated state and each segment state:

$$\mathrm{Agg}(\mathbf{H}_{inn_i}) = \frac{1}{\sum e^{\mathbf{M}_j}} \sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * e^{\mathbf{M}_j},$$

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{\mathrm{Agg}(\mathbf{H}_{inn_i})\mathbf{H}_{inn_{i,j}}^T}{\sqrt{d}}\right), j \in [1, B],$$

where $\mathbf{M}_j$ is the attention mask that is $-\inf$ for non-attended tokens and 0 for the rest.

To obtain the sub-layer output $\tilde{\mathbf{H}}_{inn}$, we use an attention-based pooling method:

$$\tilde{\mathbf{H}}_{inn_i} = \mathbf{W}_p\left(\sum_{j=1}^{B} \mathbf{H}_{inn_{i,j}} * \alpha_{ij}\right)^T + \mathbf{b}_p,$$

$$\tilde{\mathbf{H}}_{inn} = [\tilde{\mathbf{H}}_{inn_1}, \tilde{\mathbf{H}}_{inn_2}, ..., \tilde{\mathbf{H}}_{inn_{n/B}}],$$

where $\mathbf{W}_p \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_p \in \mathbb{R}^d$ are parameters of linear transformation.

We then apply the self-attention mechanism to get the inter-segment hidden states $\mathbf{H}_{int} = \mathrm{SA}(\tilde{\mathbf{H}}_{inn})$. The inner-segment and inter-segment self-attention share the same set of parameters. Next, we use an updater to update B hidden states in each segment with the corresponding inter-segment representation $\mathbf{H}_{int_i} \in \mathbb{R}^{1 \times d}$:

$$\tilde{\mathbf{H}}_{seg_{i,j}} = \beta_{i,j} * \mathbf{H}_{int_i} + \mathbf{H}_{inn_{i,j}},$$

$$\beta_{i,j} = \mathrm{softmax}\left(\frac{\mathbf{H}_{inn_{i,j}}\mathbf{H}_{int_i}^T}{\sqrt{d}}\right), j \in [1, B].$$

The segment representations are concatenated and fed to the feed-forward layer to get the output $\{\mathbf{D}_1^l, \mathbf{D}_2^l, ..., \mathbf{D}_n^l\}$. We then apply a residual connection to the output and $\mathbf{D}^{l-1}$ with layer normalization. Note that HSSA model does not include extra parameters, thus we can fully initialize the model with pretrained language models, such as BERT. Moreover, the segment-based attention reduces the computational burden. HSSA can reduce the memory cost from $O(n^2)$ to $O(nB + (\frac{n}{B})^2 + n)$, which also improves the training and inference efficiency.

3 Experiments

To verify the effectiveness of DIALOG-POST, we conduct extensive experiments on both dialogue representation and understanding tasks. We first introduce the experimental setup, then elaborate implementation details, and finally illustrate the main experimental results.

3.1 Experimental Setup

Post-Training Data. For fair comparison (Xu and Zhao, 2021; Zhang and Zhao, 2021; Xu et al., 2021), we utilize two public dialogue datasets, JDDC (Chen et al., 2020) and ECD (Zhang et al., 2018), to conduct all experiments. JDDC[1] is a large-scale multi-turn dialogue corpus released by

[1]Dataset is available at http://jddc.jd.com.

Method	JDDC			ECD		
	Corr.	MAP	MRR	Corr.	MAP	MRR
BERT (Devlin et al., 2019)	72.60	53.03	66.99	74.26	59.32	76.89
ELECTRA (Clark et al., 2020)	71.05	52.21	66.30	73.07	56.07	76.14
ERNIE (Sun et al., 2019, 2020)	72.73	52.96	66.79	74.29	59.11	76.87
UMS (Whang et al., 2021)	74.69	56.39	70.33	75.23	60.99	78.06
TOD-BERT (Wu et al., 2020)	78.43	60.15	74.32	80.17	65.78	80.22
PLATO (Bao et al., 2020b, 2021)	73.48	53.86	68.00	74.65	60.52	77.16
DialBERT (Zhang et al., 2021)	76.55	58.83	72.09	78.65	62.23	78.64
DomainAP (Wu et al., 2021)	76.54	59.27	72.36	78.99	62.85	79.08
DialCSE (Liu et al., 2021)	81.22	68.02	79.52	83.94	69.32	81.20
DIALOG-POST-BERT	82.78	69.91	79.83	**83.96**	**71.78**	**81.78**
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	81.72

Table 2: Evaluation results on semantic retrieval (SR) and dialogue-based semantic textual similarity (D-STS) tasks.

JD[2], which contains more than 1 million real conversations between users and customer service staff in E-commerce scenario. ECD is a large-scale dialogue corpus collected from Taobao[3]. Finally, 2,044,196 dialogues with 27,951,337 utterances in total are used for post-training.

Evaluation Tasks. Two typical groups of evaluation are considered to verify the effectiveness of DIALOG-POST. The first group is evaluation on dialogue representation, which uses utterance embeddings obtained by the dialogue encoder to fulfill two tasks, the semantic retrieval (**SR**) and the dialogue-based semantic textual similarity (**D-STS**) (Liu et al., 2021). The SR task is a retrieval task that ranks utterance candidates by calculating the semantic similarity between embeddings of a query utterance and those candidates. The D-STS task aims to classify each utterance pair into five degrees ranging from 1 to 5 according to their semantic relevance. We utilize the public evaluation sets of JDDC and ECD release by Liu et al. (2021).

The second group of evaluation consists of four popular downstream tasks of dialogue understanding, which are Intent Classification (**IC**), Sentiment Recognition (**Senti**), Context-Question Matching (**CtxQ**), and Context-Response Matching (**CtxR**). CtxQ and CtxR are two critical tasks for retrieval-based dialogue systems, and we formulate them as binary classification problem here. The downstream understanding tasks usually rely on the domain of dialogue corpus. To avoid domain inconsistency, we construct four datasets for the above tasks by re-annotating the data sampled from JDDC. Please refer to Appendix A for more details of the annotation.

Task	Class	Metric	Train	Test
J/D-STS	-	Corr.	-	2,000
J/SR	-	MAP/MRR	-	6,970
E/D-STS	-	Corr.	-	1,000
E/SR	-	MAP/MRR	-	4,243
IC	30	F1	4.7K	988
Senti	7	ACC	2.7K	342
CtxQ	2	AUC	4.1K	620
CtxR	2	AUC	4K	593

Table 3: Details of evaluation tasks. "J" and "E" represent JDDC and ECD.

Evaluation Metrics. Following Liu et al. (2021), we report the mean average precision (MAP) and mean reciprocal rank (MRR) scores for SR, and the Spearman's Correlation (denoted as Corr.) score for D-STS. For four understanding tasks, we calculate Macro-F1 (denoted as F1) for IC, Accuracy (denoted as ACC) for Senti, and AUC (Area under the ROC plot) for CtxQ and CtxR. To avoid the impact of randomness in neural networks, we report the evaluation results of 5 runs in the format "avg$\pm$std.dev". Details of each evaluation task are illustrated in Table 3.

Baselines. We choose two branches of models as our baselines. The first branch is PLMs post-trained with dialogue data via original objectives, including: (1) **BERT** (Devlin et al., 2019), which utilizes Masked Language Modeling and Next Sentence Prediction (NSP) objectives for pre-training. (2) **ERNIE** (Sun et al., 2019, 2020), which leverages external knowledge base to mask entities and phrases. (3) **ELECTRA** (Clark et al., 2020), which devises the replaced token detection task to pre-

[2] http://www.jd.com.
[3] http://www.taobao.com.

Method	IC	Senti	CtxQ	CtxR	Average
BERT (Devlin et al., 2019)	86.0±0.3	71.9±1.8	87.9±1.1	80.0±0.9	81.5
ELECTRA (Clark et al., 2020)	87.4±0.5	72.5±0.6	88.9±0.5	81.7±1.5	82.6
ERNIE (Sun et al., 2019, 2020)	87.2±0.3	73.4±1.0	89.2±1.2	82.9±0.4	83.2
UMS (Whang et al., 2021)	86.8±0.3	71.2±1.0	88.8±0.8	84.0±0.1	82.7
TOD-BERT (Wu et al., 2020)	87.4±0.9	74.8±1.2	87.8±0.7	82.8±0.5	83.2
PLATO (Bao et al., 2020b, 2021)	86.5±0.4	73.1±0.1	88.9±0.4	82.2±0.4	82.7
DialBERT (Zhang et al., 2021)	88.5±0.4	73.5±0.5	87.5±0.4	81.9±0.5	82.8
DomainAP (Wu et al., 2021)	87.9±0.4	73.8±0.5	89.1±0.4	83.7±0.2	83.6
DialCSE (Liu et al., 2021)	86.8±0.3	73.6±0.5	90.7±0.8	85.6±0.2	84.2
DIALOG-POST-BERT	91.3±0.7	**78.3**±0.9	92.0±0.6	87.3±0.8	87.2
DIALOG-POST	**91.8**±0.5	78.1±0.5	**92.4**±0.7	**87.9**±0.5	**87.5**

Table 4: Evaluation results on dialogue understanding tasks (all with significance value $p < 0.05$).

train the language model as a discriminator.

The second branch is the dialogue-adaptive post-training models, including: (4) **UMS** (Whang et al., 2021), which proposes three utterance manipulation strategies for dialogues to promote response selection and context understanding. (5) **PLATO** (Bao et al., 2020b, 2021), which utilizes UniLM (Dong et al., 2019) to pre-train dialogue encoder with a discrete latent variable via act recognition and response generation tasks. (6) **TOD-BERT** (Wu et al., 2020), which combines the contrastive learning loss and MLM to train the dialogue encoder. (7) **DialCSE** (Liu et al., 2021), which designs the matching-guided embedding and turn aggregation with contrastive learning to obtain the context-aware utterance representation. (8) **DialBERT** (Zhang et al., 2021), which proposes several dialogue-specific self-supervised tasks to train a dialogue encoder. (9) **DomainAP** (Wu et al., 2021), which combines the pre-training objectives of SpanBERT (Joshi et al., 2020) and perturbation masking objective to enhance the model performance in downstream dialogue tasks.

All above baselines are back-boned with BERT[4] (Devlin et al., 2019). For fair comparison, we post-train all models with the same training data as mentioned before. For SR and D-STS tasks, we infer the utterance embeddings by feeding utterances into the model without fine-tuning. For IC, Senti, CtxQ and CtxR tasks, we fine-tune all models[5] with the corresponding datasets, then conduct the performance evaluation.

[4]We choose the whole word masking Chinese BERT (Cui et al., 2020) as the base model.

[5]Our code and dataset can be found from: https://github.com/zhangzhenyu13/dialogue-post

3.2 Implementation Details

Hyper-parameters of HSSA. Previous research (Zhong et al., 2021) indicates that segment-based attention and full self-attention are complementary on catching the local and global dialogue semantics. Inspired by this, we take the hybrid manner for HSSA implementation, i.e., the first 10 layers are the HSSA blocks with segment size of 8, 16, 32, 32, 64, 64, 64, 128, 128, 128, while the last 2 layers are the original Transformer blocks. We use the self-attention layer weights from a Chinese BERT with whole word masking (Cui et al., 2020) to initialize both the inner-segment and inter-segment self-attention sub-layers in HSSA. The input embedding layer is the same as that of BERT. Therefore, HSSA has no extra parameters compared to BERT. Moreover, we also post-train a BERT model (denoted as DIALOG-POST-BERT) with the multi-level SSOs. Unless otherwise specified, the model base of DIALOG-POST in our work is HSSA.

3.3 Experimental Results

Evaluation on Dialogue Representation. As a novel method for dialogue representation, we first verify the performance of our model on SR and D-STS tasks. Previous research (Liu et al., 2021) shows that using the average of all token embeddings is better than using the "[CLS]" token embedding for utterance representation, thus we utilize the average token embedding in our experiments. The results in Table 2 shows that: (1) All models post-trained with dialogue-adaptive methods surpass the general-purpose PLMs by a large margin, which indicates the advantages of various self-supervised training objectives to catch the dialogue characteristics during representation learning. (2) Among the baselines, DialCSE (Liu et al., 2021) has the best performance, we argue that the advantage mainly comes from the context-aware response-based contrastive learning, which benefits the semantic matching tasks naturally by eliminating the gap between training and evaluation. (3) Our proposed method DIALOG-POST beats all baselines on both datasets, demonstrating the superiority of multi-level SSOs during post-training, which can generate better representations for dialogue utterances by catching the multi-facet attributes of dialogues. For SR on ECD, the performance of DIALOG-POST-BERT is slightly better than DIALOG-POST. We conjecture that it is because the ECD corpus has shorter dialogue contexts, which may limit the ability of HSSA.

Evaluation on Dialogue Understanding. We evaluate our method on four popular downstream tasks for dialogue understanding, including IC, Senti, CtxQ, and CtxR. Unlike the evaluation on dialogue representation, we fine-tune the models with task-specific datasets. The results in Table 4 show that: (1) Compared to all baselines, DIALOG-POST yields substantial improvements across four understanding tasks, achieving **3.9%**, **3.5%**, **1.7%**, and **2.3%** absolute improvements against previous SOTA approaches on IC, Senti, CtxQ, and CtxR tasks, respectively. (2) DIALOG-POST also leads to further improvement (+0.3% average) compared with DIALOG-POST-BERT, revealing the capacity of HSSA on grasping the structure of dialogues. We argue that, understanding dialogues (e.g., the intents and emotions) relies on deep semantic meanings from the hierarchical dialogue structure, which requires the model to catch the multi-granularity semantic relations from the tokens, utterances, and the whole dialogue. By harnessing the multi-level SSOs and HSSA model, our method can better understand the intrinsic dialogue structure, and finally boost performance on downstream tasks.

4 Discussion

In this section, we conduct further in-depth discussions to analyse the HSSA model, the contribution of each SSO, and visualize the training process of CMTL. Due to space limitation, we report the results of JDDC/SR, JDDC/D-STS, IC and Senti.

4.1 Ablation Study of HSSA

As mentioned in Section 2.2, we stack 10 layers of HSSA blocks and 2 layers of Transformer blocks in our model. The last 2 Transformer layers are devised to capture the full dialogue semantics based on the global self-attention (SA) mechanism. Here, we first replace the last 2 Transformer layers with 2 HSSA layers (denoted as "w/o trs"). Table 5 shows that the performance degrades significantly on D-STS, SR, and IC, indicating the necessity of global self-attention. It is notable that the performance of Senti becomes slightly better with all HSSA blocks. Since the input of Senti task is an utterance without context, it is possible that the 12-layer HSSA focusing on the local attention has some advantages. Moreover, we also try to remove the updater ("w/o updater"), the inter-segment self attention ("w/o $\mathbf{H}_{int}$"), or the inner-segment self attention ("w/o $\mathbf{H}_{inn}$") sub-layer from HSSA. The results in Table 5 demonstrate that all variants lead to a pronounced performance degradation, which proves the rationality of each sub-layer.

Model	D-STS	SR	IC	Senti
HSSA	**82.90**	**69.95/79.87**	**91.8**	78.1
w/o trs	78.92	65.40/76.31	91.0	**78.5**
w/o updater	74.20	65.61/74.35	88.6	77.6
w/o $\mathbf{H}_{int}$	58.75	49.83/65.74	86.8	75.2
w/o $\mathbf{H}_{inn}$	45.97	48.64/63.22	76.6	68.9

Table 5: The ablation results of HSSA model.

Method	D-STS	SR	IC	Senti
DIALOG-POST	**82.90**	**69.95**/79.87	**91.8**	**78.1**
w/o DRM	82.84	69.93/**79.90**	91.2	77.9
w/o DSM	82.76	69.16/78.65	91.0	77.4
w/o DUC	81.96	69.25/79.69	89.7	77.4
w/o DUP	81.75	68.99/79.13	91.0	77.8
w/o DCL	77.98	61.21/75.33	89.0	77.0

Table 6: The ablation results of SSOs in DIALOG-POST.

4.2 Ablation Study of SSOs

Here, we conduct the ablation study for five SSOs. We follow the same training order (DRM → DSM → DUC → DUP → DCL) as CMTL mentioned in Table 1, but remove one training objective each time while keeping the remaining four. Table 6 shows that each training objective contributes to the overall performance to some extent, indicating the multi-level SSOs are complementary. Besides, DCL brings the most benefits, which implies the effectiveness of DCL on capturing the content-relatedness of context-context pairs.

4.3 Visualization of CMTL Training

Figure 3 illustrates the curves of training loss for each task across different training steps. The lines

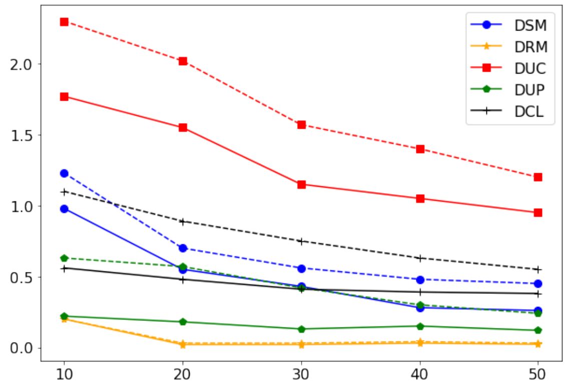

Figure 3: Visualization of training process. Horizontal/vertical axis represents the training steps (K)/loss. The solid lines and dashed lines represent the CMTL training and the single task training respectively.

with the same color represent the training loss of with the same objective. Here, we compare the training loss of CMTL and single task training. For example, for the red lines, the solid one is much lower than that of the dashed one, which indicates that the training process converges much faster by applying CMTL. It also shows the former training tasks may facilitate the latter, and finally promote the stability of the whole model training.

5 Related Work

Dialogue Encoding Networks. To handle the particularity of dialogue structure, previous works have proposed several typical networks for dialogue encoding, including hierarchical attention-based models (Jiao et al., 2019; Zhu et al., 2020; Li et al., 2020b), recurrence-based models (Shen et al., 2021b; Yang et al., 2019), and long conversation oriented models (Zhong et al., 2021). For example, Jiao et al. (2019) and Zhu et al. (2020) encode each utterance at first, and then leverage LSTMs and Transformers to aggregate the utterances, while Shen et al. (2021b) use memory caches to encode utterances sequentially. Huang et al. (2021) integrate sparse attention to encode long dialogue sequences (e.g., with 5000 words).

In this work, we propose a novel dialogue encoding network HSSA to capture the semantic structure of dialogues. It takes the dialogue as input and leverages inner-segment and inter-segment self-attention to capture the hierarchical dependencies. Finally, we devise an updater to obtain the contextual encoding of dialogues by aggregating the inner- and inter-segment representations.

Dialogue Post-training. With the booming of PLMs (Devlin et al., 2019; Radford et al., 2019; Bao et al., 2020a), researchers try to apply PLMs in the field of dialogues. An intuitive idea is to conduct a second-stage pre-training with massive dialogue corpora, but without changing the training objectives (Zhang et al., 2020; Xu et al., 2021). Recently, some works (Jiao et al., 2019; Feng et al., 2020; Zhang et al., 2021; Xu and Zhao, 2021) are proposed to design several new objectives for dialogue-adaptive post-training and achieve astonishing performance on downstream tasks of dialogue understanding. PLATOs (Bao et al., 2020b, 2021) leverage the large-size unified language model (Dong et al., 2019) to fulfill context encoding and response generation tasks with curriculum learning. Response selection is widely used as a self-supervised post-training task due to the convenience of constructing training data (Mehri et al., 2019; Su et al., 2021; Liu et al., 2021; Whang et al., 2021). Wu et al. (2021) propose Span Boundary Objective and Perturbation Masking Objective to capture the dialogue semantics in span and token levels. Above works either focus on token-level or utterance-level semantics, and only consider a small set of dialogue attributes.

Differently, we propose five self-supervised objectives in token, utterance and dialogue levels, aiming to modeling multi-facet dialogue attributes, including fact-awareness, speaker-shift, coherence, and content-relatedness of both utterance-response pairs and context-context pairs.

6 Conclusion and Future Work

In this paper, we propose a novel dialogue-adaptive post-training method, DIALOG-POST, by devising five multi-level training objectives and a hierarchical dialogue encoder network. These training objectives capture the multi-facet attributes of dialogues by leveraging token-level, utterance-level, and dialogue-level self-supervised signals. The dialogue encoder learns the hierarchical semantic structure of dialogues. To validate the effectiveness of our method, extensive experiments on dialogue representation and understanding tasks are conducted. Experimental results demonstrate the competitiveness of our method against strong baselines of both tasks. In the future, we will explore more efficient model architectures and try to pretrain the dialogue-oriented PLMs from scratch.

Limitations

Although the proposed method achieves exciting results, there are still some issues that need to be addressed in the future: (1) When designing the structure of HSSA layers, we assume that humans tend to understand a dialogue from the local to global perspective, which supports the existence of inner- and inter-segment self-attention layers. (2) We use 2 public Chinese corpora, JDDC and EDC, for post-training. Though there are diverse topics in them, it is desired to introduce other corpora from different domains and languages. (3) SSL tasks are arranged in post-training via CMTL (Sun et al., 2020) based on the intuitive understanding of their semantic levels and difficulties. Therefore, to combine the power of each SSL task more effectively, new training strategies need to be explored.

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020a. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020b. PLATO: pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 85–96. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2513–2525. Association for Computational Linguistics.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 459–466. European Language Resources Association.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Shaoxiong Feng, Hongshen Chen, Kan Li, and Dawei Yin. 2020. Posterior-gan: Towards informative and coherent response generation with posterior generative adversarial network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7708–7715. AAAI Press.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Matthew Henderson, Iñigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP*

2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2161–2174. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2019. Pt-code: Pre-trained context-dependent encoder for utterance-level emotion recognition. *CoRR*, abs/1910.08916.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020a. Task-specific objectives of pre-trained language models for dialogue adaptation. *CoRR*, abs/2009.04984.

Tianda Li, Jia-Chen Gu, Xiaodan Zhu, Quan Liu, Zhen-Hua Ling, Zhiming Su, and Si Wei. 2020b. Dialbert: A hierarchical pre-trained model for conversation disentanglement. *CoRR*, abs/2004.03760.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2396–2406. Association for Computational Linguistics.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskénazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3836–3845. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.

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.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021a. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1598–1608.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021b. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13789–13797. AAAI Press.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1740–1751. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of* the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 9929–9939. PMLR.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. *Proc. Interspeech 2020*, pages 1585–1589.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14041–14049. AAAI Press.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 917–929. Association for Computational Linguistics.

Han Wu, Kun Xu, Linfeng Song, Lifeng Jin, Haisong Zhang, and Linqi Song. 2021. Domain-adaptive pre-training methods for dialogue understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 665–669. Association for Computational Linguistics.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14158–14166. AAAI Press.

Yi Xu and Hai Zhao. 2021. Dialogue-oriented pre-training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2663–2673. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

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

Zhenyu Zhang, Tao Guo, and Meng Chen. 2021. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3647–3651. ACM.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3740–3752. Association for Computational Linguistics.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pre-training for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5134–5145.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Who did they respond to? conversation structure modeling using masked hierarchical transformer. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9741–9748. AAAI Press.

A Appendix

A.1 Details of Dialogue Understanding Tasks

In this section, we introduce the details of dataset annotation and show some examples from the dialogue understanding task. The original JDDC (Chen et al., 2020) corpus provides intent labels for each utterance, and three challenging sets of response generation. Considering intent classification, sentiment recognition, context-query matching, and context-response matching are very common tasks of dialogue applications in industry, we construct an evaluation dataset for dialogue understanding, which consists of 4 downstream tasks.

We sample 5,000 dialogues from JDDC and invite 4 graduate students to finish the annotation. For each data sample, at least three people finish the annotation and the majority voting is applied to decide the final label. The annotation agreement (Fleiss' Kappa (Fleiss and Cohen, 1973) score) is 0.83, showing the good quality of the annotation. The evaluation sets are derived from JDDC corpus, and we hope they can facilitate the dialogue understanding for future research.

We list the description of each task below, and show some examples in Table 8. Note that for a dialogue $\mathcal{D} = \{u_1, s_1, u_2, s_2, ..., u_{n-1}, s_{n-1}, u_n, s_n\}$, u_n and s_n are the current user query and staff response, and the previous utterances are denoted as context. u and s represent the user and service staff.

- Intent Classification (**IC**) aims to predict the intent of user query based on the dialogue context. Since the JDDC corpus is in E-commerce scenario, the intents are related to E-commerce activities and actions, such as "Warranty and return policy", "Delivery duration", "Change order information", and "Check order status". Understanding user intents is the foundation of industrial dialogue systems. Since context plays a critical role in intent classification, we combine the current user query and last two user utterances before it as an unit, and annotate the intent label for the current user query.

- Sentiment Recognition (**Senti**) aims to detect the emotions from user utterances. The categories include "happy", "sad", "angry", "feared", "disappointed", "anxious", and "other". For this task, each user utterance is considered individually for annotation.

- Context-Question Matching (**CtxQ**) aims to determine whether the semantic meanings are similar given a context-question pair. CtxQ is widely used to find a question in "frequent asked question (FAQ)", which is highly-relevant to a context, and return its answer as the response to context. Before that, the standard question-answer (QA) pairs are stored in the database.

- Context-Response Matching (**CtxR**) aims to determine whether an utterance can be the appropriate response to a given context. The task is also denoted as response selection if multiple response candidates were given.

The classes in training set are uniformly distributed with each class holds nearly same amount of examples. For test sets, the largest class holds 90 examples and the rest classes each hold about 30 examples. For test set of Senti, each class holds roughly 40 to 50 examples. The $|positive|$: $|negative|$ (examples) are 317:276 and 301:319 for CtxR and CtxQ respectively.

A.2 Training Efficiency and Memory Cost

Table 7 illustrates the comparison of BERT and HSSA on memory cost and training speed, and HSSA is more computationally efficient, especially for long dialogues. We post-train the models on Tesla P40 GPUs with batch size of 16.

	Memory (MiB)		Speed (steps/s)	
Length	Ours	BERT	Ours	BERT
128	5,407	5,537	1.90	2.06
256	7,799	8,817	1.39	1.38
384	10,405	13,095	1.07	1.02
512	13,615	18,737	0.84	0.78

Table 7: The memory cost and training speed comparison between DIALOG-POST (ours) and DIALOG-POST-BERT given different dialogue lengths.

A.3 Complete Ablation Study of HSSA

In Section 4.1, we only show the experimental results on 4 tasks due to space limitation. Here, we supplement the complete experimental results on all test sets in Table 9 and 10 to demonstrate the contribution of each module in HSSA.

A.4 Complete Ablation Study of SSOs

We illustrate the complete experimental results for the ablation study of SSOs mentioned in Section 4.2. Table 11 and 12 show the results on dialogue representation and understanding tasks.

Task	Chinese	English	Label
IC	u_1: 保修多长时间?	u_1: How long is the warranty period?	-
	u_2: 我想把地址换一下。	u_2: I wanna change my post address.	-
	u_3: 我忘改地址了。	u_3: Because I forget to change the address.	Change order information
Senti	u: 发票还没给我呀?	u: I haven't received my invoice yet.	anxiety
	u: 为什么刚买完就降价?	u: Why do you cut price just after I bought it?	disappointed
CtxQ	u_1: 你好	u_1: Hi.	-
	s_1: 您好，国庆节快乐，有什么可以帮您?	s_1: Hi, Happy National Day. How can I help you?	-
	u_2: 安装和架子多少钱?	u_2: How much is the installation and shelf?	-
	q: 支架多少钱?	q: How much is the shelf?	Matched
CtxR	u_1: 请问怎么调节冰箱温度去除结霜?	u_1: How can I adjust the temperature of the fridge to remove the frost?	-
	s_1: 定期除霜就可以了哦	s_1: You just need to defrost on time.	-
	u_2: 是不是调这个?	u_2: Should I set this?	-
	r: 洗衣机4个底脚都可以调整，范围在1cm左右	r: All the feet of the washing machine can be adjusted within 1cm.	Mismatched

Table 8: Examples of four dialogue understanding tasks. For CtxQ and CtxR, q and r represent the candidate question and response respectively.

Model	JDDC			ECD		
	Corr.	MAP	MRR	Corr	MAP	MRR
HSSA	**82.90**	**69.95**	**79.87**	**83.91**	**71.65**	**81.72**
w/o trs	78.92	65.40	76.31	79.84	68.25	78.86
w/o updater	74.20	65.61	74.35	75.67	67.33	77.85
w/o $\mathbf{H}_{int}$	58.75	49.83	65.74	56.92	59.86	74.99
w/o $\mathbf{H}_{inn}$	45.97	48.64	63.22	29.65	49.57	69.02

Table 9: Experimental results of HSSA Ablation Study on all dialogue representation tasks.

Model	IC	Senti	CtxQ	CtxR	Average
HSSA	**91.8**	78.1	**92.4**	**87.9**	**87.5**
w/o trs	91.0	**78.5**	91.2	87.2	87.0
w/o updater	88.6	77.6	90.5	86.5	85.8
w/o $\mathbf{H}_{int}$	86.8	75.2	87.9	82.7	83.2
w/o $\mathbf{H}_{inn}$	76.6	68.9	82.4	73.0	75.2

Table 10: Experimental results of HSSA Ablation Study on all dialogue understanding tasks.

Method	JDDC			ECD		
	Corr.	MAP	MRR	Corr.	MAP	MRR
DIALOG-POST	**82.90**	**69.95**	**79.87**	83.91	71.65	**81.72**
w/o DRM	82.84	69.93	79.90	**83.95**	71.64	81.72
w/o DSM	82.76	69.16	78.65	83.62	**71.69**	81.24
w/o DUC	81.96	69.25	79.69	83.91	71.64	81.72
w/o DUP	81.75	68.99	79.13	83.58	71.18	81.71
w/o DCL	77.98	61.21	75.33	80.16	67.35	79.06

Table 11: Experimental results of SSOs Ablation Study on all dialogue representation tasks.

Method	IC	Senti	CtxQ	CtxR	Average
DIALOG-POST	**91.8**	**78.1**	**92.4**	**87.9**	**87.5**
w/o DRM	91.2	77.9	91.8	87.0	87.0
w/o DSM	91.0	77.4	90.9	86.9	86.6
w/o DUC	89.7	77.4	90.3	85.1	85.6
w/o DUP	91.0	77.8	91.2	86.7	86.7
w/o DCL	89.0	77.0	89.6	86.5	85.5

Table 12: Experimental results of SSOs Ablation Study on all dialogue understanding tasks.