Semantic-based Pre-training for Dialogue Understanding

Xuefeng Bai[*♠◇], **Linfeng Song**[♣], **Yue Zhang**[♠◇]
[♠] School of Engineering, Westlake University, China
[♣] Tencent AI Lab, Bellevue, WA, USA
[◇] Institute of Advanced Technology, Westlake Institute for Advanced Study, China

Abstract

Pre-trained language models have made great progress on dialogue tasks. However, these models are typically trained on surface dialogue text, thus are proven to be weak in understanding the main semantic meaning of a dialogue context. We investigate Abstract Meaning Representation (AMR) as explicit semantic knowledge for pre-training models to capture the core semantic information in dialogues during pre-training. In particular, we propose a semantic-based pre-training framework that extends the standard pre-training framework (Devlin et al., 2019) by three tasks for learning 1) core semantic units, 2) semantic relations and 3) the overall semantic representation according to AMR graphs. Experiments on the understanding of both chit-chats and task-oriented dialogues show the superiority of our model. To our knowledge, we are the first to leverage a deep semantic representation for dialogue pre-training.

1 Introduction

Dialogue systems have attracted increasing attention from both academia and industry researches (Chen et al., 2017; Deriu et al., 2021; Gao et al., 2021a). The tasks can be commonly divided into two categories: task-oriented dialogue systems (Wen et al., 2017; Dinan et al., 2019; Mehri et al., 2020) and chit-chat dialogue systems (Ritter et al., 2011; Li et al., 2017; Yu et al., 2020; Cui et al., 2020; Chen et al., 2021, 2022; Song et al., 2022). The former aims to interact in the context of a specific task, while the latter chats with users without task and domain restrictions. Despite differences in goals, a common challenge for both tasks is understanding the semantic information conveyed in a dialogue history.

Recently, semantic representations from pre-trained language models have achieved remarkable success on a spectrum of dialogue tasks (Wen et al., 2015; Zhang et al., 2020; Wu et al., 2020; Gu et al., 2021; Zeng et al., 2021; Zhang and Zhao, 2021; Cui et al., 2021), where knowledge learned in pre-training over large-scale dialogue corpora can be transferred to downstream applications. Current pre-training techniques typically focus on the surface text. However, they do not explicitly consider deep semantic clues beyond text, which leads to some unexpected behavior, such as paying attention to meaningless words (Mudrakarta et al., 2018), and suffering from spurious feature associations (Kaushik et al., 2020) and adversarial attacks (Jia and Liang, 2017).

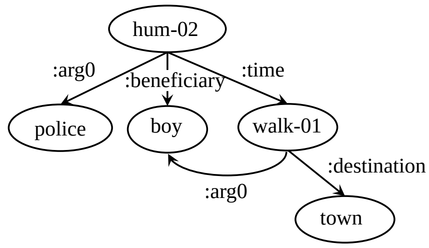


Figure 1: An AMR graph for sentence "*The police hummed to the boy as he walked to town.*"

Incorporating semantic information into dialogue systems has been shown to be helpful for many downstream tasks, such as dialogue intent prediction (Gupta et al., 2018), dialogue state tracking (Cheng et al., 2020), and dialogue relation extraction (Bai et al., 2021). These methods first parse dialogue turns into semantic structures, and then incorporate them as extra features into neural systems. However, they 1) only focus on domain-specific benchmark data, leaving the general potentiality of semantic structures unexploited; 2) require either human annotations or an external parser to obtain semantic structures, raising costs or/and causing error propagation for real applications.

We present **SARA**, a **S**emantic-graph-based

[*] Work done as an intern at Tencent AI Lab.

Proceedings of the 29th International Conference on Computational Linguistics, pages 592–607
October 12–17, 2022.

pre-tr**A**ining f**R**amework for di**A**logues, aiming to endow a pre-trained dialogue model with a stronger ability to infer semantic structures from conversations by using explicit semantic structures for more fine-grained supervisions. In particular, we exploit the abstract meaning representation (AMR; Banarescu et al. 2013), a fine-grained deep structure widely adopted in semantic parsing (Lyu and Titov, 2018; Zhang et al., 2019; Cai and Lam, 2020; Bevilacqua et al., 2021; Bai et al., 2022) and generation (Konstas et al., 2017; Song et al., 2018; Zhu et al., 2019; Bai et al., 2020; Ribeiro et al., 2021). As shown in Figure 1, AMR represents a sentence using a rooted directed graph, highlighting the core semantic units (e.g., "*police*", "*hum*", "*boy*") in a sentence and connecting them with semantic relations (e.g., "*:arg0*", "*:time*").

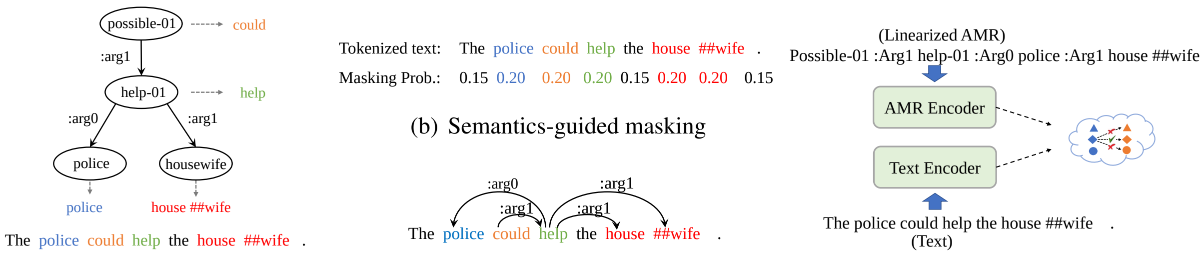
We explicitly leverage AMR graphs for pre-training our dialogue model. As shown in Figure 2, SARA consists of three pre-training sub-tasks: 1) semantic-based mask language modeling, which extends the standard mask language modeling task (Devlin et al., 2019) by paying more attention to core semantic units in a dialogue; 2) semantic relation prediction, which aims to learn semantic relations between words; 3) semantic agreement, which optimizes the overall similarity between a dialogue and its corresponding AMR graph. The SARA combines strengths of both powerful contextualized representation of pre-trained models and explicit semantic knowledge, while eliminating the requirement of an external semantic parser in downstream applications.

We choose BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models as backbone, which are then continual pre-trained on a large-scale conversation dataset using our framework. Experiments show that our semantic-based framework gives better results than current pre-training methods that use much more training data, achieving new state-of-the-art results on both chit-chat understanding (dialogue relation extraction) and task-oriented dialogue understanding tasks (DialoGLUE benchmark). Our method also gives better results than previous semantic-base systems on downstream tasks, without using an external parser. Further analysis suggests that semantic information introduced by AMR can help our model to better understand semantically complex dialogues. To our knowledge, we are the first to leverage deep semantic representation for dialogue pre-training. Our code and the pre-trained models are available at `https://github.com/goodbai-nlp/Sem-PLM`.

2 Related Work

Pre-training for Dialogue. Inspired by the success of pre-trained language models in the general domain (Peters et al., 2018; Radford and Narasimhan, 2018; Devlin et al., 2019; Lewis et al., 2020), various pre-trained models have been proposed in the domain of dialogue. DialoGPT (Zhang et al., 2020) continual pre-trains a GPT-2 (Radford et al., 2019) model directly on Reddit comments data. ConvRT (Henderson et al., 2019) pre-trains a dual Transformer encoder for the response selection task. PLATO (Bao et al., 2020) introduces a latent variable-based model for dialogue response generation pre-training. TOD-BERT (Wu et al., 2020) pre-trains a Transformer encoder on task-oriented dialogue corpus for task-oriented dialogue applications. MPC-BERT (Gu et al., 2021) continues to pre-train a BERT model with self-supervised tasks based on the interactions among utterances and interlocutors. SPIDER (Zhang and Zhao, 2021) continues to pre-train a BERT model with auxiliary tasks to predict the utterance order and understand the sentence backbone. DialogLM (Zhong et al., 2022) pre-trains a generative Transformer encoder on long conversations with window-based pre-training tasks. Our work is similar in that we also pre-train a model on the dialogue corpora. However, unlike these previous studies, which focus on text level distributions, we additionally enhance the model with semantic structures.

Semantics for dialogue. Semantic knowledge has been used for both social chat and task-oriented dialogues systems. PEGASUS (Zue et al., 1994) transforms a sentence into a semantic frame which is then used for travel planing. Wirsching et al. (2012) design a dialogue system which performs database operations based on semantic features. Gupta et al. (2018) and Aghajanyan et al. (2020) integrate intents and slots into a semantic tree and solve intent classification and slot-filling tasks as semantic parsing. Cheng et al. (2020) represent task-oriented dialogue as a semantic graph to perform dialogue state tracking. A most related work is Bai et al. (2021), who build dialogue-level AMR graphs for both social chat understanding and dialogue response generation. Our work is similar



(a) Example AMR graph. (b) Semantics-guided masking (c) Semantic relation prediction (d) Semantic agreement

Figure 2: The semantic-based pre-training framework.

in showing the effect of semantic knowledge for improving dialogue understanding. However, different from them, we focus on enhancing the language model with semantic knowledge during pre-training, and our model does not require an external AMR parser in downstream applications.

3 Method

Figure 2 illustrates our semantic-based pre-training framework for dialogues. We take a pre-trained Transformer (Vaswani et al., 2017) encoder as the backbone, using AMR as explicit semantic knowledge to continuously pre-train the model on dialogues in a multitask setting. In particular, the following three semantic-aware tasks are designed:

- Semantics-based masking (Section 3.1).
- Semantic relation prediction (Section 3.2).
- Semantic agreement (Section 3.3).

The former two learn semantic knowledge from AMR nodes and AMR edges, respectively. The last task regularizes the overall representation of a dialogue using graph-level semantic features.

We follow Bai et al. (2021) and construct dialogue-level AMR graphs by 2 steps: 1) building utterance-level AMR graphs by independently transforming utterances into AMR using a pre-trained AMR parser. 2) connecting utterance-level AMR graphs with a root node, where edges are labeled with the corresponding speaker.

Formally, denote an input dialogue sequence[1] as $\boldsymbol{x} = [x_1, x_2, ..., x_n]$, where n is the number of tokens in the dialogue. The corresponding AMR is a directed acyclic graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ denotes a set of nodes (i.e., AMR concepts) and $\mathcal{E}$ (i.e., AMR relations) denotes a set of labeled edges. An edge can be further represented by a triple $\langle v_i, r_{ij}, v_j \rangle$, meaning that the edge is from node v_i to v_j with label r_{ij}.

3.1 Task 1: Semantics-guided Masking

We first formally present the vanilla mask language modeling (MLM) setup, before introducing the semantic-guided masking strategy.

Vanilla MLM. Given a sequence of tokens $\boldsymbol{x}$, the standard masking strategy (Devlin et al., 2019) selects a set fraction of tokens positions (denoted as $\boldsymbol{m} = [m_1, m_2, ..., m_k]$) for masking independently at random, and use these "*selected*" tokens $\{x_i | i \in \boldsymbol{m}\}$ as supervisions to train a language model. Formally, denoting the masked text as $\tilde{\boldsymbol{x}}$, vanilla MLM optimizes the following training objective:

$$\ell_{vanilla\_mlm} = -\sum_{i \in \boldsymbol{m}} \log P(x_i | \tilde{\boldsymbol{x}}), \quad (1)$$

where the conditional probability $P(x_i|\tilde{\boldsymbol{x}})$ is generated by an encoder model with a softmax layer.

Semantics-guided Masking. A salient limitation of vanilla MLM is that it treats all tokens equally, thus potentially wasting resources on tokens that provide little signal (e.g., punctuations and stop words). We introduce a semantic-guided masking strategy, encouraging model to give more attention on *semantic-aware* units, which are expected to have more influence on text understanding. As shown in Figure 2(b), our semantic-guided masking strategy gives a higher masking probability for tokens (e.g. "*police*", "*could*", "*help*") that contain important semantic information. Formally, we define a token as a *semantic-aware* unit when it is aligned with an AMR node, according to the AMR-to-text alignment $\mathcal{A}$[2] (An example is given in Figure 2(a)). Since pre-trained models typically use a vocabulary with sub-word units (Sennrich et al., 2016), for an alignment pair $\langle v_i, x_j \rangle$, we extend the alignment as $\left\langle v_i, \{x_j^1, x_j^2 ..., x_j^l\} \right\rangle$, where the AMR node v_i is aligned a set of all tokens $\{x_j^1, x_j^2 ..., x_j^l\}$

[1]Please refer to Appendix B for dialogue input format.

[2]$\mathcal{A}$ is a one-to-K mapping ($K \in [1, \ldots, n]$).

which are sub-words of word w_j. For example, in Figure 2, the AMR node "*housewife*" is aligned with sub-tokens "*house*" and "*##wife*".

Denoting $\boldsymbol{m'} = [m'_1, m'_2, ..., m'_k]$ as token indices selected by the proposed semantic-guided masking strategy, the training objective is:

$$\ell_{sem\_mlm} = -\sum_{i \in \boldsymbol{m'}} \log P(x_i | \tilde{\boldsymbol{x}}). \quad (2)$$

We follow RoBERTa (Liu et al., 2019) and use the dynamic masking, where we generate the masking pattern every step instead of performing masking during data preprocessing.

3.2 Task 2: Semantic Relation Prediction

The semantic relation prediction task is designed for learning the semantic relations between words. To this end, we project the edges of each input AMR graph onto the corresponding sentence according to their node-to-word alignments (as shown in Figure 2(c)), before training a predictor to generate the projected edges.

Relation Projection. Since AMR relations are defined on AMR nodes instead of words in the dialogue text, we use a node-to-word alignment $\mathcal{A}$ to project the AMR edges $\mathcal{E}$ onto text with following rules:

$$\hat{r}_{ij} = \begin{cases} r_{i'j'}, & \text{if } x_i \in \mathcal{A}(v_{i'}), x_j \in \mathcal{A}(v_{j'}), \\ None, & \text{otherwise.} \end{cases} \quad (3)$$

The same strategy in Section 3.1 is used to deal with sub-word tokens.

Relation Prediction. We first use a Transformer encoder to generate contextualized word hidden states $\boldsymbol{h} = [h_1, h_2, ..., h_n]$. Based on that, a deep biaffine neural parser (Dozat and Manning, 2017) is used to predict the relations between words. To determine whether a directed edge (or *arc*) from x_i to x_j exists, the biaffine parser first uses two separate MLPs (denoted as $\texttt{MLP}^H$ and $\texttt{MLP}^D$) to obtain two lower-dimensional representation vectors for each position, then calculates scores via a biaffine operation:

$$\begin{aligned} r_i^H, r_j^D &= \texttt{MLP}^H(h_i), \texttt{MLP}^D(h_j), \\ s_{ij}^{arc} &= \begin{bmatrix} r_j^D \\ 1 \end{bmatrix}^T W^{arc} r_i^H, \\ P(y_{ij}^{arc}|\boldsymbol{x}) &= \text{softmax}_j(s_i^{arc}), \end{aligned} \quad (4)$$

where r_i^H is the representation vector of x_i as a head word, and r_j^D denotes the vector of x_j as a dependent word. $P(y_{ij}^{arc}|\boldsymbol{x})$ is the probability of the *arc* (i, j), and W^{arc} is a parameter matrix. To calculate the probability of assigning a label l to the *arc*(i, j), which is denoted as $P(y_{ijl}^{label}|\boldsymbol{x})$, the biaffine parser uses the same scorer as in Equation 4 but with different parameters for MLPs and biaffines.[3]

The training objective of relation prediction is:

$$\ell_{rel} = -\sum_{\langle x_i, \hat{r}_{ij}, x_j \rangle \in \mathcal{E}'} \log P(y_{ij}^{arc}|\boldsymbol{x}) P(y_{ij\hat{r}_{ij}}^{label}|\boldsymbol{x}), \quad (5)$$

where $\mathcal{E}'$ represents the *projected* AMR edges.

3.3 Task 3: Semantic Agreement

We encourage the model to learn the overall agreement of a dialogue and its corresponding AMR graph. As shown in Figure 2(d), we use an auxiliary network to encode the AMR, and maximize the similarity score between the hidden states of text and AMR. Following previous work (Konstas et al., 2017), we linearize AMR graphs into a sequence (refer to Figure 2(d) for an example) and use a pre-trained encoder to transform AMR into a set of hidden states.[4]

Formally, defining the linearized AMR graph as $\boldsymbol{g} = [g_1, g_2, ..., g_m]$, the vector representation of text and its corresponding AMR is calculated as:

$$\begin{aligned} h^{\text{text}} &= \texttt{Pooling}(\texttt{TextEnc}(\boldsymbol{x})), \\ h^{\text{amr}} &= \texttt{Pooling}(\texttt{TextEnc}(\boldsymbol{g})), \end{aligned} \quad (6)$$

where $\texttt{TextEnc}(\cdot)$ and $\texttt{TextEnc}(\cdot)$ are text encoder and AMR encoder, respectively. They are initialized with the same weights but updated separately during training. $\texttt{Pooling}(\cdot)$ is a function that reduces that sequence of vectors into one vector. Following BERT (Devlin et al., 2019), we feed the hidden state of the first input token into a MLP layer to get the "pooled" vector.

We use the cosine similarity as a distance scoring function and adopt the contrastive learning framework (Hadsell et al., 2006; Frosst et al., 2019; Gao et al., 2021b; Luo et al., 2022) to train our model, with the aim to pulling semantically close text-AMR pairs and pushing apart unpaired examples. In particular, for a given text $\boldsymbol{x}$, the *positive* example is its corresponding AMR graph $\boldsymbol{g}$, the *negative* examples are the AMR graphs of its

[3]The biaffine parameter for label scoring is a three dimensional tensor.

[4]We also tried a structure-aware encoder but without observing significant improvements.

Dataset	DialogRE	BANKING77	HWU64	CLINC150	REST8K	DSTC8	TOP	MULTIWOZ
train	5,997	8,622	8,954	15,000	7,244	5,023	31,279	56,774
dev	1,914	1,540	1,076	3,000	1,000	602	4,462	7,374
test	1,862	3,080	1,076	4,500	3,731	1,813	9,042	7,372

Table 1: Statistics of datasets.

neighbor dialogues in the corpus. Formally, let h_i^{text} and h_i^{amr} denote the representations of the ith $\langle text, AMR\rangle$ pair in the dataset, the training objective is:

$$\ell_{sim} = -\log \frac{\exp(\text{sim}(h_i^{\text{text}}, h_i^{\text{amr}})/\tau)}{\sum_{j\in\mathcal{N}(i)} \exp(\text{sim}(h_i^{\text{text}}, h_j^{\text{amr}})/\tau)}, \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity, $\mathcal{N}(i)$ collects neighbor index of the ith example, and $\tau > 0$ denotes the temperature hyper-parameter.

3.4 Training

Our model is trained by optimizing the total loss of above 3 tasks:

$$\ell_{total} = \ell_{sem\_mlm} + \alpha\ell_{rel} + \beta\ell_{sim}, \tag{8}$$

where α and β are weighting hyper-parameters for ℓ_{rel} and ℓ_{sim}, respectively. To make the computational requirements feasible, we do not train our model from scratch, but rather continue training a model that has been pre-trained on textual inputs. Our framework is architecture-flexible and can be be applied to different models such as BERT, RoBERTa, and BART.

4 Experiments

We evaluate the effectiveness of our semantic pre-training model on 8 dialogue tasks and compare the results with the state-of-the-art pre-trained and semantic-enriched models.

4.1 Dataset

Pre-training Corpus. We continual pre-train our model on the Reddit (Henderson et al., 2019) corpus. After sampling and filtering (refer Appendix A), the dataset comprises 5,864,254 dialogue instances, in total 397 million words. We adopt the state-of-the-art AMRBART (Bai et al., 2022) parser [5] to transform the text into AMR graphs. To obtain the AMR-to-text alignment, we use the JAMR aligner[6] released by Flanigan et al. (2014).

Dialogue task datasets. We evaluate our model on both chitchat and task-oriented understanding tasks. For chitchat, we focus on the dialogue relation extraction task which aims to predict the relationship between an given entity pair. We report results on both original (v1) and updated (v2) English version of **DialogRE** (Yu et al., 2020).

For task-oriented dialogue, we report results on the DialoGLUE (Mehri et al., 2020) benchmark, which consists of 7 different datasets spanning 4 different tasks: 1) intention prediction, including **BANKING77** (Casanueva et al., 2020), **CLINC150** (Larson et al., 2019) and **HWU64** (Liu et al., 2021); 2) slot filling, including **RESTAURANT8K** (Coope et al., 2020) and **DSTC8** (Rastogi et al., 2019); 3) semantic parsing, **TOP** (Gupta et al., 2018); and 4) dialogue state tracking, **MULTIWOZ2.1** (Eric et al., 2020).

Table 1 shows the statistics of above datasets.

4.2 Settings

Model Configuration. We take BERT-base and RoBERTa-base as our backbone model. For Pre-training, AdamW (Loshchilov and Hutter, 2019) is used as an optimizer, with an initial learning rate of 1×10^{-5}. We reduce the learning rate according to a linear scheduler. The batch size is 2048, and the maximum input sequence length is 512. For the hyper-parameters, we empirically set $\alpha = 0.1, \beta = 1.0, \tau = 1.0$ in our experiments. The pre-training of our model is carried out on 8 Nvidia Telsa V100 32G GPU for 5 epochs, taking about 2 days to reach convergence. For fine-tuning, we follow previous works to set hyper-parameters. More details can be found in Appendix C.

Metrics. We use macro F1 and macro $F1_c$ for dialogue relation extraction (DialogRE), following Yu et al. (2020). For intent prediction (BANKING77, CLINC150, HWU64), we report the accuracy. Macro F1 (Coope et al., 2020) is adopted for slot filling tasks (RESTAURANT8K, DSTC8). For TOP, we use exact-match, which measures how often the model generates the exact reference structure. For MULTIWOZ, we use the joint goal accuracy following Budzianowski et al. (2018).

[5] https://github.com/muyeby/AMRBART
[6] https://github.com/jflanigan/jamr

Model	data-v1				data-v2			
	dev		test		dev		test	
	F1(δ)	F1$_c$(δ)	F1(δ)	F1$_c$(δ)	F1(δ)	F1$_c$(δ)	F1(δ)	F1$_c$(δ)
GDPNet	67.1 (1.0)	61.5 (0.8)	64.9 (1.1)	60.1 (0.9)	-	-	-	-
TUCORE-GCN	-	-	-	-	66.8 (0.7)	61.0 (0.5)	65.5 (0.4)	60.2 (0.6)
TSP	66.8 (0.9)	61.5 (1.0)	65.5 (0.7)	60.5 (0.8)	-	-	-	-
BERT	60.6 (1.2)	55.4 (0.9)	58.5 (2.0)	53.2 (1.6)	59.4 (0.7)	54.7 (0.8)	57.9 (1.0)	53.1 (0.7)
BERT$_s$	63.0 (1.5)	57.3 (1.2)	61.2 (0.9)	55.4 (0.9)	62.2 (1.3)	57.0 (1.0)	59.5 (2.1)	54.2 (1.4)
BERT$_c$	66.8 (0.9)	60.9 (1.0)	66.1 (1.1)	60.2 (0.8)	66.2 (0.9)	60.5 (1.1)	65.1 (0.8)	59.8 (1.2)
RoBERTa	68.0 (1.0)	60.3 (1.0)	66.0 (0.6)	59.6 (0.7)	67.6 (0.8)	61.0 (0.7)	65.8 (1.0)	59.6 (0.5)
SARA-BERT	68.1 (1.0)	62.1 (0.9)	67.5 (0.7)	61.4 (0.9)	68.0 (0.8)	62.1 (0.6)	67.3 (1.0)	61.3 (0.8)
SARA-RoBERTa	**69.3** (0.9)	**62.3** (0.8)	**68.1** (0.8)	**61.7** (1.0)	**69.5** (0.7)	**62.4** (0.5)	**67.8** (0.8)	**61.5** (0.7)

Table 2: Performance on DialogRE. We report the average and the standard deviation computed from 5 runs, best results are marked in **bold**.

4.3 Compared Models

For Dialogue relation extraction, we compare the proposed model with BERT-based models: **BERT** takes a pre-trained BERT as the dialogue encoder and predicts relation labels using the hidden state of the [CLS] token. **BERT$_s$** (Yu et al., 2020) enhances the speaker representation by marking speaker arguments with special tokens. **BERT$_c$** (Bai et al., 2021) concatenates hidden states of the [CLS] token and entity tokens for classification. For completeness, we also include recent methods which give the state-of-the-art results, such as **GDPNet** (Xue et al., 2020), **TUCORE-GCN** (Lee and Choi, 2021), **TSP** (Zhao et al., 2021) and **Hier** (Bai et al., 2021). We follow the implementation and hyper-parameters of BERT$_c$ to evaluate our model.

For DialoGLUE, the compared models include: **BERT** (Devlin et al., 2019) pre-trains a Transformer encoder on large-scale monotonic text. **USE** (Yang et al., 2020) pre-trains a dual Transformer encoder model on multilingual corpus using retrieval focused training tasks. **ConveRT (654M)** (Henderson et al., 2020) pre-trains a dual Transformer encoder on the full 2015-2019 Reddit data comprising 654M $\langle$*context*, *response*$\rangle$ training pairs using response selection task. **ConvBERT (700M)** (Mehri et al., 2020) fine-tunes a BERT model on 700M Reddit conversational data. We adopt the same implementation and hyper-parameters of ConvBERT (700M) to conduct experiments on DialoGLUE.

To verify the scalability of the proposed method, we also report results based on the RoBERTa model for all tasks. The model architectures for about tasks is given in Appendix D.

4.4 Main Results

Results on DialogRE. Table 2 lists the results of different systems on DialogRE. Among BERT-based models (i.e., BERT, BERT$_s$, BERT$_c$), BERT$_c$ reports the best results. Compared with BERT$_c$, SARA-BERT gives significantly ($p < 0.001$) better results on both datasets. In particular, SARA-BERT improves BERT$_c$ by 1.4 and 2.2 points in terms of F1 score on two test sets, respectively, indicating that our semantic pre-training framework is beneficial for dialogue relation extraction. The main reason can be that SARA improves the model capacity of understanding entities (which are core semantic units) and the semantic relations between them during pre-training stage.

SARA-BERT achieves better F1 scores than the other state-of-the-art methods. In addition, when using RoBERTa as the backbone, SARA gives consistent improvements. In particular, SARA-RoBERTa achieves 68.1 and 67.8 F1 scores on the test set of data-v1 and data-v2, respectively. To our best knowledge, these are the best-reported results based on RoBERTa-base.

Results on DialoGLUE. We report the results of different methods on the DialoGLUE benchmark in Table 3. Compared with BERT, SARA-BERT (6M) gives consistently better results on all 7 datasets, with an improvement of 1.1 point in average. In particular, SARA-BERT (6M) outperforms BERT by 2.1 and 3.0 points on HWU64 and MultiWOZ, respectively, showing that our SARA framework can benefit task-oriented dialogue systems.

Compared with the other state-of-the-art systems, SARA-BERT (6M) obtains better results than USE, because SARA-BERT (6M) is pre-trained on large-scale dialogue corpus. In addition, SARA-BERT (6M) gives highly competitive results than

Model	BANK	HWU64	CLINC150	REST8K	DSTC8	TOP	MULTIWOZ	Avg
USE	92.81	91.25	95.06	-	-	-	-	-
CONVERT (654M)	93.01	91.24	97.16	-	-	-	-	-
USE+CONVERT (654M)	93.36	**92.62**	**97.16**	-	-	-	-	-
CONVBERT (700M)	93.44	92.38	97.11	95.44	91.20	82.08	56.56	86.89
BERT	93.02	89.87	95.93	95.53	90.05	81.90	56.30	86.08
ROBERTA	93.16	91.30	96.09	96.27	90.78	81.80	54.95	86.28
SARA-BERT (6M)	93.47	92.01	96.24	95.92	91.57	82.05	**59.33**	87.23
SARA-ROBERTA (6M)	**93.64**	92.29	96.60	**96.74**	**92.02**	**82.78**	57.52	**87.37**

Table 3: Performance on DialoGLUE, best results are in **bold**. REST8K and BANK stands for RESTAURANT8K and BANKING77, respectively.

Model	DialogRE	DSTC8
ROBERTA	67.6	93.98
ROBERTA (6M)	68.2	94.17
SARA-ROBERTA (6M)	**69.5**	**95.24**
w/o sem_mlm	69.0	95.01
w/o rel_pred	68.6	94.63
w/o sem_agree	68.8	94.72

Table 4: Validation F1 of DialogRE and DSTC8.

CONVERT (654M), USE+CONVERT (654M) and CONVBERT (700M), using significantly fewer data (about 1% than others). This indicates that our semantic-based pre-training framework is more data-efficient. Finally, similar to SARA-BERT (6M), SARA-ROBERTA (6M) significantly ($p < 0.001$) outperforms ROBERTA, giving the best results on BANKING77, REST8K, DSTC8 and TOP.

5 Analysis

5.1 Ablation Study

We compare our full system with the following models: ROBERTA (6M) is continuously pre-trained on the exact same training corpus as our model using corresponding standard pre-training objectives; w/o sem_mlm, w/o rel_pred, and w/o sem_agree denote the models which are trained without the semantics-guided masking, semantic relation prediction, and semantic agreement task, respectively. Table 4 shows the F1 scores on the validation sets of DialogRE and DSTC8. First of all, using dialogue domain data (ROBERTA v.s. ROBERTA (6M)) for pre-training leads to improvements on both tasks. This meets previous observations (Gururangan et al., 2020; Mehri et al., 2020). Also, the semantic-based mask language modeling task (sem_mlm) gives an obvious improvement on DialogRE and a small one on DSTC8. The reason can be that DSTC8 has an average length of 8 tokens, making it easy to understand core semantic units in dialogues. In addition, the performance drops significantly without the relation prediction task (rel_pred), indicating that the rel_pred task is important for dialogue understanding. Furthermore, the semantic agreement task (sem_agree) is helpful for both datasets, showing that the AMR is beneficial to improve the overall semantic representation of dialogue. Finally, by combining dialogue domain data and all pre-training tasks, our final model achieves the best performance on both datasets.

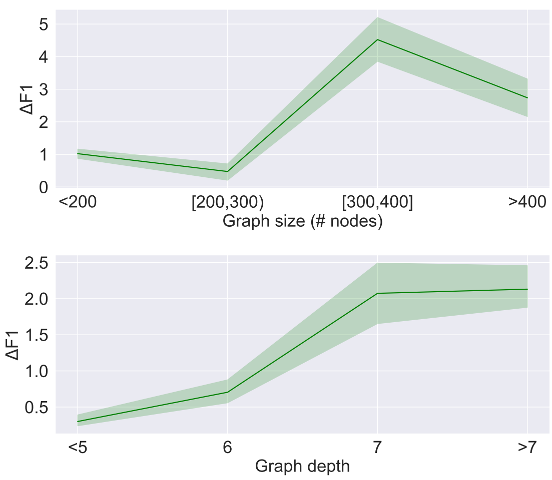


Figure 3: Performance improvement (ΔF1) over two aspects: (top) graph size and (bottom) graph depth.

5.2 Effect of Semantic-based Pre-training

To further understand the effectiveness of our semantic-based pre-training framework, we split the test set of DialogRE (v2) into different groups according to semantic complexity and report the performance improvement of SARA-ROBERTA over ROBERTA. In particular, two metrics are considered to measure the semantic complexity of a dialogue: 1) graph size (i.e., the number of nodes in the AMR graph) which records the number of

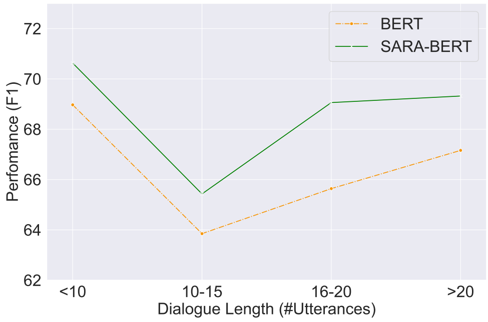


Figure 4: Test F1 on DialogRE (v2).

semantic units in the dialogue; 2) graph depth which is defined as the longest distance between the AMR node and root node. An AMR graph has a deeper depth means that its corresponding dialogue has more long-range dependencies.

As shown in the top sub-figure of Figure 3, SARA-ROBERTA gives consistent improvements over ROBERTA in different graph groups. In particular, the improvements are more considerable on graphs with more than 300 nodes, showing that SARA-ROBERTA has better capacity than ROBERTA in understanding dialogues which contain more semantic units. The reason can be that the semantic-based MLM task enhances the model ability to capture core semantic features, which helps in reducing negative impacts of meaningless tokens in dialogue text. With respect to graph depth, SARA-ROBERTA also outperforms ROBERTA on all groups, and larger improvements are observed on deeper graphs. It can be that the relation prediction task helps to establish semantic associations between non-neighbor words, thus benefiting long-range dependencies understanding.

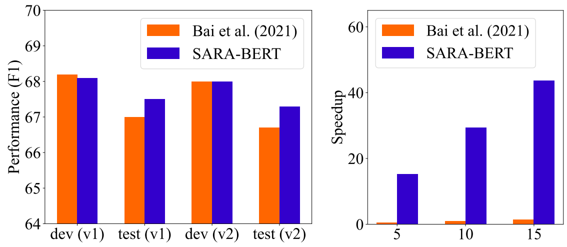
We also compare the model performance in terms of dialogue length. In particular, we split the test set of DialogRE (v2) into 4 groups according to the utterance number of each dialogue, and compare the performance of ROBERTA and SARA-ROBERTA. As shown in Figure 4, SARA-ROBERTA consistently gives better results than ROBERTA on all groups. The performance gap is bigger when the input dialogue has more than 16 utterances. The reason is that SARA encourages the model to understand core semantics, which is helpful for learning long dialogues.

5.3 Impact of AMR Features

AMR is a deep semantic structure which consists of both backbone relations and fine-grained semantic relations. To study the contribution of such features, we simplify an AMR graph by masking the fine-grained semantic relations, resulting in a graph with frame arguments relations (e.g., *:arg0, :arg1, :arg2*). We use the simplified graph as explicit semantic knowledge for pre-training and compare it with the standard AMR graph under the same framework.

Model	DialogRE	DSTC8
ROBERTA	65.8	90.78
SARA-ROBERTA (full)	**67.8**	**92.02**
SARA-ROBERTA (simplified)	67.3	91.34

Table 5: F1 on the test set of DialogRE and DSTC8.



(a) (b)

Figure 5: (a) Comparison of performance on DialogRE; (b) Comparison of inference speed regarding to dialogue length (measured by number of utterances).

Table 5 lists the performance of two systems. It can be observed that both *simplified* graphs and *full* AMR graphs lead to better performance. Compared with *simplified* graphs, using *full* AMR graph for pre-training leads to better results on both DialogRE and DSTC8, showing that the fine-grained semantic features can further improve the model performance.

5.4 Comparison with explicit AMR

Figure 5(a) compares the performance of our model with the method of Bai et al. (2021)[7] which use explicit AMR structures for dialogue applications. We report the F1 score on the test set of DialogRE. Compared with the system of Bai et al. (2021), our model gives comparable results on the validation set, and better results on the test set, without using an external AMR parser. This indicates that 1) our pre-training framework can efficiently transfer the learned semantic information to downstream tasks; 2) large-scale semantic-aware pre-training can give further improvement compared with using

[7]We choose the `Hier` model which has comparable parameters to our model.

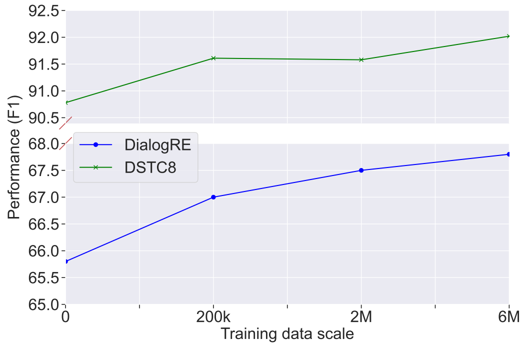


Figure 6: Impact of pre-training data.

semantic information in downstream tasks.

As shown in Figure 5(b), our system is significantly faster than the method of Bai et al. (2021) which relies on an external parser. As the dialogue length increases, the performance gap is more obvious. In particular, our system obtains about a 45 times speedup when the input dialogues have an average utterance number of 15.

5.5 Impact of Training Data Scale

Figure 6 shows the model performance regarding different scales of pre-training data. The performance on both DialogRE and DSTC8 datasets increases as the scale of training data grows bigger, with a margin of about 2.0 F1 score on DialogRE. Due to the limitation of computational resources, we do not conduct experiments on larger training corpus and models, and we leave this for a future work.

5.6 Case Study

Figure 7 shows an example conversation from DialogRE dataset. The baseline model (RoBERTa) is misled by sentences last three utterances (marked with underline) where *Speaker2* shows an negative emotions towards Rachel Green, and thus incorrectly predicting the relationship between two speakers as *negative_impression*. In contrast, our model (SARA-RoBERTa) predicts the correct relationship, suggesting that our semantic-based pre-training framework helps model to better understand the relationship between entity pairs and avoid focusing on spurious features.

Figure 8 presents a case of dialogue intent prediction. The baseline system pays much attention on word "*alarm*" while ignores other two core semantic units "*minutes*" and "*bake*", giving an incorrect prediction. Our system successfully predicts the gold intent, because AMR guides our

Speaker1: Wanna give me a hand? Speaker2: Sure! Monica, I can't get over how great you look! Speaker1: Oh umm, I meant to tell you, Ross is coming. Speaker2: Ross is coming. Great! I love Ross! Speaker1: Good, and Rachel Green too. Speaker2: Oh. Speaker1: Is there a problem? Speaker2: Nope. Uh, it's okay. It's just uh, God I hate her. Speaker1: What? Speaker2: Yeah, I hate her. She was horrible to me in school.
Ground-Truth: per:positive_impression (Speaker1, Speaker2)
Baseline: per:negative_impression (Speaker1, Speaker2) Ours: per:positive_impression (Speaker1, Speaker2)

Figure 7: An example of dialogue relation extraction.

How many minutes should I set an alarm for this bake?
Ground-Truth intent: cook_time
Baseline: alarm Ours: cook_time

Figure 8: An example of dialogue intent prediction.

model to discover the core semantic units in the dialogue text.

6 Conclusion

We investigated the abstract meaning representation as explicit semantic clues for dialogue pre-training, using a semantic-based pre-training framework. Experiments on two benchmarks show that the proposed framework is highly effective on both chit-chat understanding and task-oriented dialogue understanding. Our method gives the best results on multiple datasets.

Acknowledgments

Yue Zhang is the corresponding author. We would like to thank anonymous reviewers for their insightful comments. This work is supported by the Zhejiang Province Key Project 2022SDX-HDX0003 and the Tencent AI Lab Rhino-Bird Focused Research Program.

References

Armen Aghajanyan, Jean Maillard, Akshat Shrivastava, Keith Diedrick, Michael Haeger, Haoran Li, Yashar Mehdad, Veselin Stoyanov, Anuj Kumar, Mike Lewis, and Sonal Gupta. 2020. Conversational semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(*EMNLP*), pages 5026–5035, Online. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. 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 4430–4445, Online. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Xuefeng Bai, Linfeng Song, and Yue Zhang. 2020. Online back-parsing for AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1206–1219, Online. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

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

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor.*, 19(2):25–35.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yulong Chen, Ming Zhong, Xuefeng Bai, Naihao Deng, Jing Li, Xianchao Zhu, and Yue Zhang. 2022. The cross-lingual conversation summarization challenge. *CoRR*, abs/2205.00379.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 107–121, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2328–2337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Nicholas Frosst, Nicolas Papernot, and Geoffrey E. Hinton. 2019. Analyzing and improving representations with the soft nearest neighbor loss. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2012–2020. PMLR.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021a. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. Simcse: Simple 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 6894–6910. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation 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 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3682–3692. Association for Computational Linguistics.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 35–44. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2021–2031. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yun Luo, Fang Guo, Zihan Liu, and Yue Zhang. 2022. Mere contrastive learning for cross-domain sentiment analysis. *CoRR*, abs/2208.08678.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 397–407, Melbourne, Australia. Association for Computational Linguistics.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Mathias Lambert. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 97–105, Minneapolis, Minnesota. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Linfeng Song, Chunlei Xin, Shaopeng Lai, Ante Wang, Jinsong Su, and Kun Xu. 2022. Casa: Conversational aspect sentiment analysis for dialogue understanding. *J. Artif. Int. Res.*, 73.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Günther Wirsching, Markus Huber, Christian Kölbl, Robert Lorenz, and Ronald Römer. 2012. Semantic dialogue modeling. In Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent C. Müller, editors, *Cognitive behavioural systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011*, volume 7403.

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)*, pages 917–929, Online. Association for Computational Linguistics.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2020. Gdpnet: Refining latent multi-view graph for relation extraction. *CoRR*, abs/2012.06780.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 87–94. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Jiali Zeng, Yongjing Yin, Yang Liu, Yubin Ge, and Jinsong Su. 2021. Domain adaptive meta-learning for dialogue state tracking. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2493–2501.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. Broad-coverage semantic parsing as transduction. 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)*, pages 3786–3798, Hong Kong, China. Association for Computational Linguistics.

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

Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. 2021. Enhancing dialogue-based relation extraction by speaker and trigger words prediction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4580–4585, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 11765–11773. AAAI Press.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better AMR-to-text generation. 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)*, pages 5459–5468, Hong Kong, China. Association for Computational Linguistics.

Victor Zue, Stephanie Seneff, Joseph Polifroni, Michael Phillips, Christine Pao, David Goddeau, James Glass, and Eric Brill. 1994. PEGASUS: A spoken language interface for on-line air travel planning. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Param. Name	Value
Batch Size	2048
Optimizer	AdamW
Learning Rate (lr)	1e-5
Lr Scheduler	linear
Warmup Step	0
Max Training Epoch	5
Semantic Masking Prob.	0.2
Extended Vocabulary Size	30,774
Max Length (dialogue)	256
Max Length (AMR)	512
Mix Precision	fp16
Parameters (Pre-training)	219M
Parameters (downstream tasks)	110M
Training Time	about 45h

Table 6: Hyper-parameters of our models.

Appendix

A Data Pre-processing

For pre-training, we randomly sample 10 million dialogue from Reddit (Henderson et al., 2019) corpus and filter the data by removing the instances where

- dialogue contains special markers;
- dialogue contains more than 10 non-English tokens;
- dialogue is longer than 150 words;
- dialogue has more than 15 turns.

We also replace the URLs in dialogues with a special token <*url*>.

B Model Input Format

Take BERT-based model as an example, given a dialogue $\boldsymbol{x}$ which consists of n utterances, we concatenate all utterances as a single consecutive token sequence with special tokens separating them: $\boldsymbol{x} = \{\texttt{[CLS]}$ $\texttt{[Utter}_1\texttt{]}$ $\texttt{Speaker}_1$ U_1 $\texttt{[Utter}_2\texttt{]}$ $\texttt{Speaker}_2$ $U_2 \ldots \texttt{[Utter}_n\texttt{]}$ $\texttt{Speaker}_n$ $U_n\texttt{[SEP]}\}$, where U_1, U_2, U_n are utterance sequences. [CLS] and [SEP] mark the start and end of the dialogue. $\texttt{[Utter}_1\texttt{]}$, $\texttt{[Utter}_2\texttt{]}$, and $\texttt{[Utter}_n\texttt{]}$ mark the utterance numbers. $\texttt{Speaker}_1$ denotes the speaker of the first utterance. For RoBERTa, we use <s> and </s> to surround the dialogue sequences.

C Model Hyper-Parameters

Table 6 lists all model hyper-parameters used for our experiments. The proposed model is implemented based on *Pytorch* and *Huggingface Transformers*[8]. Our source code and pre-trained models is released at https://github.com/goodbai-nlp/Sem-PLM.

D Architecture for Downstream Tasks

For all downstream dialogue understanding tasks, we use the pre-trained dialogue model as a dialogue encoder and make prediction based on the encoded hidden states. Taking the BERT-based model as an example, the model architecture of downstream task are:

Dialogue Relation Extraction: We concatenate the hidden states of two entities (denoted as e_1 and e_2) as well as the pooled representation of the [CLS] token into a linear classifier to predict the relation label as:

$$y = \texttt{MLP}_c([\texttt{pool}(h^{\texttt{[CLS]}}); vec(e_1); vec(e_2)]), \quad (9)$$

where $\texttt{MLP}_c$ is a linear classifier, and *vec*$(\cdot)$ selects the encoded representation of the input token. $\texttt{pool}(h^{\texttt{[CLS]}})$ passes the hidden state of the [CLS] token through a linear layer.

Intent Prediction: We solve the task as a sequence classification problem, by feeding the pooled hidden state of [CLS] token into a linear classifier to predict the relation label as:

$$y = \texttt{MLP}_c(\texttt{pool}(h^{\texttt{[CLS]}})). \quad (10)$$

Slot Filling: We represent the problem as IOB tagging, by feeding all hidden state of the input dialogue (denoted by H) into a linear classifier and predict the relation label as:

$$Y = \texttt{MLP}_c(H), \quad (11)$$

where H denotes the output hidden states, and Y is the output tag sequence.

Semantic Parsing: We solve the problem as joint sequence classification and sequence labeling task. Specifically, we predict the intent and slots label as:

$$\begin{aligned} y_{intent} &= \texttt{MLP}_{intent}(\texttt{pool}(h^{\texttt{[CLS]}})), \\ Y_{slot} &= \texttt{MLP}_{slot}(H), \end{aligned} \quad (12)$$

[8]https://github.com/huggingface/transformers

where H denotes the output hidden states, and Y is the output tag sequence.

Dialogue State Tracking: We follow the TripPy (Heck et al., 2020) framework make prediction, which uses BERT model as encoder and combines BERT with a triple copy strategy to perform state tracking. Please refer the original paper for more details.