Structural Pre-training for Dialogue Comprehension

Zhuosheng Zhang[1,2,3], Hai Zhao[1,2,3,*]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
`zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn`

Abstract

Pre-trained language models (PrLMs) have demonstrated superior performance due to their strong ability to learn universal language representations from self-supervised pre-training. However, even with the help of the powerful PrLMs, it is still challenging to effectively capture task-related knowledge from dialogue texts which are enriched by correlations among speaker-aware utterances. In this work, we present SPIDER, Structural PretraIned DialoguE Reader, to capture dialogue exclusive features. To simulate the dialogue-like features, we propose two training objectives in addition to the original LM objectives: 1) utterance order restoration, which predicts the order of the permuted utterances in dialogue context; 2) sentence backbone regularization, which regularizes the model to improve the factual correctness of summarized subject-verb-object triplets. Experimental results on widely used dialogue benchmarks verify the effectiveness of the newly introduced self-supervised tasks.

1 Introduction

Recent advances in large-scale pre-training language models (PrLMs) have achieved remarkable successes in a variety of natural language processing (NLP) tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019a; Yang et al., 2019; Clark et al., 2020; Zhang et al., 2020d). Providing fine-grained contextualized embedding, these pre-trained models are widely employed as encoders for various downstream NLP tasks. Although the PrLMs demonstrate superior perfor-

U_1: *how can i connect to URL streaming server ?*
U_2: *i think mplayer does it*
U_3: *i use kino , could be ?*
U_4: *kino is editing software . i don't think it supports media .*
U_5: *sorry EMOJI , i mean movie player*
U_6: *not sure about it . if it can't try vlc and mplayer too .*
U_7: *is mplayer aptable ?*
U_8: *yes but i 'm not sure if it is in the main repo .*
U_9: *have you ever updated your sources.list ?*
U_{10}: *i have no idea..i use adept on kubuntu .*

Figure 1: A multi-turn dialogue example. Different colors indicate the utterances from different speakers.

mance due to their strong representation ability from self-supervised pre-training, it is still challenging to effectively adapt task-related knowledge during the detailed task-specific training which is usually in a way of fine-tuning (Gururangan et al., 2020). Generally, those PrLMs handle the whole input text as a linear sequence of successive tokens and implicitly capture the contextualized representations of those tokens through self-attention. Such fine-tuning paradigm of exploiting PrLMs would be suboptimal to model dialogue task which holds exclusive text features that plain text for PrLM training may hardly embody. Therefore, we explore a fundamental way to alleviate this difficulty by improving the training of PrLM. This work devotes itself to designing the natural way of adapting the language modeling to the dialogue scenario motivated by the natural characteristics of dialogue contexts.

As an active research topic in the NLP field, multi-turn dialogue modeling has attracted great interest. The typical task is response selection (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018) that aims to select the appropriate response according to a given dialogue context containing a number of utterances, which is the focus in this work. How-

ever, selecting a coherent and informative response for a given dialogue context remains a challenge. The multi-turn dialogue typically involves two or more speakers that engage in various conversation topics, intentions, thus the utterances are rich in interactions, e.g., with criss-cross discourse structures (Li et al., 2020a; Bai and Zhao, 2018; Qin et al., 2016, 2017). A critical challenge is the learning of rich and robust context representations and interactive relationships of dialogue utterances, so that the resulting model is capable of adequately capturing the semantics of each utterance, and the relationships among all the utterances inside the dialogue.

Inspired by the effectiveness for learning universal language representations of PrLMs, there are increasing studies that employ PrLMs for conversation modeling (Mehri et al., 2019; Zhang et al., 2020b; Rothe et al., 2020; Whang et al., 2020; Han et al., 2021). These studies typically model the response selection with only the context-response matching task and overlook many potential training signals contained in dialogue data. Although the PrLMs have learned contextualized semantic representation from token-level or sentence-level pre-training tasks like MLM, NSP, they all do not consider dialogue related features like speaker role, continuity and consistency. One obvious issue of these approaches is that the relationships between utterances are harder to capture using word-level semantics. Besides, some latent features, such as user intent and conversation topic, are under-discovered in existing works (Xu et al., 2021). Therefore, the response retrieved by existing dialogue systems supervised by the conventional way still faces critical challenges, including incoherence and inconsistency.

In this work, we present SPIDER (Structural PretraIned DialoguE Reader), a structural language modeling method to capture dialogue exclusive features. Motivated to efficiently and explicitly model the coherence among utterances and the key facts in each utterance, we propose two training objectives in analogy to the original BERT-like language model (LM) training: 1) utterance order restoration (UOR), which predicts the order of the permuted utterances in dialogue context; 2) sentence backbone regularization (SBR), which regularizes the model to improve the factual correctness of summarized subject-verb-object (SVO) triplets. Experimental results on widely used benchmarks show

that SPDER boosts the model performance for various multi-turn dialogue comprehension tasks including response selection and dialogue reasoning.

2   Background and Related Work

2.1   Pre-trained Language Models

Recent works have explored various architecture choices and training objectives for large-scale LM pre-training (Zhou et al., 2020b,a; Xu et al., 2020a,b; Li et al., 2021, 2020b). Most of the PrLMs are based on the encoder in Transformer, among which Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019b) is one of the most representative work. BERT uses multiple layers of stacked Transformer Encoder to obtain contextualized representations of the language at different levels. BERT has helped achieve great performance improvement in a broad range of NLP tasks (Bai and Zhao, 2018; Zhang et al., 2020a; Luo and Zhao, 2020; Zhang et al., 2021). Several subsequent variants have been proposed to further enhance the capacity of PrLMs, such as XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020). For simplicity and convenient comparison with public studies, we select the most widely used BERT as the backbone in this work.

There are two ways of training PrLMs on dialogue scenarios, including open-domain pre-training and domain-adaptive post-training. Some studies perform training on open-domain conversational data like Reddit for response selection or generation tasks (Wolf et al., 2019; Zhang et al., 2020c; Henderson et al., 2020; Bao et al., 2020), but they are limited to the original pre-training tasks and ignore the dialogue related features. For domain-adaptive post-training, prior works have indicated that the order information would be important in the text representation, and the well-known next-sentence-prediction (Devlin et al., 2019b) and sentence-order-prediction (Lan et al., 2020) can be viewed as special cases of order prediction. Especially in the dialogue scenario, predicting the word order of utterance, as well as the utterance order in the context, has shown effectiveness in the dialogue generation task (Kumar et al., 2020; Gu et al., 2020b), where the order information is well recognized (Chen et al., 2019). However, there is little attention paid to dialogue comprehension tasks such as response selection (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018). The potential difficulty

is that utterance order restoration involves much more ordering possibilities for utterances that may have a quite flexible order inside dialogue text than NSP and SOP which only handle the predication of two-class ordering.

Our work is also profoundly related to auxiliary multi-task learning, whose common theme is to guide the language modeling Transformers with explicit knowledge and complementing objectives (Zhang et al., 2019; Sun et al., 2019b; Xu et al., 2020a). A most related work is Xu et al. (2020a), which introduces four self-supervised tasks including next session prediction, utterance restoration, incoherence detection and consistency discrimination. Our work differs from Xu et al. (2020a) by three sides. 1) Motivation: our method is designed for a general-purpose in broad dialogue comprehension tasks whose goals may be either utterance-level discourse coherence or inner-utterance factual correctness, instead of only motivated for downstream context-response matching, whose goal is to measure if two sequences are related or not. 2) Technique: we propose both sides of intra- and inter- utterance objectives. In contrast, the four objectives proposed in Xu et al. (2020a) are natural variants of NSP in BERT, which are all utterance-level. 3) Training: we empirically evaluate domain-adaptive training and multi-task learning, instead of only employing multi-task learning, which requires many efforts of optimizing coefficients in the loss functions, which would be time-consuming.

In terms of factual backbone modeling, compared with the existing studies that enhance the PrLMs by annotating named entities or incorporating external knowledge graphs (Eric et al., 2017; Liu et al., 2018), the SVO triplets extracted in our sentence backbone predication objective (SBP) method, appear more widely in the text itself. Such triplets ensure the correctness of SVO and enable our model to discover the salient facts from the lengthy texts, sensing the intuition of "who did what".

2.2 Multi-turn Dialogue Comprehension

Multi-turn dialogue comprehension aims to teach machines to read dialogue contexts and solve tasks such as response selection (Lowe et al., 2015; Wu et al., 2017; Zhang et al., 2018) and answering questions (Sun et al., 2019a; Cui et al., 2020), whose common application is building intelligent human-computer interactive systems (Chen et al., 2017a; Shum et al., 2018; Li et al., 2017; Zhu et al., 2018b). Early studies mainly focus on the matching between the dialogue context and question (Huang et al., 2019; Zhu et al., 2018a). Recently, inspired by the impressive performance of PrLMs, the mainstream is employing PrLMs to handle the whole input texts of context and question, as a linear sequence of successive tokens and implicitly capture the contextualized representations of those tokens through self-attention (Qu et al., 2019; Liu et al., 2020). Such a way of modeling would be suboptimal to capture the high-level relationships between utterances in the dialogue history. In this work, we are motivated to model the structural relationships between utterances from utterance order restoration and the factual correctness inside each utterance in the perspective of language modeling pre-training instead of heuristically stacking deeper model architectures.

3 Approach

This section presents our proposed method SPIDER (Structural Pre-traIned DialoguE Reader). First, we will present the standard dialogue comprehension model as the backbone. Then, we will introduce our designed language modeling objectives for dialogue scenarios, including utterance order restoration (UOR) and sentence backbone regularization (SBR). In terms of model training, we employ two strategies, i.e., 1) domain adaptive post-training that first trains a language model based on newly proposed objectives and then fine-tunes the response selection task; 2) multi-task fine-tuning that trains the model for downstream tasks, along with LM objectives.

3.1 Transformer Encoder

We first employ a pre-trained language model such as BERT (Devlin et al., 2019a) to obtain the initial word representations. The utterances and response are concatenated and then fed into the encoder. Given the context C and response R, we concatenate all utterances in the context and the response candidate as a single consecutive token sequence with special tokens separating them: $X = \{[\text{CLS}] R [\text{SEP}] U_1 [\text{EOU}] \ldots [\text{EOU}] U_n [\text{SEP}]\}$, where $[\text{CLS}]$ and $[\text{SEP}]$ are special tokens. $[\text{EOU}]$ is the "End Of Utterance" tag designed for multiturn context. X is then fed into the BERT encoder, which is a deep multi-layer bidirectional Transformer, to obtain a contextualized
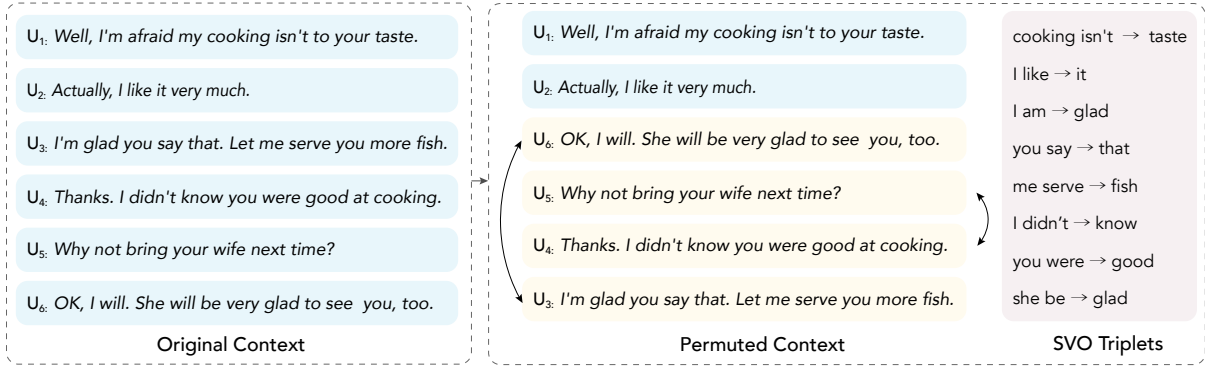
Figure 2: Structural language modeling manipulations.

representation H.

In detail, let $X = \{x_1, \ldots, x_n\}$ be the embedding of the sequence, which are features of encoding sentence words of length n. The input embeddings are then fed into the multi-head attention layer to obtain the contextual representations.

The embedding sequence X is processed to a multi-layer bidirectional Transformer for learning contextualized representations, which is defined as

$$H = \text{FFN}(\text{MultiHead}(K, Q, V)), \qquad (1)$$

where K,Q,V are packed from the input sequence representation X. As the common practice, we set $K = Q = V$ in the implementation.

For the following part, we use $H = \{h_1, \ldots, h_n\}$ to denote the last-layer hidden states of the input sequence.

3.2 SPIDER Training Objectives

To simulate the dialogue-like features, we propose two pre-training objectives in addition to the original LM objectives: 1) utterance order restoration, which predicts the order of the permuted utterances in dialogue context; 2) sentence backbone regularization, which regularizes the model to improve the factual correctness of summarized subject-verb-object triplets. The utterance manipulations are shown in Figure 2. The following subsections describe the objectives in turn.

3.2.1 Utterance Order Restoration

Coherence is an essential aspect of conversation modeling. In a coherent discourse, utterances should respect specific orders of relations and logic. The ordering of utterances in a dialogue context determines the semantic of the conversation. Therefore, learning to order a set of disordered utterances

in such a way that maximizes the discourse coherence will have a critical impact in learning the representation of dialogue contexts.

However, most previous studies focused on semantic relevance between context and response candidate. Here we introduce utterance-level position modeling, i.e., utterance order restoration to encourage the model to be aware of the semantic connections among utterances in the context. The idea is similar to autoencoding (AE) which aims to reconstruct the original data from corrupted input (Yang et al., 2019). Given permuted dialogue contexts that comprise utterances in random orders, we maximize the expected log-likelihood of a sequence of the original ground-truth order.

The goal of the utterance order restoration is to organize randomly shuffled utterances of a conversation into a coherent dialogue context. We extract the hidden states of [EOU] from H as the representation of each utterance. Formally, given an utterance sequence denoted as $C' = [H_{u_1}; H_{u_2}; \ldots; H_{u_K}]$ with order $o = [o_1; o_2; \ldots; o_K]$, where K means the number of maximum positions to be predicted. We expect an ordered context $C^* = [u_{o_1^*}; u_{o_2^*}; \ldots; u_{o_K^*}]$ is the most coherent permutation of utterances.

As predicting the permuted orders is a more challenging optimization problem than NSP and SOP tasks due to the large searching space of permutations and causes slow convergence in preliminary experiments, we choose to only predict the order of the last few permuted utterances by a permutation ratio δ to control the maximum number of permutations: $K' = K * \delta$. The UOR training objective is then formed as:

$$\mathbb{L}_{uor} = -\sum_{k=1}^{K'} [o_k \log \hat{o}_k], \qquad (2)$$

5137

where $\hat{o}_k$ denotes the predicted order.

3.2.2 Sentence Backbone Regularization

The sentence backbone regularization objective is motivated to guide the model to distinguish the internal relation of the fact triplets that are extracted from each utterance, which would be helpful to improve the ability to capture the key facts of the utterance as well as the correctness. First, we apply a fact extractor to conduct the dependency parsing of each sentence. After that, we extract the subject, the root verb, and the object tokens as an SVO triplet corresponding to each utterance. Inspired by Bordes et al. (2013) where the embedding of the tail entity should be close to the embedding of the head entity plus some vector that depends on the relationship, we assume that given the dialogue input, in the hidden representation space, the summation of the subject and the verb should be close to the object as much as possible, i.e.,

$$h_{subject} + h_{verb} \rightarrow h_{object}. \quad (3)$$

Consequently, based on the sequence hidden states h_i where $i = 1, ..., L_y$, we introduce a regularization for the extracted facts:

$$\mathbb{L}_{sbr} = \sum_{k=1}^{m}(1 - \cos(h_{subj_k} + h_{verb_k}, h_{obj_k})), \quad (4)$$

where m is the total number of fact tuples extracted from the summary and k indicates the k-th triplet. "$subj_k$", "$verb_k$", and "obj_k" are indexes of the k-th fact tuple's subject, verb, and object.

In our implementation, since PrLMs take subwords as input while the SVO extraction performs in word-level, we use the first-token hidden state as the representation of the original word following the way in Devlin et al. (2019a) for named entity recognition.

4 Use of SPIDER Objectives

In this section, we introduce two training methods to take the newly proposed language modeling objectives into account, namely domain-adaptive post-training and multi-task fine-tuning, as illustrated in Figure 3.

4.1 Domain Adaptive Post-training

Similar to BERT, we also adopt the masked language model (MLM) and the next sentence prediction (NSP) as LM-training tasks to enable our model to capture lexical and syntactic information

from tokens in text. More details of the LM training tasks can be found from Devlin et al. (2019a). The overall post-training loss is the sum of the MLM, NSP, UOR, and SBR loss.

Our full model is trained by a joint loss by combining both of the objectives above:

$$\mathbb{L} = \lambda_1(\mathbb{L}_{mlm} + \mathbb{L}_{nsp}) + \lambda_2\mathbb{L}_{uor} + \lambda_3\mathbb{L}_{sbr}, \quad (5)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters.

After post-training the language model on the dialogue corpus, we load the pre-trained weights as the same way of using BERT (Devlin et al., 2019a), to fine-tune the downstream tasks such as response selection and dialogue reasoning as focused in this work (details in Section 5.1).

4.2 Multi-task Fine-tuning

Since our objectives can well share the same input as the downstream tasks, there is an efficient way of using multi-task fine-tuning (MTF) to directly train the task-specific models along with our SPIDER objectives. Therefore, we feed the permuted context to the dialogue comprehension model and combine the three losses for training:

$$\mathbb{L} = \beta_1\mathbb{L}_{dm} + \beta_2\mathbb{L}_{uor} + \beta_3\mathbb{L}_{sbr}, \quad (6)$$

where $\beta_1, \beta_2, \beta_3$ are hyper-parameters.

In order to train a task-specific model for dialogue comprehension, the hidden states H will be fed into a classifier with a fully connected and softmax layer. We learn model $g(\cdot, \cdot)$ by minimizing cross entropy loss with dataset $\mathcal{D}$. Let Θ denote the parameters, for binary classification like the response selection task, the objective function $\mathcal{L}(\mathcal{D}, \Theta)$ can be formulated as:

$$\mathbb{L}_{dm} = -\sum_{i=1}^{N}[y_i \log(g(c_i, r_i)) + (1 - y_i) \log(1 - g(c_i, r_i))].$$

where N denotes the number of examples. For multiple choice task like MuTual, the loss function is:

$$\mathbb{L}_{dm} = -\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,c} \log(g(c_i, r_{i,k})).$$
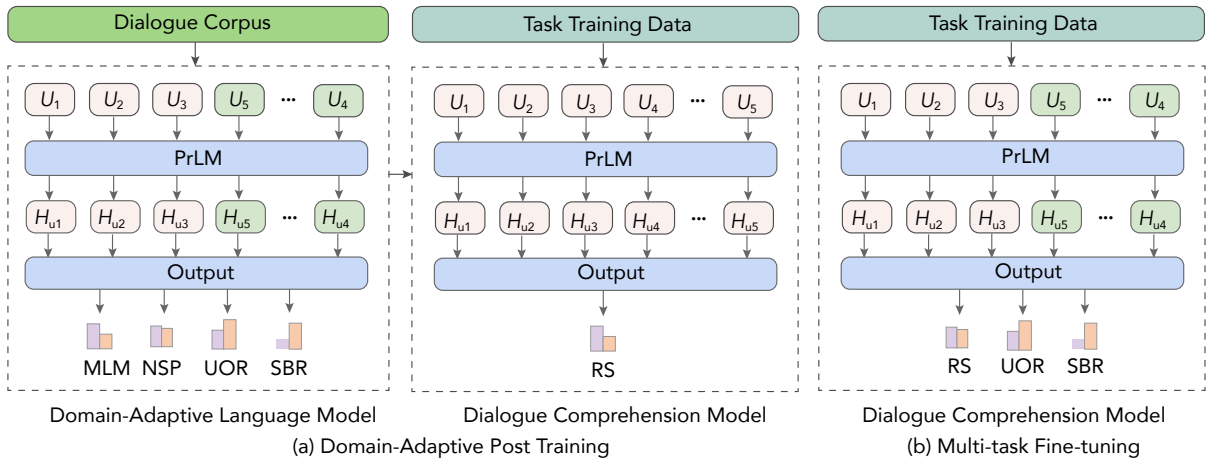
where C is the number of choice.

Figure 3: Model training flow of domain-adaptive post-training (a) and multi-task fine-tuning (b).

| | Ubuntu | | | Douban | | | ECD | | |
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
# context-response pairs	1M	500K	500K	1M	50K	10K	1M	10K	10K
# candidates per context	2	10	10	2	2	10	2	2	10
Avg # turns per context	10.13	10.11	10.11	6.69	6.75	6.45	5.51	5.48	5.64
Avg # words per utterance	11.35	11.34	11.37	18.56	18.50	20.74	7.02	6.99	7.11

Table 1: Data statistics of Ubuntu, Douban, and ECD datasets.

5 Experiments

5.1 Datasets

We evaluated our model on two English datasets: Ubuntu Dialogue Corpus (Ubuntu) (Lowe et al., 2015) and Multi-Turn Dialogue Reasoning (MuTual) (Cui et al., 2020),[1] and two Chinese datasets: Douban Conversation Corpus (Douban) (Wu et al., 2017) and E-commerce Dialogue Corpus (ECD) (Zhang et al., 2018).

5.1.1 Ubuntu Dialogue Corpus

Ubuntu (Lowe et al., 2015) consists of English multi-turn conversations about technical support collected from chat logs of the Ubuntu forum. The dataset contains 1 million context-response pairs, 0.5 million for validation and 0.5 million for testing. In training set, each context has one positive response generated by human and one negative response sampled randomly. In validation and test sets, for each context, there are 9 negative responses and 1 positive response.

5.1.2 Douban Conversation Corpus

Douban (Wu et al., 2017) is different from Ubuntu in the following ways. First, it is an open domain where dialogues are extracted from Douban Group. Second, response candidates on the test set are collected by using the last turn as the query to retrieve 10 response candidates and labeled by humans. Third, there could be more than one correct response for a context.

5.1.3 E-commerce Dialogue Corpus

ECD (Zhang et al., 2018) dataset is extracted from conversations between customer and service staff on Taobao. It contains over 5 types of conversations based on over 20 commodities. There are also 1 million context-response pairs in the training set, 0.5 million in the validation set, and 0.5 million in the test set.

5.1.4 Multi-Turn Dialogue Reasoning

MuTual (Cui et al., 2020) consists of 8860 manually annotated dialogues based on Chinese student English listening comprehension exams. For each context, there is one positive response and three negative responses. The difference compared to the above three datasets is that only MuTual is reasoning-based. There are more than 6 types of reasoning abilities reflected in MuTual.

---

[1] Actually, MuTual is a retrieval-based dialogue corpus in form, but the theme is English listening comprehension exams, thus we regard as a reading comprehension corpus in this work. Because the test set of MuTual is not publicly available, we conducted the comparison with our baselines on the Dev set for convenience.

Model	Ubuntu Corpus			Douban Conversation Corpus						E-commerce Corpus		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
SMN	72.6	84.7	96.1	52.9	56.9	39.7	23.3	39.6	72.4	45.3	65.4	88.6
DUA	75.2	86.8	96.2	55.1	59.9	42.1	24.3	42.1	78.0	50.1	70.0	92.1
DAM	76.7	87.4	96.9	55.0	60.1	42.7	25.4	41.0	75.7	-	-	-
IoI	79.6	89.4	97.4	57.3	62.1	44.4	26.9	45.1	78.6	-	-	-
MSN	80.0	89.9	97.8	58.7	63.2	47.0	29.5	45.2	78.8	60.6	77.0	93.7
MRFN	78.6	88.6	97.6	57.1	61.7	44.8	27.6	43.5	78.3	-	-	-
SA-BERT	85.5	92.8	98.3	**61.9**	**65.9**	49.6	31.3	48.1	**84.7**	70.4	**87.9**	98.5
Multi-task Fine-tuning												
BERT	81.7	90.4	97.7	58.8	63.1	45.3	27.7	46.4	81.8	61.7	81.1	97.0
+ SPIDER	83.1	91.3	98.0	59.8	63.8	45.9	28.5	48.7	82.6	62.6	82.7	97.1
Domain Adaptive Post-training												
BERT	85.7	93.0	98.5	60.5	64.7	47.4	29.1	47.8	84.9	66.4	84.8	97.6
+ SPIDER	**86.9**	**93.8**	**98.7**	60.9	65.0	47.5	29.6	**48.8**	83.6	**70.8**	85.3	**98.6**

Table 2: Performance comparison on Ubuntu, Douban and E-Commerce datasets.

5.2 Implementation Details

For the sake of computational efficiency, the maximum number of utterances is specialized as 20. The concatenated context, response, `[CLS]` and `[SEP]` in one sample is truncated according to the "longest first" rule or padded to a certain length, which is 256 for MuTual and 384 for the other three datasets. For the hyper-parameters, we empirically set $\lambda_1 = \lambda_2 = \lambda_3 = \beta_1 = \beta_2 = 1$.

Our model is implemented using Pytorch and based on the Transformer Library.[2] We use BERT (Devlin et al., 2019a) as our backbone model. AdamW (Loshchilov and Hutter, 2019) is used as our optimizer. The batch size is 24 for MuTual, and 64 for others. The initial learning rate is 4×10^{-6} for MuTual and 3×10^{-5} for others. The ratio is set to 0.4 in our implementation by default. We run 3 epochs for MuTual and 2 epochs for others and select the model that achieves the best result in validation. The training epochs are 3 for DAP.

Our domain adaptive post-training for the corresponding response selection tasks is based on the three large-scale dialogue corpus including Ubuntu, Douban, and ECD, respectively.[3] The data statistics are in Table 1. Since domain adaptive post-training is time-consuming, following previous studies (Gu et al., 2020a), we use *bert-base-uncased*, and *bert-base-chinese* for the English and

Chinese datasets, respectively. Because there is no appropriate domain data for the small-scale Mutual dataset, we only report the multi-task fine-tuning results with our SPIDER objectives, and also present the results with other PrLMs such as ELECTRA (Clark et al., 2020) for general comparison.

5.3 Baseline Models

We include the following models for comparison:

• **Multi-turn matching models**: Sequential Matching Network (SMN) (Wu et al., 2017), Deep Attention Matching Network (DAM) (Zhou et al., 2018), Deep Utterance Aggregation (DUA) (Zhang et al., 2018), Interaction-over-Interaction (IoI) (Tao et al., 2019b) have been stated in Section 2.2. Besides, Multi-Representation Fusion Network (MRFN) (Tao et al., 2019a) matches context and response with multiple types of representations. Multi-hop Selector Network (MSN) (Yuan et al., 2019) utilizes a multi-hop selector to filter necessary utterances and matches among them.

• **PrLMs-based models**: BERT (Devlin et al., 2019b), SA-BERT (Gu et al., 2020a), and ELECTRA (Clark et al., 2020).

5.4 Evaluation Metrics

Following (Lowe et al., 2015; Wu et al., 2017), we calculate the proportion of true positive response among the top-k selected responses from the list of n available candidates for one context, denoted as $R_n@k$. Besides, additional conventional metrics of information retrieval are employed on Douban: Mean Average Precision (MAP) (Baeza-Yates et al.,

---

[2] Our source code is available at https://github.com/cooelf/SPIDER.

[3] Since phrases are quite common in Chinese, making it inaccurate to calculate the SVO relations according to Eq. 3, thus we did not use the SBR objective for the two Chinese tasks in this work.

Model	MRR	$R_4@1$	$R_4@2$
$BERT_{base}$	80.0	65.3	86.0
+ UOR	80.7	66.1	86.7
+ SBR	81.3	67.4	87.1
+ SPIDER	81.6	67.6	87.3
$BERT_{large}$	82.2	69.1	87.9
+ UOR	82.8	69.8	88.6
+ SBR	83.4	71.0	89.4
+ SPIDER	83.9	71.8	89.2
$ELECTRA_{base}$	86.5	76.2	91.6
+ UOR	86.9	76.6	91.8
+ SBR	87.6	77.1	92.0
+ SPIDER	88.2	79.2	92.3
$ELECTRA_{large}$	94.9	90.6	97.7
+ UOR	95.3	91.3	97.8
+ SBR	95.5	91.6	97.8
+ SPIDER	**95.6**	**92.0**	**97.9**

Table 3: Results on MuTual dataset.

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
$SPIDER_{DAP}$	**86.9**	**93.8**	**98.7**
w/o UOR	86.2	93.3	98.6
w/o SBR	86.4	93.5	98.6
w/o Both	85.7	93.0	98.5
$SPIDER_{MTF}$	83.1	91.3	98.0
w/o UOR	82.6	91.0	97.9
w/o SBR	82.3	90.8	97.8
w/o Both	81.7	90.4	97.7

Table 4: Ablation study on the Ubuntu dataset.

1999), Mean Reciprocal Rank (MRR) (Voorhees et al., 1999), and precision at position 1 (P@1).

5.5 Results

Tables 2-3 show the results on the four benchmark datasets. We have the following observations:

1) Generally, the previous models based on multi-turn matching networks perform worse than simple PrLMs-based ones, illustrating the power of contextualized representations in context-sensitive dialogue modeling. PrLM can perform even better when equipped with our SPIDER objectives, verifying the effectiveness of dialogue-aware language modeling, where inter-utterance position information and inner-utterance key facts are better exploited. Compared with SA-BERT that involves more complex architecture and more parameters by injecting extra speaker-aware embeddings, SPIDER keeps the same model size as the backbone BERT, and even surpasses SA-BERT on most of the metrics.

2) In terms of the training methods, DAP generally works better than MTF, with the merits of two-step procedures including the pure LM-based post-training. According to the ablation study in Table 4, we see that both of the dialogue-aware LM objectives are essentially effective and combining them (SPIDER) gives the best performance, which verifies the necessity of modeling the utterance order and factual correctness. We also notice that UOR shows better performance than SBR in DAP, while gives relative descent in MFT. The most plau-

sible reason would be that UOR would permute the utterances in the dialogue context which helps the language model learn the utterance in UOR. However, in MFT, the major objective is the downstream dialogue comprehension task. The permutation of the context would possibly bring some negative effects to the downstream task training.

5.6 Influence of Permutation Ratio

For the UOR objective, a hyper-parameter δ is set to control the maximum number of permutations (as described in Section 3.2.1), which would possibly influence the overall model performance. To investigate the effect, we set the permutation ratio from [0, 20%, 40%, 60%, 80%, 100%]. The result is depicted in Figure 4, in which our model outperforms the baseline in general, showing that the permutation indeed strengthens the baseline.

5.7 Comparison with Different Context Length

Context length can be measured by the number of turns and average utterance length in a conversation respectively. We split test instances from the Ubuntu dataset into several buckets and compare SPIDER with UOR with the BERT baseline. According to the results depicted in Figure 5, we observe that SPIDER performs much better on contexts with long utterances, and it also performs robustly and is significantly and consistently superior to the baseline. The results indicate the benefits of modeling the utterance order for dialogue comprehension.

5.8 Human Evaluation about Factual Correctness

To compare the improvements of SPIDER over the baseline on factual correctness, we extract the error cases of the BERT baseline on MuTual (102 in total) and 42 (41.2%) are correctly answered
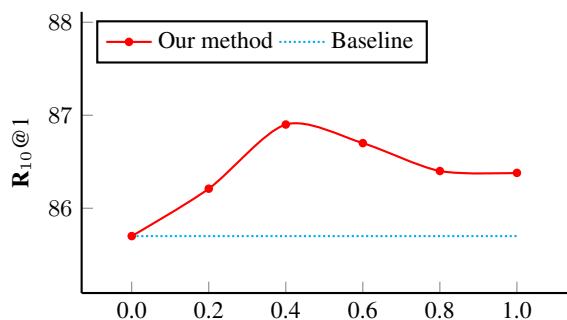
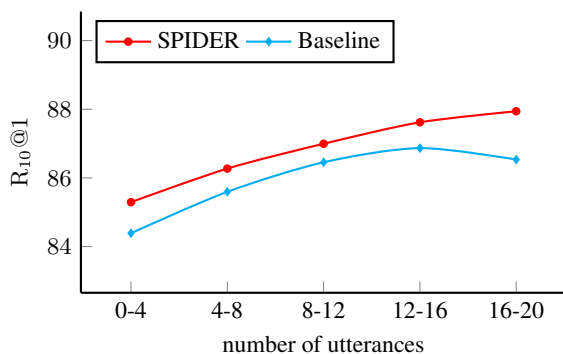Figure 4: Influence of the permutation ratio δ.

Figure 5: $R_{10}@1$ of SPIDER and the baseline BERT on different numbers of utterances.

by SPIDER. Among the 42 solved cases, 33/42 (78.6%) are entailed with SVO facts in contexts, indicating the benefits of factual correctness.

6 Conclusion

In this paper, we focus on the task-related adaptation of the pre-trained language models and propose SPIDER (Structural Pre-traIned DialoguE Reader), a structural language modeling method to capture dialogue exclusive features. To explicitly model the coherence among utterances and the key facts in each utterance, we introduce two novel dialogue-aware language modeling tasks including utterance order restoration and sentence backbone regularization objectives. Experiments on widely-used multi-turn dialogue comprehension benchmark datasets show the superiority over baseline methods. Our work reveals a way to make better use of the structure learning of the contextualized representations from pre-trained language models and gives insights on how to adapt the language modeling training objectives in downstream tasks.

References

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463.

ACM press New York.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa Fe, New Mexico, USA. 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.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

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

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Neural machine translation with reordering embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1787–1799, Florence, Italy. Association for Computational Linguistics.

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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020a. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020b. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *arXiv preprint arXiv:2012.01775*.

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.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. 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 8115–8122. AAAI Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2495–2498.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020a. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020b. Task-specific objectives of pre-trained language models for dialogue adaptation. *arXiv preprint arXiv:2009.04984*.

Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2021. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chuang Liu, Deyi Xiong, Yuxiang Jia, Hongying Zan, and Changjian Hu. 2020. Hisbert for conversational reading comprehension. In *2020 International Conference on Asian Language Processing (IALP)*, pages 147–152. IEEE.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Ying Luo and Hai Zhao. 2020. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418, Online. Association for Computational Linguistics.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Matthew 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.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270, Austin, Texas. Association for Computational Linguistics.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1133–1136. ACM.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. ERNIE: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019a. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 267–275. ACM.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2020. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. *arXiv preprint arXiv:2009.04703*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2020a. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *arXiv preprint arXiv:2009.06265*.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Nan Duan, and Daxin Jiang. 2020b. Syntax-enhanced pre-trained model. *arXiv preprint arXiv:2012.14116*.

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.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. 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 111–120, Hong Kong, China. Association for Computational Linguistics.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020a. DCMN+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*, volume 34, pages 9563–9570.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. 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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. 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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

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*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020d. Semantics-aware BERT 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 9628–9635. AAAI Press.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Junru Zhou, Zuchao Li, and Hai Zhao. 2020a. Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020b. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018a. SDNet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. 2018b. Lingke: a fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 108–112, Santa Fe, New Mexico. Association for Computational Linguistics.