Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues

Ruijian Xu[1], **Chongyang Tao**[2], **Daxin Jiang**[2], **Xueliang Zhao**[1], **Dongyan Zhao**[1], **Rui Yan**[1,3]
[1] Peking University, Beijing, China
[2] Microsoft Corporation, Beijing, China
[3] Beijing Academy of Artificial Intelligence, Beijing, China
{xurj,xl.zhao,zhaody,ruiyan}@pku.edu.cn,{chongyang.tao,djiang}@microsoft.com

Abstract

Building an intelligent dialogue system with the ability to select a proper response according to a multi-turn context is a great challenging task. Existing studies focus on building a context-response matching model with various neural architectures or pretrained language models (PLMs) and typically learning with a single response prediction task. These approaches overlook many potential training signals contained in dialogue data, which might be beneficial for context understanding and produce better features for response prediction. Besides, the response retrieved from existing dialogue systems supervised by the conventional way still faces some critical challenges, including incoherence and inconsistency. To address these issues, in this paper, we propose learning a context-response matching model with auxiliary self-supervised tasks designed for the dialogue data based on pre-trained language models. Specifically, we introduce four self-supervised tasks including next session prediction, utterance restoration, incoherence detection and consistency discrimination, and jointly train the PLM-based response selection model with these auxiliary tasks in a multi-task manner. By this means, the auxiliary tasks can guide the learning of the matching model to achieve a better local optimum and select a more proper response. Experiment results on two benchmarks indicate that the proposed auxiliary self-supervised tasks bring significant improvement for multi-turn response selection in retrieval-based dialogues, and our model achieves new state-of-the-art results on both datasets.

Introduction

Building a dialogue system that can converse with people naturally and meaningfully is one of the most challenging problems towards high-level artificial intelligence, and has been drawing increasing interests from both academia and industry area. Most existing dialogue systems are either generation-based (Vinyals and Le 2015; Serban et al. 2016) or retrieval-based (Wang et al. 2013; Lowe et al. 2015; Wu et al. 2017; Tao et al. 2019a). Given the dialogue context, generation-based approaches synthesize a response word by word with a conditional language model, while retrieval-based methods select a proper response from a candidate pool. In this paper, we focus on retrieval-based approaches that are superior in providing informative responses and have been widely applied in many commercial products.

We consider the response selection task in multi-turn dialogues, where the retrieval model ought to select a proper response by measuring the matching degree between a multi-turn dialogue context and a number of response candidates. Earlier studies (Wang et al. 2013; Lowe et al. 2015) concatenate the context to a single utterance and calculate the matching score with the utterance-level representations. Later, most response selection models (Zhou et al. 2016; Wu et al. 2017; Zhang et al. 2018) perform context-response matching within the representation-matching-aggregation paradigm, where each turn of utterance is represented individually and sequential information is aggregated among a sequence of utterance-response matching features. To further improve the performance, some recent approaches consider multiple granularities (or layers) of representations (Zhou et al. 2018; Tao et al. 2019a) for matching or propose more complicated interaction mechanisms between the context and the response (Tao et al. 2019b).

Recently, a wide range of studies have shown that PLMs, such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), on the large corpus can learn universal language representations, which are helpful for various downstream natural language processing tasks and can get rid of training a new model from scratch. To adapt pre-trained models for multi-turn response selection, Whang et al. (2020) and Gu et al. (2020) make the first attempt to utilize BERT (Devlin et al. 2019) to learn a matching model, where context and the candidate response are first concatenated and then fed into the PLMs for calculating the final matching score. These pre-trained language models can well capture the interaction information among inter-utterance and intra-utterance through multiple transformer layers. Although PLM-based response selection models demonstrate superior performance due to its strong representation ability, it is still challenging to effectively learn task-related knowledge during the training process, especially when the size of training corpora is limited. Naturally, these studies typically learn the response selection model with only the context-response matching task and overlook many potential training signals contained in dialogue data. Such training signals might be beneficial for context understanding and produce better features for response prediction. Besides, the response retrieved by existing dialogue systems supervised by the conventional way still faces some critical challenges, including incoherence and inconsistency.

On account of the above issues, in this paper, instead of configuring complex context-response matching models, we propose learning the context-response matching model with auxiliary self-supervised tasks designed for dialogue data based on pre-trained language models (e.g., BERT). Specifically, we introduce four self-supervised tasks including *next session prediction*, *utterance restoration*, *incoherence detection* and *consistency discrimination*, and jointly train the PLM-based response selection model with these auxiliary tasks in a multi-task manner. On the one hand, these auxiliary tasks help improve the capability of the response selection model to understand the dialogue context and measure the semantic relevance, consistency or coherent between the context and the response candidates. On the other hand, they can guide the matching model to effectively learn task-related knowledge with a fixed amount of train corpora and produce better features for response prediction.

We conduct experiments on two benchmark data sets for multi-turn response selection: the Ubuntu Dialog Corpus (Lowe et al. 2015) and the E-commerce Dialogue Corpus (Zhang et al. 2018). Evaluation results show that our proposed approach is significantly better than all state-of-the-art models on both datasets. Compared with the previous state-of-the-art methods, our model achieves 2.9% absolute improvement in terms of $R_{10}@1$ for the Ubuntu dataset and 4.8% absolute improvement for the E-commerce dataset. Furthermore, we applied our proposed self-supervised learning schema to some non-PLM-based response selection models, e.g., dual LSTM (Lowe et al. 2015) and ESIM (Chen and Wang 2019). Experimental results indicate that our learning schema can also bring consistent and significant improvement to the performance of the existing matching models. Surprisingly, with self-supervised learning, a simple ESIM even performs better than BERT on the ubuntu dataset, demonstrating that our approach is beneficial for various matching architectures.

In summary, our contributions are three-fold:

- We propose learning a context-response matching model with auxiliary self-supervised tasks to fully utilize training signals in the multi-turn dialogue context.
- We design four self-supervised tasks to enhance the capability of a PLM-based response prediction model in capturing the semantic relevance, coherence or consistency.
- We achieve new state-of-the-art results on two benchmark datasets. Besides, with the help of auxiliary self-supervised tasks, a simple ESIM model can even achieve better performance than BERT on the Ubuntu dataset.

Model

Task Formalization

Suppose that there is a multi-turn dialogue dataset $\mathcal{D} = \{c_i, r_i, y_i\}_{i=1}^N$, where $c_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,m_i}\}$ denotes a dialogue context with $u_{i,t}$ representing the utterance of the t-th turn, r_i denotes a response candidate, and $y_i \in \{0, 1\}$ denotes a label with $y_i = 1$ indicating that r_i is a proper response for c_i (otherwise, $y_i = 0$). The task is to learn a matching model $g(\cdot, \cdot)$ from $\mathcal{D}$ so that for any new context $c = \{u_1, u_2, \ldots, u_m\}$ and a response candidate r, $g(c, r) \in [0, 1]$ can measure the matching degree between c and r.

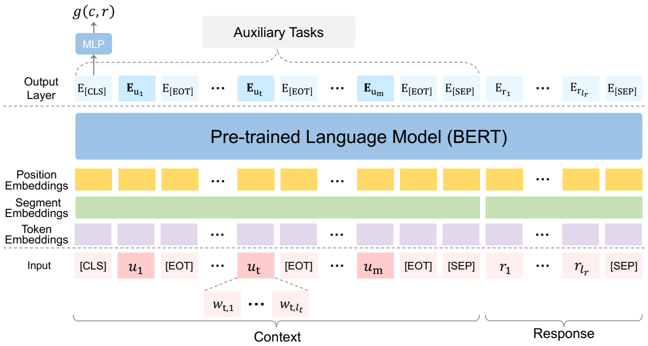


Figure 1: Overall architecture of our model.

Matching with PLMs

We consider building the context-response matching model with the PLMs, as it is trained on large amounts of unlabelled data and provides strong "universal representations" that can be finetuned on task-specific training data to achieve good performance on downstream tasks. Consistent with previous studies (Gu et al. 2020; Whang et al. 2020), we select BERT as the base model for a fair comparison.

Specifically, given a context $c = \{u_1, u_2, \ldots, u_m\}$, where the t-th utterance $u_t = \{w_{t,1}, \ldots, w_{t,l_t}\}$ is a sequence with l_t words, a response candidate $r = \{r_1, r_2, \ldots, r_{l_r}\}$ consisting of l_r words and a label $y \in \{0, 1\}$, we first concatenate all utterances in the context and the response candidate as a single consecutive token sequence with special tokens separating them, which can be formulated as $x = \{\texttt{[CLS]}, u_1, \texttt{[EOT]}, u_2, \texttt{[EOT]}, \ldots, \texttt{[EOT]}, u_m, \texttt{[EOT]}, \texttt{[SEP]}, r, \texttt{[SEP]}\}$. Here $\texttt{[CLS]}$ and $\texttt{[SEP]}$ are the classification symbol and the segment separation symbol of BERT, $\texttt{[EOT]}$ is the "End Of Turn" tag designed for multi-turn context. For each word of x, *token*, *position* and *segment* embeddings of x are summated and fed into pre-trained transformer layer (a.k.a. BERT), giving us the contextualized embedding sequence $\{E_{\texttt{[CLS]}}, E_2, \ldots, E_{l_x}\}$. $E_{\texttt{[CLS]}}$ is an aggregated representation vector that contains the semantic interaction information for the context-response pair. We then fed $E_{\texttt{[CLS]}}$ into a multi-perception layer to obtain the final matching score for the context-response pair:

$$g(c, r) = \sigma(W_2 \cdot f(W_1 E_{\texttt{[CLS]}} + b_1) + b_2) \tag{1}$$

where $W_{\{1,2\}}$ and $b_{\{1,2\}}$ are trainable parameters for response prediction task, $f(\cdot)$ is a $\texttt{tanh}$ activation function, $\sigma(\cdot)$ stands a sigmoid function.

Finally, cross-entropy loss function is utilized as the training objective of the *context-response matching* task:

$$\mathcal{L}_{\texttt{crm}} = -y \log(g(c, r)) - (1 - y) \log(1 - g(c, r)) \tag{2}$$

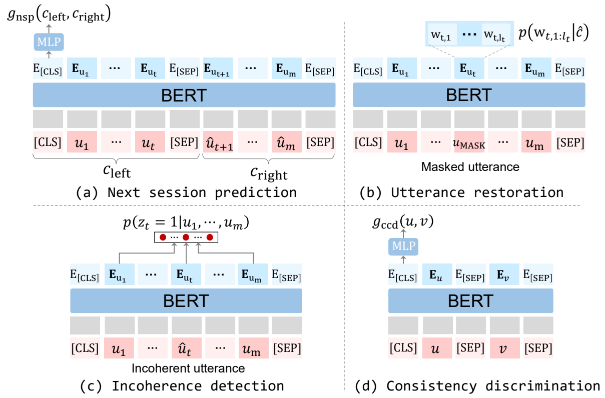

$g_{nsp}(c_{left}, c_{right})$
MLP
$E_{[CLS]}$ E_{u_1} ⋯ E_{u_t} $E_{[SEP]}$ $E_{u_{t+1}}$ ⋯ E_{u_m} $E_{[SEP]}$
BERT
[CLS] u_1 ⋯ u_t [SEP] $\hat{u}_{t+1}$ ⋯ $\hat{u}_m$ [SEP]
c_{left}
c_{right}
(a) Next session prediction
$w_{t,1}$ ⋯ w_{t,l_t} $p(w_{t,1:l_t}|\hat{c})$
$E_{[CLS]}$ E_{u_1} ⋯ E_{u_t} ⋯ E_{u_m} $E_{[SEP]}$
BERT
[CLS] u_1 ⋯ u_{MASK} ⋯ u_m [SEP]
Masked utterance
(b) Utterance restoration
$p(z_t = 1|u_1, \cdots, u_m)$
$E_{[CLS]}$ E_{u_1} ⋯ E_{u_t} ⋯ E_{u_m} $E_{[SEP]}$
BERT
[CLS] u_1 ⋯ $\hat{u}_t$ ⋯ u_m [SEP]
Incoherent utterance
(c) Incoherence detection
$g_{ccd}(u, v)$
MLP
$E_{[CLS]}$ E_u $E_{[SEP]}$ E_v $E_{[SEP]}$
BERT
[CLS] u [SEP] v [SEP]
(d) Consistency discrimination

Figure 2: Sketches of four types of self-supervised tasks. Gray square stands for various embeddings for each token.

Before the fine-tuning procedure with the above context-response matching task, for a fair comparison, we follow previous studies (Whang et al. 2020; Gu et al. 2020; Gururangan et al. 2020) and carry out domain-adaptive post-training to incorporate in-domain knowledge into BERT. In the rest of this section, we will introduce our proposed four auxiliary self-supervised tasks, and then present the final learning objective of our model.

Self-Supervised Tasks

Heading for a matching model that can effectively learn domain knowledge with a fixed amount of training corpora and produce better features for response prediction, we design four auxiliary self-supervised tasks, i.e. *session-level matching*, *utterance restoration*, *incoherence detection* and *consistency classification*. These self-supervised tasks try to enhance the capability of the model to measure the semantic relevance, coherent, and consistency between the context and the response candidate. On the other hand, they can also guide the learning of the model to achieve a better local optimum. Figure 2 illustrates the sketches of four types of self-supervised tasks.

Next Session Prediction Due to the natural sequential relationship between dialogue turns, the latter turns usually show a strong semantic relevance with the previous turns in the context. Inspired by such a characteristic, we design a more general response prediction task with the dialogue context, name *next session prediction* (NSP), to fully utilize the sequential relationship of the dialogue data and enhance the capability of the model to measure the semantic relevance. Specifically, the next session prediction task requires the model to predict whether two sequences are consecutive and relevant. However, instead of matching a context with a response utterance, the model needs to calculate the matching degree between two pieces of dialogue session.

Formally, given a context $c = \{u_1, u_2, \ldots, u_m\}$, we randomly[1] split c into two consecutive pieces $c_{\texttt{left}} = \{u_1, \ldots, u_t\}$ and $c_{\texttt{right}} = \{u_{t+1}, \ldots, u_m\}$. Then, with a 50% chance, we replace $c_{\texttt{left}}$ or $c_{\texttt{right}}$ with a piece of context sampled from the whole training corpus[2]. If one of the two piece is replaced, we set the label $y_{\texttt{nsp}} = 0$, otherwise $y_{\texttt{nsp}} = 1$. The next session prediction task requires the model to discriminate whether $c_{\texttt{left}}$ and $c_{\texttt{right}}$ can form a consecutive context.

To train PLMs with the proposed self-supervised task, we first concatenate all utterances of each piece as a single sequence with [EOT] appended to the end of each utterance. Similar to the main task, we fed two segments into BERT encoder and obtain the aggregated representation of the piece pair $E^{\texttt{nsp}}_{\texttt{[CLS]}}$. We further compute the final matching score $g_{\texttt{nsp}}(c_{\texttt{left}}, c_{\texttt{right}})$ with a non-linear transformation. Finally, the objective function of context alignment task can be formulated as

$$\begin{aligned}\mathcal{L}_{\texttt{nsp}} &= -y_{\texttt{nsp}} \log(g_{\texttt{nsp}}(c_{\texttt{left}}, c_{\texttt{right}})) \\ &\quad - (1 - y_{\texttt{nsp}}) \log(1 - g_{\texttt{nsp}}(c_{\texttt{left}}, c_{\texttt{right}}))\end{aligned} \tag{3}$$

Utterance Restoration As one of the common self-supervised tasks in PLMs, token-level masked language modeling is usually utilized to guide the model to learn semantic and syntactic features of word sequences with the bidirectional context. Here we further introduce utterance-level masked language modeling, i.e. *utterance restoration* (UR) task to encourage the model to be aware of the semantic connections among utterances in the context. Specifically, we mask all the tokens in an utterance randomly sampled from the dialogue session and let the model restore it with the information from the rest context. By learning to predict a proper utterance that fits its surrounding dialogue context, the model can produce better representations that can well adapt to dialogues, similar to the idea of continuous bag-of-words model (Mikolov et al. 2013).

Formally, given a context $c = \{u_1, u_2, \ldots, u_m\}$, we randomly select an utterance u_t and replace all tokens in the utterance with a special token [MASK]. The model is required to restore u_t based on $\hat{c} = \{u_1, \ldots, u_{t-1}, u_{\text{mask}}, u_{t+1}, \ldots, u_m\}$. To adapt the task in BERT, we formulate input of BERT encoder as $x_{\texttt{ur}} = \{\texttt{[CLS]}, u_1, \texttt{[EOT]}, \ldots, u_{\texttt{mask}}, \texttt{[EOT]}, \ldots, u_m, \texttt{[EOT]}, \texttt{[SEP]}\}$, where $u_{\texttt{mask}}$ consists of only [MASK] tokens and has the same length with u_t. After being processed by BERT, the top layer output a representation sequence $E_{\texttt{ur}} = \{E_{\texttt{[CLS]}}, E_{1,1}, \ldots, E_{1,l_1}, E_{\texttt{[EOT]}}, \ldots, E_{m,1}, \ldots, E_{m,l_m}, E_{\texttt{[EOT]}}, E_{\texttt{[SEP]}}\}$, where l_t is the length of the t-th utterance. The model predict the masked utterance conditioned on the contextualized representations of each word. The probability distribution of each masked word can be calculated as

$$\begin{aligned}E'_{t,j} &= \text{GLEU}(W_{ur} E_{t,j} + b_{ur}) \\ p(w_{t,j}|\hat{c}) &= \text{softmax}\big(W'_{ur} E'_{t,j} + b'_{ur}\big)\end{aligned} \tag{4}$$

where W_{ur}, W'_{ur}, b_{ur}, b'_{ur} are trainable parameters, $w_{t,j}$ is the j-th token of the t-th utterance, and GLEU$(\cdot)$ is an activation function. Then, the training objective of *utterance*

[1] In this work, all random sampling operations are carried out according to uniformly distribution.

[2] If $c_{\texttt{left}}$ is replaced, the new piece should be the left part of another context with a random length, and vice versa.

restoration task is to minimize the following negative log-likelihood (NLL):

$$\mathcal{L}_{\mathtt{ur}} = -\frac{1}{l_t}\sum_{j=1}^{l_t}\log p(w_{t,j}|\hat{c}) \tag{5}$$

Incoherence Detection Inspired by the concept of discourse coherence (Jurafsky 2000) in linguistics, we further introduce the *incoherence detection* (ID) task which requires the model to recognize the incoherent utterance within a dialogue session, so as to enhance the capability of a model on capturing the sequential relationship among utterances and selecting coherent response candidates. Specifically, given a dialogue context $c = \{u_1, \ldots, u_m\}$, we randomly select one of the utterances $u_k \in \{u_1, \ldots, u_m\}$ and replace it with an utterance randomly sampled from the whole training corpus. Then, the model should find the incoherent utterance among the context. For each sample, we define a one-hot label $\{z_1, \ldots, z_m\}$, where $z_t = 1$ if $t = k$, indicating that the t-th utterance is been replaced, otherwise $z_t = 0$.

To model this task, BERT encoder takes an input $x_{\mathtt{id}} = \{\mathtt{[CLS]}, u_1, \mathtt{[EOT]}, \ldots, u_m, \mathtt{[EOT]}, \mathtt{[SEP]}\}$ and outputs $E_{\mathtt{id}} = \{E_{\mathtt{[EOT]}}, E_{1,1}, \ldots, E_{m,l_m}, E_{\mathtt{[SEP]}}\}$, where $E_{t,j}$ denotes the contextualized embedding of the j-th word in the k-th utterance and l_t is the length of t-th utterance. We calculate the aggregated representation of the k-th utterance by fusing the mean and max value of the embedding sequence $\{E_{t,1}, \ldots, E_{t,l_t}\}$, which can be formulated as

$$U_t = \big[\frac{1}{l_t}\sum_{j=1}^{l_t} E_{t,j}; \max_{1\le j\le l_t} E_{t,j}\big] \tag{6}$$

Then, the model makes a prediction based on the aggregated representations of each utterance, the probability of the t-th utterance being replaced is

$$\begin{aligned} p(z_t = 1|u_1, \ldots, u_m) &= \text{softmax}(W_{\mathtt{id}}U_t + b_{\mathtt{id}}) \\ &= \frac{\exp(W_{\mathtt{id}}U_t + b_{\mathtt{id}})}{\sum_{s=1}^{m}\big(\exp(W_{\mathtt{id}}U_s) + b_{\mathtt{id}}\big)} \end{aligned} \tag{7}$$

where $W_{\mathtt{id}}$ and $b_{\mathtt{id}}$ are trainable parameters.

Finally, the learning objective of inconsistency detection task is defined as

$$\mathcal{L}_{\mathtt{id}} = -\sum_{t=1}^{m} z_t \log p(z_t = 1|u_1, \ldots, u_m) \tag{8}$$

Consistency Discrimination Selecting responses that are consistent with the dialogue context is one of the major challenges in building engaging conversational agents. However, most previous studies focused on modeling the semantic relevance between the context and the response candidate. Intuitively, utterances from the same dialogue session tend to share similar topics, and utterances from the same interlocutor tend to share the same personality or style. According to the characteristics, we propose a *consistency discrimination* (CD) task to enhance the ability of a response prediction model to measure the consistency among dialogue utterances with a self-supervised discriminative training schema.

Formally, given a dialog context $c = \{u_1, u_2, \ldots, u_m\}$, we sample two utterances from the same interlocutor[3], and denote them as u and v respectively. Then, we randomly sample an utterance $\tilde{v}$ from another context in the training corpus. The model is required to measure the consistency degree of $\langle u, v\rangle$ and $\langle u, \tilde{v}\rangle$ and give a higher score to $\langle u, v\rangle$. Since u and v are not consecutive in the dialogue context, the model is forced to capture the features about the consistency (such as topic, personality and style) between two sequences, rather than coherence and semantic relevance.

To calculate the consistency score of a sequence pair $\langle u, v\rangle$, we first concatenate the two utterances as $x_{\mathtt{cd}} = \{\mathtt{[CLS]}, u, \mathtt{[SEP]}, v, \mathtt{[SEP]}\}$, and then fed the sequence into BERT. As described in previous tasks, BERT returns an aggregated representation $E^{\mathtt{cd}}_{\mathtt{[CLS]}}$. Then, the consistency score $g_{\mathtt{cd}}(u, v)$ is computed with a non-linear transformation over $E^{\mathtt{cd}}_{\mathtt{[CLS]}}$. Likewise, we can obtain the consistency score of $\langle u, \tilde{v}\rangle$, i.e. $g_{\mathtt{cd}}(u, \tilde{v})$. Finally, we would like $g_{\mathtt{cd}}(u, v)$ to be larger than $g_{\mathtt{cd}}(u, \tilde{v})$ by at least a margin Δ and define the learning objective as a hing loss function:

$$\mathcal{L}_{\mathtt{cd}} = \max\{0, \Delta - g_{\mathtt{cd}}(u, v) + g_{\mathtt{cd}}(u, \tilde{v})\} \tag{9}$$

Learning Objective

We adopt a multi-task learning manner and define the final objective function as:

$$\begin{aligned} \mathcal{L}_{\mathtt{final}} &= \mathcal{L}_{\mathtt{crm}} + \alpha\mathcal{L}_{\mathtt{self}} \\ \mathcal{L}_{\mathtt{self}} &= \mathcal{L}_{\mathtt{nsp}} + \mathcal{L}_{\mathtt{ur}} + \mathcal{L}_{\mathtt{id}} + \mathcal{L}_{\mathtt{cd}} \end{aligned} \tag{10}$$

where α is a hyper-parameter as a trade-off between the objective of the main task and those of the auxiliary tasks.

Experiments

Datasets and Evaluation Metrics

we evaluate the proposed method on two benchmark datasets for multi-turn dialogue response selection. The first dataset is the ***Ubuntu Dialogue Corpus (v1.0)*** (Lowe et al. 2015), which consists of multi-turn English dialogues about technical support and is collected from chat logs of the Ubuntu forum. We use the copy shared by Gu et al. (2020), in which numbers, paths and URLs are replaced by placeholders. The Ubuntu dataset contains 1 million context-response pairs for training, and 0.5 million pairs for validation and test. The ratio of positive candidates and negative candidates is $1:1$ in the training set, and $1:9$ in the validation set and the test set. The second dataset is the ***E-commerce Dialogue Corpus*** (Zhang et al. 2018), which consists of real-world multi-turn dialogues between customers and customer service staff on Taobao[4], the largest e-commerce platform in China. The E-commerce dataset contains 1 million context-response pairs for training, and 10 thousand pairs for validation and test. The ratio of positive candidates and negative candidates is $1:1$ in the training set and the validation set, and $1:9$ in the test set.

[3] We assume that utterances in a dialogue context are posed one by one, therefore we can simply sample utterances from only the odd turns or even turns.

[4] https://www.taobao.com

Metrics		Ubuntu Corpus				E-commerce Corpus		
Models		R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
Non-PLM-based Models	DualLSTM (Lowe et al. 2015)	0.901	0.638	0.784	0.949	0.365	0.536	0.828
	Multi-View (Zhou et al. 2016)	0.908	0.662	0.801	0.951	0.421	0.601	0.861
	SMN (Wu et al. 2017)	0.926	0.726	0.847	0.961	0.453	0.654	0.886
	DUA (Zhang et al. 2018)	-	0.752	0.868	0.962	0.501	0.700	0.921
	DAM (Zhou et al. 2018)	0.938	0.767	0.874	0.969	0.526	0.727	0.933
	MRFN (Tao et al. 2019a)	0.945	0.786	0.886	0.976	-	-	-
	IMN (Gu, Ling, and Liu 2019)	0.946	0.794	0.889	0.974	0.621	0.797	0.964
	ESIM (Chen and Wang 2019)	0.950	0.796	0.874	0.975	0.570	0.767	0.948
	IoI (Tao et al. 2019b)	0.947	0.796	0.894	0.974	0.563	0.768	0.950
	MSN (Yuan et al. 2019)	-	0.800	0.899	0.978	0.606	0.770	0.937
PLM-based Models	BERT (Whang et al. 2020)	0.952	0.814	0.902	0.977	0.631	0.826	0.964
	SA-BERT (Gu et al. 2020)	0.965	0.855	0.928	0.983	0.704	0.879	0.985
	BERT-VFT (Whang et al. 2020)	-	0.855	0.928	0.985	-	-	-
	BERT-VFT (Ours)	0.969	0.867	0.939	0.987	0.717	0.884	0.986
	BERT-SL	**0.975***	**0.884***	**0.946***	**0.990***	**0.776***	**0.919***	**0.991**
	BERT-SL w/o. NSP	0.973	0.879	0.944	0.989	0.760	0.914	0.988
	BERT-SL w/o. UR	0.974	0.881	0.945	0.990	0.763	0.916	0.991
	BERT-SL w/o. ID	0.972	0.877	0.942	0.989	0.755	0.911	0.987
	BERT-SL w/o. CD	0.973	0.880	0.945	0.989	0.742	0.897	0.986

Table 1: Evaluation results on the two data sets. Numbers marked with $*$ mean that the improvement is statistically significant compared with the baseline (t-test with p-value < 0.05). Numbers in bold indicate the best strategies for the corresponding models on specific metrics.

Following Lowe et al. (2015) and Zhang et al. (2018), we employ $R_n@k$s as evaluation metrics, where $R_n@k$ denotes recall at position k in n candidates and measures the probability of the positive response being ranked in top k positions among n candidates.

Baseline Models

We compared BERT-SL with the following models:

DualLSTM (Lowe et al. 2015): the model concatenates all utterances in the context to form a single sequence and calculates a matching score based on the representations produced by an LSTM.

Multi-View (Zhou et al. 2016): the model measures the matching degree between the context and the response candidate in both a word view and an utterance view.

SMN (Wu et al. 2017): the model lets each utterance in the context interacts with the response candidate, and the matching vectors of all utterance-response pairs are aggregated with an RNN to calculate a final matching score.

DUA (Zhang et al. 2018): the model formulates previous utterances into context using a deep utterance aggregation model, and performs context-response similar to SMN.

DAM (Zhou et al. 2018): the model is similar to SMN, but utterances in the context and the response candidate are represented with stacked self-attention and cross-attention layers. The matching vectors are aggregated with a 3-D CNN.

MRFN (Tao et al. 2019a): the model employs multiple types of representations for context-response interaction, where each type encodes semantics of units from a kind of granularity or dependency among the units.

ESIM (Chen and Wang 2019): the model first concatenates all utterances in the context into a single sequence, and then employs ESIM structure derived from NLI for context-response matching.

IMN (Gu, Ling, and Liu 2019): following Wu et al. (2017), the model enhances the representations at both the word- and sentence-level and collects matching information of utterance-response pairs bidirectionally.

IoI (Tao et al. 2019b): the model lets the context-response matching process goes deep along the interaction block chain via representations in an iterative fashion.

MSN (Yuan et al. 2019): the model utilizes a multi-hop selector to select the relevant utterances in context and then matches the filtered context with the response candidate to obtain a matching score.

BERT (Whang et al. 2020): the model fine-tunes the BERT with the concatenation of the context and the response candidates as the input.

BERT-VFT (Whang et al. 2020): before fine-tuning, the model also carries out a post-training on training corpora in the same manner as BERT.

SA-BERT (Gu et al. 2020): the model follows BERT-VFT, and further incorporates speaker-aware embeddings.

Implementation Details

Following Gu et al. (2020), we select English uncased $\text{BERT}_{\text{base}}$ (110M) as the context-response matching model for the Ubuntu dataset and Chinese $\text{BERT}_{\text{base}}$ model for the E-commerce dataset. We implement the models with the code in https://github.com/huggingface/transformers. The maximum lengths of the context and response were set to 448 and 64 as the maximum length of input sequence in BERT is 512. Intuitively, the last tokens in the context and the previous tokens in the response candidate are more important, so we cut off the previous tokens for the context but do the cut-off in the reverse direction for the response candidate if the sequences are longer than the maximum length. We choose 32 as the size of mini-batches for training. On both the Ubuntu dataset and the E-commerce dataset, we

Metrics	Ubuntu Corpus				E-Commerce Corpus		
Models	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
DualLSTM (Lowe et al. 2015)	0.901	0.638	0.784	0.949	0.365	0.536	0.828
DualLSTM-SL	0.925*	0.724*	0.858*	0.969*	0.518*	0.722*	0.933*
ESIM (Chen and Wang 2019)	0.950	0.796	0.874	0.975	0.570	0.767	0.948
ESIM-SL	0.963*	0.822*	0.909*	0.980*	0.623*	0.797*	0.969*

Table 2: Evaluation results of two matching models trained with the proposed self-supervised tasks. Numbers marked with $*$ mean that the improvement is statistically significant compared with the baseline (t-test with p-value < 0.05).

applied domain adaptive post-training before the finetuning procedure following the settings of Whang et al. (2020). Training instances of auxiliary tasks are generated dynamically. We set $\Delta = 0.6$ (Equation (9)) and choose $\alpha = 1.0$ (Equation (10)) as the trade-off between the learning objectives. The model is optimized using Adam optimizer with a learning rate set as $3e - 5$. Early stopping on the validation data is adopted as a regularization strategy.

Evaluation Results

Table 1 reports the results of BERT-SL and all baseline models on the Ubuntu datasets and the E-commerce dataset. From the evaluation results, we can easily observe that the PLM-based response selection models generally perform better than the models based on various neural architectures. The phenomenon shows the advantage of the pre-trained models on providing strong universal representations for response selection. Among those PLM-based response selection models, our BERT-SL outperforms the best baseline BERT-VFT in terms of all metrics on both data sets. Specifically, compared to the previous state-of-the-art model, our BERT-SL achieves 2.9% absolute improvement in terms of $\mathrm{R}_{10}@1$ on the Ubuntu dataset and 4.8% absolute improvement on the E-commerce dataset. We conduct statistical tests, and the results indicate that the improvement on all metrics except $\mathrm{R}_{10}@5$ on the E-commerce data is statistically significant. The significant improvement demonstrates the effectiveness of our proposed self-supervised learning schema. Notably, our method does not increase the inference time compared with existing PLM-based models.

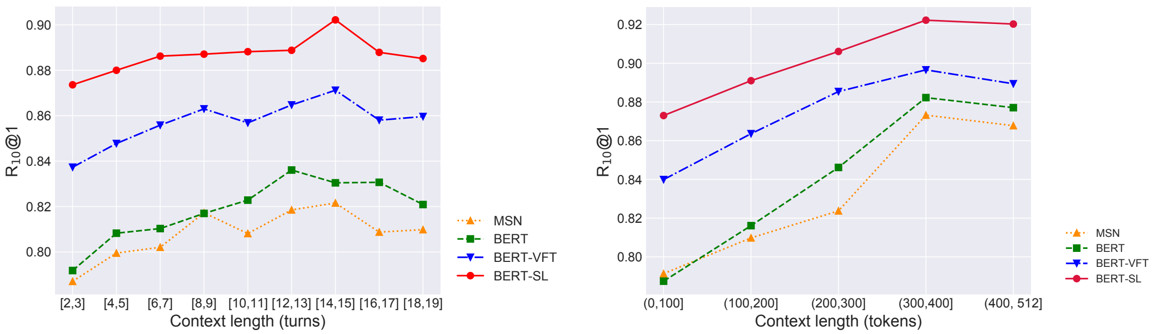
Discussions

Ablation study. To investigate the impact of different self-supervised tasks, we conducted a comprehensive ablation study. We keep the architecture of the matching model and remove each self-supervised task individually from the model, and denote the model as "BERT-SL w/o. $\mathcal{T}$", where $\mathcal{T} \in \{$NSP, UR, ID, CD$\}$ stand for next session prediction, utterance restoration, incoherence detection and consistency discrimination respectively. The detailed results are reported in the last four lines of Table 1. First of all, we find that all four self-supervised tasks are useful as removing any of them causes a performance drop on both datasets. Second, we can conclude that on the Ubuntu data, the rank of the tasks in terms of $R_{10}@1$ is that ID $>$ NSP $>$ CD $>$ UR; and on the E-commerce data, the rank of the tasks is that CD $>$ ID $>$ NSP $>$ UR[5]. Among the four tasks, ID plays an important role in improving the response selection task. The reason might be that the ID task can encourage the model to consider the coherence between the context and a response candidate, which acts as complementary to the main task. It is also noted that removing the UR task leads to the slightest decrease of the performance of response selection on both datasets, as the feature learned by UR may be redundant with that learned by the token-level mask language modeling in pre-training. Besides, the representation learned by the generative task might have a considerable discrepancy with the discrimination task. Finally, the CD task is much more important on the E-commerce data than it is on the Ubuntu data, as E-commerce corpora contain more diverse content.

Self-supervised learning for ESIM/DualLSTM. We are curious about whether the effectiveness of the proposed self-supervised learning schema depends on the architecture of the response selection model. Therefore, we test our proposed learning schema on some non-PLM-based response selection models, such as dual LSTM (Lowe et al. 2015) and ESIM (Chen and Wang 2019). The original two models treat the multi-turn context as a long sequence and are trained with only the context-response task. Thus, we implement two models and jointly train them with the proposed four self-supervised tasks in a multi-task manner. Table 2 reports the comparison results on both data sets. We observe a consistent and significant improvement of the performance for both DualLSTM and ESIM. Particularly, with the help of auxiliary self-supervised tasks, a simple ESIM model can even achieve better performance on the Ubuntu dataset than BERT, which is a bigger pre-trained model. The results imply that our learning schema is beneficial for various matching architectures, and indicate the effectiveness and generality of the proposed method.

Performance across different lengths of context. To analyze how the performance of our proposed BERT-SL varies with different context lengths, we compare BERT-SL with BERT, BERT-VFT and the state-of-the-art non-PLM-based response selection models (a.k.a. MSN). In this work, context length is measured by (1) number of turns and (2) number of all tokens in a context. Figure 3 shows how the performance of the four models varies across contexts with different lengths. We can observe that the performance of all models first increases monotonically when the context length increases, and then fluctuates or even drops when

[5] We select $R_{10}@1$ as target metrics in the study of the importance of different tasks because they are more critical than other metrics in real systems of response selection.



(a) Average turns vs. R$_{10}$@1

(b) Total context length vs. R$_{10}$@1

Figure 3: Performance of BERT-SL and its variants across different lengths of contexts. (a) context length is measured by the average number of turns; (b) context length is measured by the total length of the context.

context length keeps increasing. The reason might be that when only a few utterances are available in the context, the model could not capture enough information for matching, but when the context becomes long enough, noises will be brought to matching due to the topic shift in dialogue. Across the different lengths of the context, our BERT-SL can always achieve better performance than BERT-VFT as well as other baselines. It is worth noting that the performance of our BERT-SL is more stable than other models across different turns of the context, and drops more slightly than other models for a long context. The results imply that our learning schema improves the capability of the matching model to deal with long contexts or short context.

Related Works

With the advance of natural language processing, building an intelligent dialogue system with data-driven approaches (Vinyals and Le 2015; Lowe et al. 2015) has drawn increasing interests in recent years. Most existing approaches are either generation-based (Vinyals and Le 2015; Serban et al. 2016) or retrieval-based (Wang et al. 2013; Wu et al. 2017; Whang et al. 2020). We focus on retrieval-based methods in this paper. Earlier studies pay attention to constructing single-turn context-response matching models where only a single utterance is considered or multiple utterances in the context are concatenated into a long sequence for response selection (Wang et al. 2013; Hu et al. 2014; Lowe et al. 2015). Recently, most studies focus on the multi-turn scenario where each utterance in the context first interacts with the response candidate, and then the matching features are aggregated according to the sequential dependencies of the multi-turn context (Zhou et al. 2016; Wu et al. 2017; Yan, Song, and Wu 2016; Tao et al. 2019a; Zhou et al. 2018; Tao et al. 2019b), and they usually adopt the *representation-matching-aggregation* paradigm to build the matching models. Following the paradigm, Yuan et al. (2019) introduce a multi-hop selector to select the relevant utterances in the context for response matching.

Recently, pre-trained language models (Devlin et al. 2019; Yang et al. 2019; Liu et al. 2019) have shown impressive benefits for various downstream NLP tasks, and some researchers tried to apply them on response selection. Vig and Ramea (2019) utilizes BERT to represent each utterance-response pair and aggregate these representations to calculate the matching score. Whang et al. (2020) treat the context as a long sequence and perform context-response matching with the BERT. Besides, the model also introduces the next utterance prediction and mask language modeling tasks borrowed from BERT during the post-training on dialogue corpus to incorporate in-domain knowledge for the matching model. Following Whang et al. (2020), Gu et al. (2020) propose to heuristically incorporate speaker-aware embeddings into BERT to promote the capability of context understanding in multi-turn dialogues.

Self-supervised learning has become a significant direction in the AI community and has contributed to the success of pre-trained language models (Devlin et al. 2019; Liu et al. 2019; Yang et al. 2019). Inspired by this, some researchers propose to learn down-stream tasks with auxiliary self-supervised tasks either in pre-training manner or multi-task manner. Existing works have explored self-supervised tasks in text classification (Yu and Jiang 2016), summarization (Wang et al. 2019) and utterance generation (Zhang et al. 2019; Zhao, Xu, and Wu 2020; Wu, Wang, and Wang 2019). Besides, Mehri et al. (2019) pre-train the hierarchical context encoder with four self-supervised learning objectives respectively and transfer it to other downstream tasks. Different from previous works, we design several self-supervised tasks according to the characteristics of the dialogue data to improve the multi-turn response selection.

Conclusion

In this paper, we propose learning a context-response matching model with four auxiliary self-supervised tasks designed for the dialogue data. Jointly trained with these auxiliary tasks, the matching model can effectively learn task-related knowledge contained in dialogue data and produce better features for response selection. Experiment results on two benchmarks indicate that the proposed auxiliary self-supervised tasks bring significant improvement for multi-turn response selection in retrieval-based dialogues, and our model achieves new state-of-the-art results on both datasets.

Acknowledgments

We would like to thank the reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106600), the National Science Foundation of China (NSFC No. 61876196 and No. 61672058). Chongyang Tao and Dongyan Zhao are the corresponding authors. Rui Yan is supported as a young fellow at Beijing Academy of Artificial Intelligence (BAAI).

References

Chen, Q.; and Wang, W. 2019. Sequential matching model for end-to-end multi-turn response selection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7350–7354. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Gu, J.-C.; Li, T.; Liu, Q.; Ling, Z.-H.; Su, Z.; Wei, S.; and Zhu, X. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20. ACM. ISBN 978-1-4503-6859-9.

Gu, J.-C.; Ling, Z.-H.; and Liu, Q. 2019. Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, 2321–2324. ACM. ISBN 978-1-4503-6976-3.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 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*, 8342–8360. Online: Association for Computational Linguistics.

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, 2042–2050.

Jurafsky, D. 2000. *Speech & language processing*. Pearson Education India.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* .

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 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*, 285–294. Prague, Czech Republic: Association for Computational Linguistics.

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

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 3776–3783. AAAI Press.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 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*, 267–275.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 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*, 1–11. Florence, Italy: Association for Computational Linguistics.

Vig, J.; and Ramea, K. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Vinyals, O.; and Le, Q. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869* .

Wang, H.; Lu, Z.; Li, H.; and Chen, E. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 935–945. Seattle, Washington, USA: Association for Computational Linguistics.

Wang, H.; Wang, X.; Xiong, W.; Yu, M.; Guo, X.; Chang, S.; and Wang, W. Y. 2019. Self-Supervised Learning for Contextualized Extractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2221–2227. Florence, Italy: Association for Computational Linguistics.

Whang, T.; Lee, D.; Lee, C.; Yang, K.; Oh, D.; and Lim, H. 2020. An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In *Proc. Interspeech 2020*.

Wu, J.; Wang, X.; and Wang, W. Y. 2019. Self-Supervised Dialogue Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3857–3867. Florence, Italy: Association for Computational Linguistics.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 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), 496–505. Vancouver, Canada: Association for Computational Linguistics.

Yan, R.; Song, Y.; and Wu, H. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 55–64.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Yu, J.; and Jiang, J. 2016. Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 236–246. Austin, Texas: Association for Computational Linguistics.

Yuan, C.; Zhou, W.; Li, M.; Lv, S.; Zhu, F.; Han, J.; and Hu, S. 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)*, 111–120. Hong Kong, China: Association for Computational Linguistics.

Zhang, Y.; Gao, X.; Lee, S.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759* .

Zhang, Z.; Li, J.; Zhu, P.; Zhao, H.; and Liu, G. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3740–3752. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Zhao, Y.; Xu, C.; and Wu, W. 2020. Learning a Simple and Effective Model for Multi-turn Response Generation with Auxiliary Tasks. *arXiv preprint arXiv:2004.01972* .

Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; and Yan, R. 2016. Multi-view Response Selection for Human-Computer Conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 372–381. Austin, Texas: Association for Computational Linguistics.

Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W. X.; Yu, D.; and Wu, H. 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)*, 1118–1127. Melbourne, Australia: Association for Computational Linguistics.