Multi-turn Dialogue Reading Comprehension with Pivot Turns and Knowledge

Zhuosheng Zhang, Junlong Li, Hai Zhao

***Abstract*—Multi-turn dialogue reading comprehension aims to teach machines to read dialogue contexts and solve tasks such as response selection and answering questions. The major challenges involve noisy history contexts and especial prerequisites of commonsense knowledge that is unseen in the given material. Existing works mainly focus on context and response matching approaches. This work thus makes the first attempt to tackle the above two challenges by extracting substantially important turns as pivot utterances and utilizing external knowledge to enhance the representation of context. We propose a pivot-oriented deep selection model (PoDS) on top of the Transformer-based language models for dialogue comprehension. In detail, our model first picks out the pivot utterances from the conversation history according to the semantic matching with the candidate response or question, if any. Besides, knowledge items related to the dialogue context are extracted from a knowledge graph as external knowledge. Then, the pivot utterances and the external knowledge are combined with a well-designed mechanism for refining predictions. Experimental results on four dialogue comprehension benchmark tasks show that our proposed model achieves great improvements on baselines. A series of empirical comparisons are conducted to show how our selection strategies and the extra knowledge injection influence the results.**

***Index Terms*—Multi-turn Dialogue Comprehension, Response Selection, Utterance Selection, Commonsense Modeling.**

I. INTRODUCTION

Multi-turn dialogue reading comprehension aims to teach machines to read dialogue contexts and solve tasks such as response selection [26, 44, 54] and answering questions [35], whose common application is building intelligent human-computer interactive systems [3, 18, 33, 62]. The task of response selection requires the model to select the appropriate response from a set of candidates given the context of a conversation [26, 44, 54]. The widely-used benchmark datasets for response selection are the English Ubuntu Dialogue Corpus (Ubuntu) [26] and two Chinese datasets, namely the Douban Conversation Corpus (Douban) [44] and E-commerce Dialogue Corpus (ECD) [54]. The concerned task has evolved from single-turn matching where only the last utterance in context is used for matching a reply [14, 26, 41, 42] to multi-turn modeling that benefits from a richer multi-turn context [9, 12, 19, 22, 26, 36, 44, 54, 61, 62]. More recently, the task has further been extended to a multi-choice dialogue machine reading comprehension (MRC) style setting [31, 38] with extra questions—given a dialogue context and corresponding questions, the machine is required to select the appropriate answer accordingly [6, 35].

TABLE I
MULTI-TURN DIALOGUES HAVE DIFFERENT TOPICS (IN BOLD).

Dialogue 1
W: ***Well, I'm afraid my cooking isn't to your taste.***
M: ***Actually, I like it very much.***
W: I'm glad you say that. Let me serve you some more fish.
M: ***Thanks. I didn't know you are so good at cooking.***
W: Why not bring your wife next time?
M: OK, I will. She will be very glad to see you, too.
Question: *What does the man think of the woman's cooking?*
A. It's really terrible.
B. It's very good indeed. *
C. It's better than what he does.

The solution architectures for both of the dialogue response selection and dialogue comprehension tasks are similar, and existing models generally consist of two components: encoder and matching network. Most previous works focused on the matching of the context and response [11, 44, 61] where matching signals in each utterance–response pair are fetched from their interaction according to their representations and then aggregated as a matching score. Benefit from recent advance of pre-trained language models (PrLMs), a critical feature of current models [13, 63] is that they choose advanced pre-trained language models (PrLMs) such as BERT [7] and GPT [30] as encoder implementation.

Although the dialogue comprehension shares a similar form with the standard reading comprehension tasks [31, 38, 55, 58, 59], understanding multi-turn dialogues is much more complex for two key challenges. On the one hand, multi-turn dialogues are multi-party, multi-topic, and always have lots of turns in real-world cases (e.g., chat history in social media) [10, 46, 54]. For the scenario of multi-turn dialogue with over 20 turns of utterances, we argue that not all the utterances contribute to the final response selection. The noise in context utterances seriously hurt the matching performance [49, 54]. Therefore, matching the response with all the utterances would be suboptimal. On the other hand, people rarely state obvious commonsense explicitly in their dialogues [8], therefore only considering superficial contexts may ignore implicit meaning and cause misunderstanding.

This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU long term AI project, Cutting-edge Machine reading comprehension and language model (Corresponding author: Hai Zhao).

Zhuosheng Zhang, Junlong Li, and Hai Zhao are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, and also with Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, and also with MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University. (e-mail: zhangzs@sjtu.edu.cn; lockonn@sjtu.edu.cn; zhaohai@cs.sjtu.edu.cn).

TABLE II
COMMONSENSE IS REQUIRED IN THE QUESTION.

Dialogue 2
M: Look at the girl on the bike!
F: Oh, yes she's really a smart girl.
Question: *Where are the two persons?*
A. At home.
B. In their classroom.
*C. On the street. **

We demonstrate two examples from DREAM dataset [35] in Tables I and II. Table I shows the topic shifts in turns, and only a few of them, called pivot turns, are directly related to the question. Treating all turns equally hurts the understanding of multi-turn dialogues as shown in some previous works [50, 54]. Table II shows that commonsense knowledge is required to answer the question (e.g., bikes are always on the streets). However, such required knowledge usually cannot be obtained or inferred from the given material (passage, question, or answer options) of the task.

To refine the context with topic clues, our prior work [54] took the last utterance in the context as the pivot to mine the connections with the rest preceding utterances.[1] However, the indicative utterance is not always the last one. Besides, there would be more than one pivot utterance in a conversation and there is a lack of focus on commonsense knowledge. In this work, we propose a pivot-oriented deep selection model (PoDS) for dialogue comprehension. In detail, our model first picks out the pivot utterances from the conversation history according to the semantic matching with the candidate response or question, if any. Besides, knowledge items related to the dialogue context are extracted from a knowledge graph as external knowledge. Then, the pivot utterances and the external knowledge are combined together with a well-designed mechanism for giving predictions. Experimental results on four dialogue comprehension benchmark tasks show that our proposed model achieves great improvements on baselines. Our contributions are three folds:

1) We propose a flexible and explainable pivot-oriented deep selection model to locate the informative utterance(s) as the pivot clues to refine the context with topic clues and apply a fine-grained matching with the candidate response.

2) For more advanced dialogue comprehension that requires commonsense reasoning, we introduce external knowledge from a knowledge graph to enrich its representation.

3) Experimental results on four benchmark corpora show that the PoDS achieves substantial improvements over the baselines. A series of empirical comparisons are conducted to show how our selection strategies and the extra knowledge injection influence the results.

II. RELATED WORK

A. Pre-trained Language Model

Recently, deep contextualized language models (PrLMs) have been shown to be effective in learning universal language representations, achieving state-of-the-art results in a series of flagship natural language processing tasks. Prominent examples are Embedding from Language Models (ELMo) [29], Generative Pre-trained Transformer (OpenAI GPT) [30], BERT [7], Generalized Autoregressive Pre-training (XLNet) [48], Robustly Optimized BERT Pretraining approach (Roberta) [24], and ALBERT [17]. Providing fine-grained contextualized embedding, these pre-trained models can be either easily applied to downstream models as the encoder or used for fine-tuning.

Despite their impressive success, these PrLMs remain limited in representing the contextualized information in the domain-specific corpus because they are usually trained on a general corpus [43, 45, 56, 57]. Besides, multi-turn conversation modeling is more challenging, which requires deep interaction between the abundant context and response. The simple linear layer used in the PrLMs for downstream task prediction would not be sufficient enough. We need a better strategy of taking advantage of both sides of the pairwise modeling in PrLMs and more effective interaction to infer the relationship between the candidate response and the conversation history. One solution is to conduct post-training on task-specific datasets. Another is task-adapted model design, which is the major focus of this work. In the present paper, we extend the deep contextualized PrLMs into multi-turn dialogue modeling with a well-crafted model design.

B. Multi-turn Dialogue Modeling

Developing a dialogue system means training machines to converse with a human using natural language. Towards this end, a number of data-driven dialogue systems have been designed [26, 44, 54], and they can be categorized into two types of architectures: one concatenates all context utterances [26, 27] and the other separates and then aggregates utterances [36, 44, 61]. Existing works showed that a matching structure is beneficial for improving the connections between sequences in neural network models [20, 26, 27, 42, 51, 60]. Recent work has extended attention to modeling multi-turn response selection. DUA [54] employed self-matching attention to route the vital information in each utterance. DAM [61] proposed a method based entirely on attention achieving impressive improvement. IOI [36] improved context-to-response matching by stacking multiple interaction blocks. State-of-the-art methods show that capturing and leveraging matched information at different granularities across context and response are crucial to multi-turn response selection [11, 37].

In contrast with the above studies, the present work benefits from a deep pre-trained Transformer architecture with deep context interaction to model the relationships between context and response. Our proposed PoDS has three differences from models in the studies mentioned above. (1) The PoDS adopts a deep Transformer encoder as the backbone with dual pairwise modeling between the context and response, while the models of previous studies are based on RNN, and the inputs are encoded separately. (2) The PoDS adopts a *pack and separate* strategy to take advantage of both deep Transformer encoders and further response-aware interaction. (3) The PoDS employs more carefully selective matching between the context

---

[1] In this work, we define *pivot* as the intermediate utterance(s) used for refining context as the topic or evidence clues.

utterances and response when calculating attention weights in the interactive matching module. A series of empirical comparisons are conducted to show the influence factors.

C. Knowledge Enhanced Language Representation

Recently, researchers pay more and more attention to enhancing text models with Knowledge Graphs (KGs), since KGs obtain a great amount of systematic knowledge. Integrating background knowledge in a neural model was first proposed in the neural-checklist model by Kiddon et al. [16] for text generation of recipes. Liu et al. [23] combined knowledge triples in KGs with original texts before modeling them with BERT to get more hidden information. Mihaylov and Frank [28] attended to relevant external knowledge and combined this knowledge with the context representation in Cloze-style reading comprehension. Bosselut et al. [1] employed the triples in KGs as corpus to train GPT [30] for commonsense learning. Lin et al. [21] proposed a knowledge-aware graph network based on GCN and LSTM with a path-based attention mechanism. Zhang et al. [53] fused entity information with BERT to enhance language representation, which can take advantage of lexical, syntactic, and knowledge information simultaneously.

Some previous works have already taken either key turns [50] or commonsense knowledge [35] into account when dealing with multi-turn dialogues. However, none of them make a combination of these two important factors. In this work, we make the first attempt to utilize key turns as pivots and commonsense knowledge simultaneously to enhance the language representation of multi-turn dialogues. Pivot utterances are selected from the conversation history according to the semantic matching with the candidate response, whose representations are directly extracted from the encoded text of the original dialogue context. On the other hand, knowledge items related to the dialogue context are extracted from a knowledge graph as external knowledge, which are then encoded with the same PrLM encoder. As a result, representations of pivot utterances, commonsense knowledge, and the original text of multi-turn dialogues are in the same vector space so that they can be easily fused together.

III. PIVOT-ORIENTED DEEP SELECTION MODEL

In the model section, we first introduce our backbone framework for the widely used dialogue response task, such as the Ubuntu [26], Douban [44], and ECD [54] task, whose form is to select the proper response from a candidate list, given the context history composed from various dialogue utterances. Then, we extend the framework to recent conversational reading comprehension, whose major difference is having extra questions, thus the aim is to select the correct answer from various candidate options, given the dialogue context and the corresponding questions, as examples shown in Tables I-II.

Figure 1 shows our PoDS framework for conventional multi-turn conversation tasks, supposing that the symbol in dark-marked $\mathbf{H}_1^u$ is the selected pivot. The PoDS formalizes the context and response into a joint input to feed the Transformer encoder [39] and mines key information from the utterances and response across the encoding. The PoDS then separates the context and response, semantically matches each utterance, and response to select the pivot utterances. The pivot utterances are used to refine the original conversation context to obtain the pivot-aware contextual representation. The matching module then calculates the attentive interaction between the refined context and candidate response. In the last module, the response-aware contextual vectors are delivered to a Gated Recurrent Unit (GRU) [4] in chronological order of the utterances in the context, and the last hidden state is passed to a linear layer to obtain the final matching score.

We denote the training set as a triple $\mathcal{D} = \{(\mathbf{C}, \mathbf{R}, \mathbf{Y})_i\}_{i=1}^N$, where $\mathbf{C} = \{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ is a conversation context with $\{\mathbf{U}_k\}_{k=1}^n$ as the utterances. $\mathbf{R}$ is a response candidate while $\mathbf{Y} \in \{0, 1\}$ is a binary label, indicating whether $\mathbf{R}$ is a proper response for $\mathbf{C}$.

A. Encoding

To make the best use of the Transformer-based deep encoders, we employ a *pack and separate* method by first packing the context and response as a joint input to feed the encoder and then separate them according to the positions for further interaction.[2]

Given the context $\mathbf{C}$ and response $\mathbf{R}$, tokens are packed into a sequence:

$$\mathbf{X} = \{\texttt{[CLS]}\,\mathbf{R}\,\texttt{[SEP]}\,\mathbf{U}_1\,\texttt{[SEP]}\ldots\texttt{[SEP]}\,\mathbf{U}_n\,\texttt{[SEP]}\},$$

where `[CLS]` and `[SEP]` are special tokens. We separate $\mathbf{C}$ and $\mathbf{R}$ with `[SEP]` to guide the model to learn the relationship between the context and response. $\mathbf{X}$ is then fed into the BERT encoder, which is a deep multi-layer bidirectional Transformer, to obtain a contextualized representation $\mathbf{H}$.

B. Separation

To obtain the representation of each individual utterance and response, we split the last-layer hidden state $\mathbf{H}$ into $\mathbf{H}^R$ and $\mathbf{H}^C = \{\mathbf{H}_1^u, \ldots, \mathbf{H}_n^u\}$ as the representations of the response and context, according to its position information. All utterances $\mathbf{H}_i^u$ in the same context are padded to the maximum length l among them.

C. Selection

To select the pivot utterances from the context, this module scores each utterance with respect to the response.[3] The m top-scoring utterances are selected as the topic clues.

Since both utterance or response are ended with a special token `[SEP]`, which is supposed to learn the sentence structure after BERT's pre-training through the next sentence objective [5], we pick it out as the representation of the corresponding utterance or response. Let $\mathbf{H}_i^u (i \in [1, n])$ and $\mathbf{H}^R$ denote

---

[2] We found this strategy could better take advantage of the benefits from the pairwise interaction of the inputs in BERT. Detailed discussion is shown in Section VI-C.

[3] It is possible to calculate the similarity with other tokens, such as the special token, `[CLS]`, which is supposed to carry the global information of the whole sequence, or with the last utterance. Here we only take the response, for example. Detailed discussion is shown in Section VI-B.

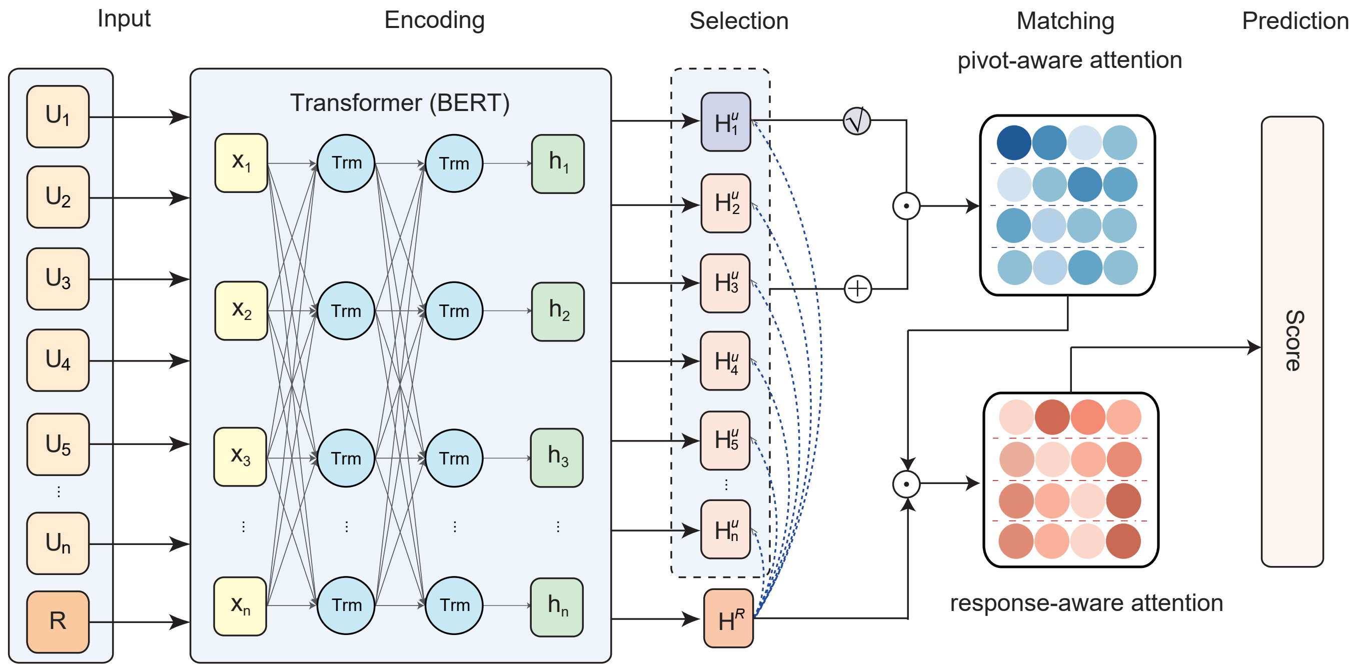


Fig. 1. Structural overview of the context—response scoring model for response selection task.

the utterance representation and response representation, respectively. We calculate the distance between each utterance-response pair to obtain the best related utterance(s):

$$\mathbf{S}_i = \text{Dist}(\mathbf{H}^R, \mathbf{H}_i^u), \tag{1}$$

where Dist$(\cdot,\cdot)$ is the distance measurement. In this work, we use cosine similarity.

After scoring each utterance, the m top-scoring sentences are selected and concatenated together following the original order to form a pivot context $\mathbf{H}^P \in \mathbb{R}^{q\times d}$, where the context length $q = m \times l$ and d denotes the dimension.

D. Matching

The matching layer is used to model the relation between the context and response, which contains two parts, 1) we first compute the *pivot-aware attention* to obtain the refined context; 2) we then calculate the *response-aware attention* to estimate the matching relationship between the refined context and response.[4]

Multi-head attention [40] is used in this work to capture the relationship between two sequences. We denote it as MHA$(\cdot)$, which is implemented as follows:

$$\begin{aligned}
&\text{Att}(E'_Q, E'_K, E'_V) = \text{softmax}(\frac{E'_Q(E'_K)^T}{\sqrt{d_{head}}})E'^V,\\
&\text{head}_i = \text{Att}(E_Q W_i^Q, E_K W_i^K, E_V W_i^V),\\
&\text{MHA}(E_Q, E_K, E_V) = \text{Concat}(\text{head}_i ... \text{head}_h),
\end{aligned} \tag{2}$$

where $W_i^Q \in \mathbb{R}^{d_{model}\times d_{head}}, W_i^K \in \mathbb{R}^{d_{model}\times d_{head}}, W_i^V \in \mathbb{R}^{d_{model}\times d_{head}}, E_Q \in \mathbb{R}^{d_q\times d_{model}}, E_K \in \mathbb{R}^{d_k\times d_{model}}, E_V \in \mathbb{R}^{d_v\times d_{model}}$, d_q, d_k, d_v and d_{head} denote the dimension of Query vectors, Key vectors, Value vectors and each head, respectively. h denotes the number of heads. We always assume $d_k = d_v$ and $d_{model} = h \times d_{head}$.

In detail, the refined context is produced by taking the pivot presentation $\mathbf{H}^P$ as the attention to the context representation $\mathbf{H}^C$:

$$\mathbf{H}^{CP} = \text{MHA}(\mathbf{H}^C, \mathbf{H}^P, \mathbf{H}^P), \tag{3}$$

where $\mathbf{H}^{CP}$ is the weighted sum of all the hidden states and it represents how the vectors in $\mathbf{H}^C$ can be aligned to each hidden state in $\mathbf{H}^P$.

Similarly, we calculate the response-aware attention by taking the response representation $\mathbf{H}^R$ as the attention to the refined context $\mathbf{H}^{CP}$. Thus, we have $\mathbf{H}^G = \text{MHA}(\mathbf{H}^{CP}, \mathbf{H}^R, \mathbf{H}^R)$ as the response-aware contextual representation.

In accumulating the response-aware contextual representation $\mathbf{H}^G \in \mathbb{R}^{u\times d}$ for the final prediction where $u = n \times l$, we employ a GRU to propagate information in $\mathbf{H}^G$ where each element in the dimension of u is seen as the time step in GRU. Supposing $\hat{\mathbf{H}} = [\hat{h}_1, \dots, \hat{h}_u]$ denotes the hidden states of the input sequence, we have

$$\hat{\mathbf{H}} = \mathbf{GRU}(\mathbf{H}^G). \tag{4}$$

The last hidden state $\hat{h}_u \in \mathbb{R}^d$ is selected for final prediction.

Note that the matching procedure can be regarded as two stages of interaction between the utterances and response. In detail, the joint pairwise input of the context and response interacts with the deep Transformer encoder where the coarse matching information can be fetched. After selecting the pivot

---

[4] To make fair comparisons, i.e., the analysis in Section VI-A, our baseline also adopts the matching mechanism by taking all the utterances as the pivots.

utterances from the context, we conduct more fine-grained two-step matching between the response and these utterances for further enhancement.

E. Prediction

We concatenate the last hidden state $\hat{h}_u$ with the first hidden state h_0 of the BERT encoder and feed the result into the output softmax layer to compute the final matching score.[5] We define $g(\hat{h}_u, h_0)$ as

$$g(\hat{h}_u, h_0) = \text{SoftMax}(\mathbf{W}_g[\hat{h}_u; h_0] + \mathbf{b}_g), \tag{5}$$

where $\mathbf{W}_g$ and $\mathbf{b}_g$ are trainable parameters. During the training phase, model parameters are updated according to a cross-entropy loss.

IV. INCORPORATION OF EXTRA KNOWLEDGE

Now, we extend the above pivot selection framework to the recent advanced conversational reading comprehension task that can benefit from extra knowledge injection. Besides the context C for the concerned multi-turn dialogue MRC represented as $[\mathbf{U}_1, \ldots, \mathbf{U}_{n_u}]$, the model is required to answer the related questions $\mathbf{Q}$, by selecting the response from an answer set $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_{n_a}]$. In this work, we treat the question and the answer option as an integral through concatenation, so that we have $\mathbf{QA}_j = [\mathbf{Q}; \mathbf{A}_j]$. Therefore the task aim is to find the most proper question-answer pair according to the context.

A. Extracting Knowledge

First, we extract the knowledge sources from an external knowledge graph, ConceptNet [34].[6] Items with weight less than a threshold or contain words that are not in the vocabulary of the chosen PrLM are removed from KG. The items are triples with the form {*relation, head, tail*}, which are rewritten as facts (*e.g.* {*causes, virus, disease*} to *virus causes disease*). These facts are encoded with our adopted PrLM and the last hidden states $\mathbf{H}_k$ ($\mathbf{H}_k \in \mathbb{R}^{n_k \times d_{model}}$ where n_k denotes the number of tokens in the fact) are taken as the output so that the representations of knowledge and context are in the same vector space. A self-attention module is used to refine the representation of each fact. We use mean-pooling in the end to aggregate the representation of each token and get a final representation r_k ($r_k \in \mathbb{R}^{d_{model}}$) for each fact.

$$\begin{aligned} \text{SelfAttention}(\mathbf{H}_k) &= \text{MHA}(\mathbf{H}_k, \mathbf{H}_k, \mathbf{H}_k), \\ r_k &= \text{mean}(\text{SelfAttention}(\mathbf{H}_k)). \end{aligned} \tag{6}$$

B. Retrieve Relevant Knowledge

Each utterance $\mathbf{U}_i$ is annotated with part-of-speech (POS) tags by NLTK [25]. For tokens with POS like adjectives, nouns, and verbs, we assume that they contain more implicit information than others; thus, items related to them are retrieved in KG. In all the chosen items, top p (a hyperparameter) ones are selected to enhance the context representation. The extracted knowledge items are denoted as $\mathbf{CK} = [r_{c_1}, ...r_{c_p}]$.

For a QA-pair $\mathbf{QA}_j$, we follow the same steps in dealing with $\mathbf{U}_i$ to get the relevant knowledge items:

$$\mathbf{QAK}_j = [r_{j_1}, r_{j_2}, ...r_{j_k}], \tag{7}$$

where j_k is the number of chosen knowledge items for $\mathbf{QA}_j$.

C. Encoding and Representation Refinement

For each $\mathbf{QA}_j$, it is concatenated with $\mathbf{C}$ as input encoded with PrLM. The last hidden states $\mathbf{H}_t$ are then separated into context representation $\mathbf{H}^C$ and QA-pair representation $\mathbf{H}^{QA}$.

Since we have questions in conversational reading comprehensions, which can serve as better indicators for selecting the pivot utterances. Therefore, the pivot selection is lightly different, which is based on the matching scores between each utterance with respect to the question instead. The representation of pivots turns $\mathbf{H}^P$ is extracted from $\mathbf{H}^C$ based on the position of key utterances as described in Section III-C. We use MHA$(\cdot)$ to calculate the pivot-refined context representation and $\mathbf{H}^P$. Simlilarly, we get the knowledge-refined representation of context and QA-pair:

$$\begin{aligned} \mathbf{H}^{CP} &= \text{MHA}(\mathbf{H}^C, \mathbf{H}^P, \mathbf{H}^P), \\ \mathbf{H}^{CK} &= \text{MHA}(\mathbf{H}^C, \mathbf{CK}, \mathbf{CK}), \\ \mathbf{H}^{QA} &= \text{MHA}(\mathbf{H}^{QA}, \mathbf{QAK}, \mathbf{QAK}). \end{aligned} \tag{8}$$

D. Representation Fusion

Following Zhu et al. [63], we use a Dual Multi-head Co-Attention (DUMA) module to fuse the representation of context and QA-pair.

$$\begin{aligned} &\text{MHA}_1 = \text{MHA}(\mathbf{H}^C, \mathbf{H}^{QA}, \mathbf{H}^{QA}), \\ &\text{MHA}_2 = \text{MHA}(\mathbf{H}^{QA}, \mathbf{H}^{QA}, \mathbf{H}^C), \\ &\text{DUMA}(\mathbf{H}^C, \mathbf{H}^{QA}) = \text{Concat}(\text{mean}(\text{MHA}_1) \\ &,\text{mean}(\text{MHA}_2)). \end{aligned} \tag{9}$$

Based on different represenation of context and QA-pair calculated above, DUMA module may give three types of outputs, the original $\mathbf{O}^O$, the pivot utterances refined $\mathbf{O}^P$ and the knowledge refined $\mathbf{O}^K$ as follow.

$$\begin{aligned} \mathbf{O}^O &= \text{DUMA}(\mathbf{H}^C, \mathbf{H}^{QA}), \\ \mathbf{O}^P &= \text{DUMA}(\mathbf{H}^{CP}, \mathbf{H}^{QA}), \\ \mathbf{O}^K &= \text{DUMA}(\mathbf{H}^{CK}, \mathbf{H}_{QA}). \end{aligned} \tag{10}$$

Then these three kinds of outputs are fused together as the final output. $\mathbf{O}^P$ and $\mathbf{O}^K$ are concatenated together and mapped to dimension of $2d_{model}$ through a linear layer to get the knowledge-pivot-utterances refined (KPR) output $\mathbf{O}^{KPR}$. Then we fuse the original output $\mathbf{O}^O$ and the KPR output $\mathbf{O}^{KPR}$ to get the final output $\mathbf{O}$. Concatenation is chosen as our fuse function.

---

[5] h_0 is regarded as the pooled representation as the BERT output [7].

[6] We selected ConceptNet because it is the most widely-used corpus in the related studies, which is also well formed as the structural commonsense knowledge network that suits our task.

E. Decoding

Our model decoder takes $\mathbf{O}$ and computes the probability distribution over answer options. Let $\mathbf{A}_i$ be the i-th answer option and $\mathbf{O}_i$ is the corresponding output of $< \mathbf{C}, \mathbf{Q}, \mathbf{A}_i >$. The loss function is computed by

$$L(\mathbf{A}_i|\mathbf{C}, \mathbf{Q}) = -\log(\frac{\exp(W^T\mathbf{O}_i)}{\sum_{j=1}^{l_a}\exp(W^T\mathbf{O}_j)}), \tag{11}$$

where W is a learnable parameter.

V. EXPERIMENT

A. Dataset

We evaluated our model on three public multi-turn dialogue response selection datasets, the English Ubuntu Dialogue Corpus (Ubuntu) [26] and two Chinese datasets, namely the Douban Conversation Corpus (Douban) [44] and E-commerce Dialogue Corpus (ECD) [54] to evaluate our backbone PoDS framework demonstrated in Section III, and one conversational comprehension dataset, i.e., DREAM [35] to assess our knowledge-enhanced variant described in Section IV.[7]

1) Dialogue Response Selection

a) Ubuntu Dialogue Corpus: Ubuntu Dialogue Corpus comprises multi-turn human-computer conversations constructed from chat logs of the Ubuntu forum. It contains 1 million context-response pairs for training and 0.5 million pairs for validation and testing. The training set contains context-response pairs labeled as a positive or negative response randomly selected on the dataset. In both validation and test sets, each context contains one positive response and nine negative responses.

b) Douban Conversation Corpus: Douban Conversation Corpus is an open-domain dataset constructed by the Douban Group, which provides a popular social networking service in China. Similarly constructed as the Ubuntu corpus, this corpus contains 1 million context-response pairs for training, 0.5 million pairs for validation, and 6670 pairs for testing.

c) E-commerce Dialogue Corpus: E-commerce Dialogue Corpus is a dataset of real-world conversations between customers and customer service staff. It contains 1 million context-response pairs for training and 10,000 pairs for both validation and testing. A topic has at least five types of conversation (e.g., commodity consultation, logistics discussions, recommendations, negotiations, and chat) relating to more than 20 commodities. The positive-to-negative ratio is 1:1 in training and validation and 1:9 in testing.

2) Conversational Reading Comprehension

d) DREAM: DREAM is a newly released dialogue-based multi-choice MRC dataset, which is collected from English exams. Each dialogue, as the given context, has multiple questions, and each question has three response options. In total, it contains 6,444 dialogues and 10,197 questions. The most important feature of the dataset is that more than 80% of the questions are non-extractive, and more than a third of the given questions involve commonsense knowledge. As a result, the dataset is small but quite challenging.

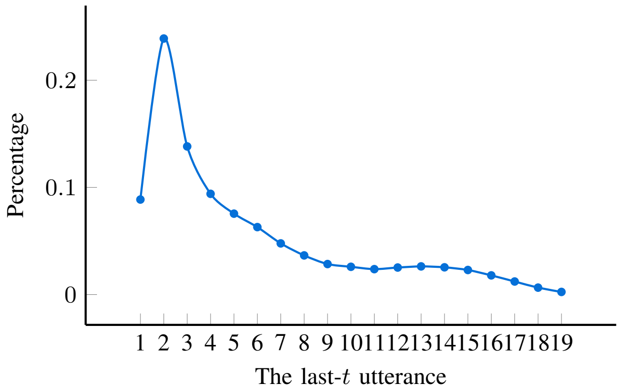


Fig. 2. Proportion of the most related utterance (last-t) for gold response.

B. Evaluation Metrics

For the dialogue response selection tasks, we used the same evaluation metrics used in previous works [44, 54]. Each model was evaluated by selecting the k best-matching responses from n available candidates for the given conversation context. We calculate the recall of the true positive replies among the k selected responses, denoted $Rn@k$, as the main evaluation metric for each model. In addition, we used the mean average precision (MAP), mean reciprocal rank (MRR), and precision at position 1 ($P@1$) for the Douban Conversation Corpus. The reason for using the additional metrics is that the Douban Conversation Corpus is different from the other three datasets as it includes multiple correct candidates for a context in its test set, which may lead to low $Rn@k$.

For the conversational reading comprehension task, the evaluation metric we use is accuracy, $acc{=}N^+/N$, where N^+ denotes the number of examples the model selects the correct answer, and N denotes the total number of evaluation examples.

C. Baseline Models

We used the pre-trained BERT as a baseline with its pairwise classification setting. In our study, we found it effective to use the context texts from the task-specific training set to fine-tune BERT with the language modeling objectives (Masked LM and Next Sentence Prediction) [7] before training the PoDS, which we call domain fine-tuning (DFT). For DFT, the hyper-parameter setting is the same with the task training as demonstrated in Section V-D.

We also compared our PoDS with the following published works.

a) Single-turn matching methods: Single-turn matching models, including the RNN [26], CNN [26], LSTM [26], BiLSTM [14], MV-LSTM [41], and Match-LSTM [42], concatenated all utterances in the context as a long document to calculate the matching score with a candidate response.

---

[7] We also tried to employ the latter method for the three conventional response task; however, we did not see any performance gains. The reason is very likely that the three datasets are concerning technical discussion, social media, and e-commerce, which do not require much commonsense to solve the task.

TABLE III
EVALUATION RESULTS OF DIFFERENT MODELS ON THE DOUBAN CONVERSATION CORPUS AND E-COMMERCE DIALOGUE CORPUS.

Model	**Ubuntu Dialogue Corpus**				**Douban Conversation Corpus**						**E-commerce Dialogue Corpus**		
	R$_2$@1	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
RNN	76.8	40.3	54.7	81.9	39.0	42.2	20.8	11.8	22.3	58.9	32.5	46.3	77.5
CNN	84.8	54.9	68.4	89.6	41.7	44.0	22.6	12.1	25.2	64.7	32.8	51.5	79.2
LSTM	90.1	63.8	78.4	94.9	48.5	53.7	32.0	18.7	34.3	72.0	36.5	53.6	82.8
BiLSTM	89.5	63.0	78.0	94.4	47.9	51.4	31.3	18.4	33.0	71.6	35.5	52.5	82.5
DL2R	89.9	62.6	78.3	94.4	48.8	52.7	33.0	19.3	34.2	70.5	39.9	57.1	84.2
MV-LSTM	90.6	65.3	80.4	94.6	49.8	53.8	34.8	20.2	35.1	71.0	41.2	59.1	85.7
Match-LSTM	90.4	65.3	79.9	94.4	50.0	53.7	34.5	20.2	34.8	72.0	41.0	59.0	85.8
Multi-View	90.8	66.2	80.1	95.1	50.5	54.3	34.2	20.2	35.0	72.9	42.1	60.1	86.1
SMN	92.6	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	93.8	76.7	87.4	96.9	55.0	60.1	42.7	25.4	41.0	75.7	52.6	72.7	93.3
IMN	94.6	79.4	88.9	97.4	57.0	61.5	44.3	26.2	45.2	78.9	62.1	79.7	96.4
MRFN	94.5	78.6	88.6	97.6	57.1	61.7	44.8	27.6	43.5	78.3	-	-	-
IOI	94.7	79.6	89.4	97.4	57.3	62.1	44.4	26.9	45.1	78.6	56.3	76.8	95.0
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
BERT	95.3	81.7	90.4	97.7	58.8	63.1	45.3	27.7	46.4	81.8	62.1	80.2	96.0
PoDS	96.0	82.8	91.2	98.1	59.8	63.6	46.0	28.7	46.8	**84.5**	63.3	81.0	96.7
+ DFT	**96.6**	**85.6**	**92.9**	**98.5**	**59.9**	**63.7**	46.0	28.7	**46.9**	83.9	**67.1**	**84.2**	**97.3**

Note: MSN [49] is the state-of-the-art model among published works. The best results are in boldface.

b) Multi-turn matching methods: Multi-turn matching models, including the multi-view model [60], DL2R [47]. SMN [44], DUA [54], deep attention matching network (DAM) [61], IMN [11], MRFN [37], IOI [36], and MSN [49], matched the response with the utterances in the context.

c) Pre-trained Language Models: Pre-trained language models, including BERT [7], XLNet [48], RoBERTa [24], and ALBERT [17].[8]

D. Implementation Details

Our implementations were based on the PyTorch version of BERT.[9] For the sake of training efficiency on the large corpora, we use `BERT-base-uncased` and `BERT-base-chinese` as initial weights on the English (Ubuntu) and Chinese datasets (Douban and ECD), respectively.[10] For the relatively smaller-scale DREAM dataset, we used ALBERT of both `base` and `xxlarge` variants [17] as our encoder, which is a recent dominant PrLM, to see if we can achieve even better performance though on such a strong PrLM model. We set the m to 12 by default, i.e., top-12 relevant utterances for context representation with careful consideration of effectiveness and efficiency (the analysis of m will be presented in Section VI-A. We set the initial learning rate in {1e-5, 2e-5, 3e-5} with a warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size was selected in {24, 32, 64}. The maximum number of epochs was set in [2, 5] depending on the dataset. Texts were tokenized using wordpieces, with a maximum length of 384 in all experiments. All our models were run on 32G NVIDIA V100 GPUs. We ran all the models up to 2 or 3 epochs and the best models on the dev set are chosen from all the checkpoints for test evaluation.

E. Preliminary Experiments

Previous works [49, 54] heuristically selected the last utterance as the directive clue to refine the context, which showed substantial benefits to the multi-turn dialogue modeling. Intuitively, the last utterance would be instructive for the subsequent response. However, there are many complex multi-turn conversations that consist of jumping topics, as shown in Table I.

Inspired by the recent studies that enjoyed adopting pre-trained language models such as BERT to measure the semantic similarity between texts [32, 52], we trained a multi-turn dialogue model using BERT to measure the average cosine similarity of each utterance and the gold response using the dev set of Ubuntu Dialogue Corpus [26]. Figure 2 shows the proportion of the most related (last-t) utterances for the gold response. We observe that, in most cases, the last three utterances are the most relevant to the intended response, which would be quite instructive for response selection. This observation verified the effectiveness of using the last utterance as the directive clue for context refinement and matching [49, 54] to some extent. However, we showed that the last one would not always be the best. This finding motivates us to investigate a more flexible way to select the most directive utterance(s) for fine-grained context modeling and context-response interactions. Besides the pivot utterances extracted from the given dialogue context, we are also interested in incorporating other indicators to improve the model capacity of conversation comprehension, such as external knowledge.

F. Main Results

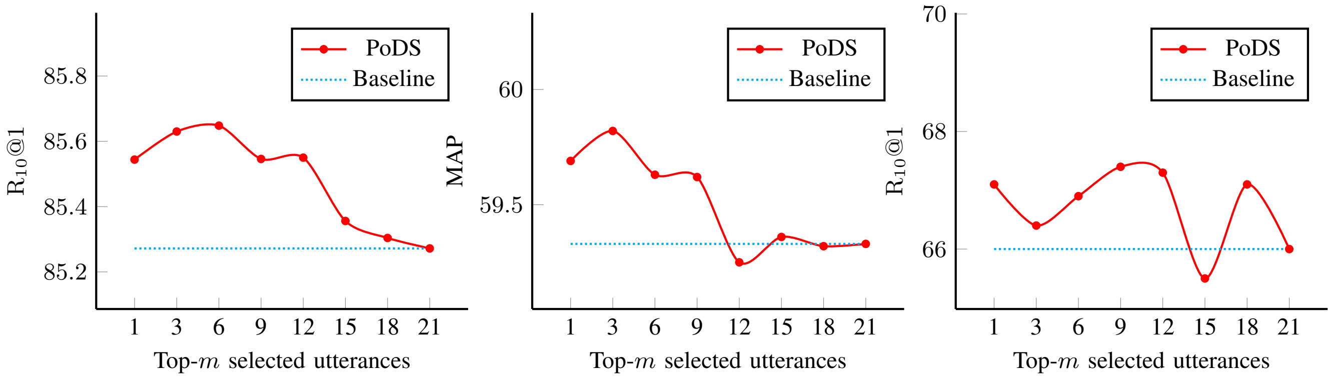
Table III gives the evaluation results for the PoDS and baseline models on the traditional response selection tasks,

---

[8] Due to high computation cost, we only use widely-used BERT as the baseline for the three response selection tasks. For the conversational response selection task, we compare all these baselines results in Table IV.

[9] https://github.com/huggingface/pytorch-pretrained-BERT.

[10] Since the corpora are quite large, training a BERT-based model requires a very long time, e.g., about 8 hours for one epoch on Ubuntu, though using the `BERT-base` models. Therefore, the SOTA results were reported on `base` models as well.



(a) Accuracy on the Ubuntu dataset (b) Accuracy on the Douban dataset. (c) Accuracy on the ECD dataset.

Fig. 3. Accuracy for a varying number of selected utterances.

TABLE IV
RESULTS ON DREAM DATASET.

Model	**Dev**	**Test**
FTLM++	58.1*	58.2*
BERT$_{large}$	66.0*	66.8*
XLNet	-	72.0*
RoBERTa$_{large}$	85.4*	85.0*
RoBERTa$_{large}$+MMM	88.0*	88.9*
ALBERT$_{xxlarge}$	89.2*	88.5*
ALBERT$_{xxlarge}$+DUMA	89.3†	**90.4**†
ALBERT$_{base}$	67.4	67.3
ALBERT$_{base}$+KPR	69.3	68.7
ALBERT$_{xxlarge}$	89.1	88.2
ALBERT$_{xxlarge}$+KPR	**90.2**	89.8

Note: Results denoted by $\star$ are from Jin et al. [13], † are from Zhu et al. [63]. MMM is short for Multi-stage Multi-task Learning for Multi-choice Reading Comprehension [13].

showing that the PoDS outperformed the other models on all metrics and datasets. In particular, our model surpassed DAM [61] by a large margin, where both models are based on a Transformer encoder. Moreover, our proposed model outperformed the strong BERT baseline substantially, achieving new state-of-the-art performance on all datasets.

We see that the recent models work relatively poorer in the Chinese datasets. The possible reason would be that those Chinese datasets are newer and more challenging datasets than the domain-specific Ubuntu dataset. Douban is an open domain conversation dataset that contains many more topics, and ECD is for the complex E-commence scenario that involves various types of conversations, e.g., commodity consultation, logistics express, recommendation, negotiation, and chitchat, over different commodities. Therefore, the Chinese tasks often involve more complex topic shifts and multiple intentions in a dialogue context, which require stronger models to solve the problems.

Table IV gives the results on the DREAM dialogue comprehension dataset. Experimental results show our model obtains a great improvement compared to the baseline and achieves state-of-the-art performance for DREAM on dev set.

TABLE V
COMPARISON OF DIFFERENT SELECTION METHODS ON THE UBUNTU DATASET.

Model	R$_{10}$@1	R$_{10}$@2	R$_{10}$@5
Baseline	85.2	92.5	98.2
PoDS	85.6	92.9	98.5
NLI	85.4	92.7	98.4
CLS	85.5	92.8	98.6
Last	85.3	92.7	98.5
Random	85.3	92.6	98.3
POS	85.5	92.7	98.5

VI. ANALYSIS

A. *Effects of the Number of Selected Utterances*

Intuitively, the number of the pivot utterances m would affect performance. We evaluated the performance of our model for different numbers of selected sentences. The comparison results are shown in Figures 3. It is seen that the performance of the PoDS was remarkably improved when m scales from 1 to 12. This indicates that the information carried by one relevant utterance is commonly insufficient to pinpoint the important part of the context and match with the response; thus, moderately selecting more utterances for matching may enhance the performance. Meanwhile, this positive effect largely dissipated as m increases from 12 to 20. This is because an excessively large number of selected utterances may incorporate irrelevant or misleading information, which hurt performance. Meanwhile, processing more utterances would result in more computational cost. We thus selected the top-12 relevant utterances for context representation with careful consideration of effectiveness and efficiency.[11]

B. *Comparison of Different Selection Methods*

Besides the simple cosine similarity to measure the distance, Natural Language Inference (NLI) models also serve as an

---

[11] We also considered setting a cosine similarity threshold and dividing the contexts into relevant and irrelevant parts. However, the scores we obtained are not quite distinguishable – they either gathered around some specific scores, i.e., 0.9 or scattered irregularly. Therefore, we decided to use the empirical way of setting the "hard threshold" with the fixed number of top-ranked utterances and achieved the performance gains.

TABLE VI
COMPARISON OF RESULTS OBTAINED FOR DIFFERENT FEEDING PATTERNS ON THE UBUNTU DATASET.

Model	R_{10}@1	R_{10}@2	R_{10}@5
Baseline	85.2	92.5	98.2
Separate	84.9	92.3	98.1
Packed (PoDS)	85.6	92.9	98.5

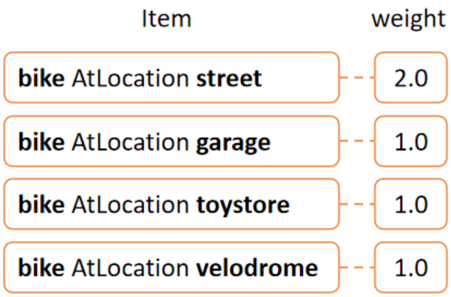


Fig. 4. Knowledge item in ConceptNet. Values on the right indicate the weight of each item in ConceptNet. A large weight means larger possibility.

effective measure of semantic similarity [15]. We trained a BERT-based NLI model on the SNLI dataset [2] with 91.1% dev accuracy, the linear layer has three output neurons for labels of *contradiction*, *entailment* and *neutral*. We apply softmax on these outputs to get the probability of *entailment* label as the similarity.

In addition to the similarity calculation algorithms, employing what kinds of patterns for selection would also matter, e.g., the special token used in BERT, `[CLS]`, which is supposed to carry the global information of the whole sequence, or with the last utterance. To investigate the influence of the selection method, we compared the results with different alternatives, 1) *CLS*: used the representation of the special token `[CLS]` to replace the response representation for calculating the cosine similarity with each utterance; 2) *Last*: directly used the last utterance as the pivot utterance; 3) *Random*: randomly sampled an utterance as the pivot utterance; 4) *Pos*: only calculated the cosine similarity for positive labeled responses during training. For the negative ones, we took the last utterance as the pivot. This aims to alleviate the noisy matching from the negative samples. Table V shows the results. We see that the selection methods basically work better than the baseline, and it is possible to adopt alternatives for simplicity – using `[CLS]` representation is a good alternative and only considering positive samples (POS) achieves a similar result. Besides, using the last utterance shows to be suboptimal and random sampling is not a good practice.

C. Discussion on the Matching Method

In this study, we found that the feeding pattern to the encoder affects the performance when it comes to PrLMs, like BERT. It is a natural idea to directly feed the separate utterance or response as individual input to the encoder, like in previous RNN-based response selection methods [36, 44, 54, 61]. Then, the encoded representations are interacted by the same selection and matching mechanisms. Table VI shows the comparison of the packed and separate inputs. We found the benefit is trivial if we feed the separate inputs to BERT or use separate BERT embeddings to feed previous state-of-the-art models. The reason might be that the deep multi-head attention in BERT is effective for modeling the interactions of the paired input. We, therefore, recommend feeding the whole input that is packed with context and response and separating them later for further matching. On the basis of such a strong baseline, we also verified that we could yield further gains with our selective matching.

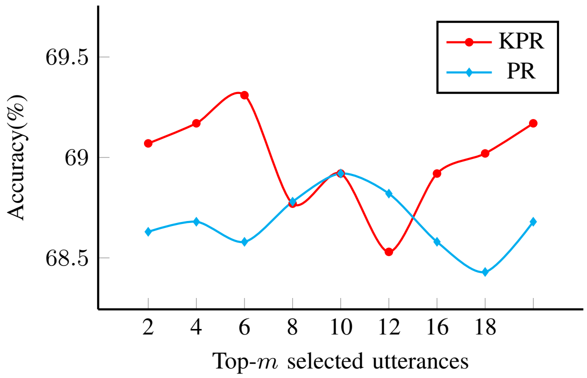


Fig. 5. Comparison between complete model (KPR) and model without knowledge-refinement on different numbers of pivot utterances (PR).

D. Effects of External Knowledge Appending

Multi-turn dialogues more or less involve implicit or explicit commonsense when the conversation goes on; therefore, understanding them requires the support of knowledge. In our model, a knowledge graph is used by adding related knowledge items. In the example in Table II (in section I), we need to know where a bike is likely to appear for answering the question. As shown in Figure 4, the knowledge item $\{$*atlocation, bike, street*$\}$ can be found in our KG, which means *Bikes are always found on the street*. It has a weight of 2, indicating it is more possible than others with a lower weight. With such a fact, the model thus can answer the question correctly.

To state the effects more clearly, we remove the knowledge refined output from our model by dropping $\mathbf{O}^K$ in Eq. 10 and evaluate it on the dev set of the DREAM dataset. The results are shown in Figure 5. We can see a general performance improvement from the knowledge refined part for different numbers of pivot utterances.

E. Effects of Pivots Utterance Extraction

As mentioned in the previous section, a challenge in understanding and modeling multi-turn dialogues is that the topic shifts in different turns, which means only a few turns are truly related to the question.

Here, we select the NLI scores for better interpreting the benefits of utterance selection, because the numbers are more informative and distinctive for each utterance than the cosine similarity scores which either gather around some specific scores, i.e. 0.9, or scatter irregularly. Besides, the NLI selection

TABLE VII
RELEVANCE SCORE FOR EACH TURN CORRESPONDING TO THE CORRECT ANSWER.

Dialogue 1	Score
W: Well, I'm afraid my cooking isn't to your taste.	-1.91
M: Actually, I like it very much.	**-1.49**
W: I'm glad you say that. Let me serve you some more fish.	-2.53
M: Thanks. I didn't know you were so good at cooking.	**-1.66**
W: Why not bring your wife next time?	-2.26
M: OK, I will. She will be very glad to see you, too.	-1.87
Question: *What does the man think of the woman's cooking?*	
A. It's really terrible.	
*B. It's very good indeed. **	
C. It's better than what he does.	

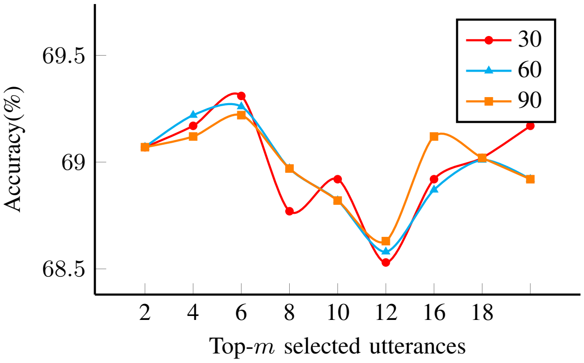


Fig. 6. Influence of number of knowledge items (K items) on the dev set of DREAM dataset.

TABLE IX
AN EXTRACTED EXAMPLE FROM THE TEST SET OF UBUNTU WHERE THE BASELINE FAILS BUT IS SUCCESSFULLY SOLVED BY OUR MODEL.

Dialogue Context
U$_1$: *how do i do a real clean remove and install*
U$_2$: *apt get/dpkg complain that its missing so i touch it*
U$_3$: *try to remove and start from scratch but the initscript is still missing*
U$_4$: *apt-get vs aptitude*
U$_5$: *thanks*
U$_6$: *yeah that worked*
PoDS: *thanks folks yeah but you re all so nice :-)*
Baseline: *maybe it cant be started from session init because of some particular status i really dont know :p ...*

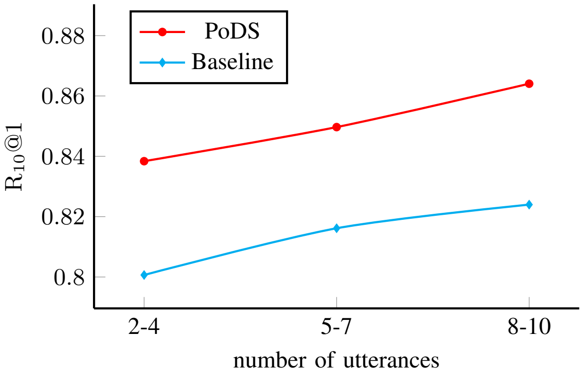


Fig. 7. R_{10}@1 of PoDS and the baseline BERT on different numbers of utterances.

method also achieves quite comparable results with the cosine one.

A higher relevance score indicates that it is more likely to conclude the QA given the corresponding turn. The example is shown in Table VII verifies our hypothesis. Turns with top 2 entailment scores can directly give the answer, while the other turns have nothing to do with the answer. This example shows that key turns are decisive for context refinement in representation, and they are suggestive for explaining the contribution of each utterance.

To address the effects more clearly, we evaluate our model by removing the key-turns refined output from our model and evaluate it. The maximum number of knowledge items is 30. The results are shown in Table VIII, which indicates significant performance loss on both dev and test sets. The results verify that using the key turns as pivot utterances is indispensable for the advanced performance.

F. Effects of Number of Knowledge Items

The number of knowledge Items can affect performance as well. So we evaluate our model on different numbers of knowledge items on the dev set of the DREAM dataset. The results are shown in Figure 6. Contrary to our expectation, the results show little difference in various numbers of knowledge items. We suppose that the attention mechanism employed to refine context with external knowledge would contribute to the result because knowledge items with small weight tend to be concerned less in the attention mechanism as well since they are less relevant to the context. Therefore, the knowledge-refined context is similar though using more numbers of knowledge items, leading to a similar final performance when choosing a different number of knowledge items.

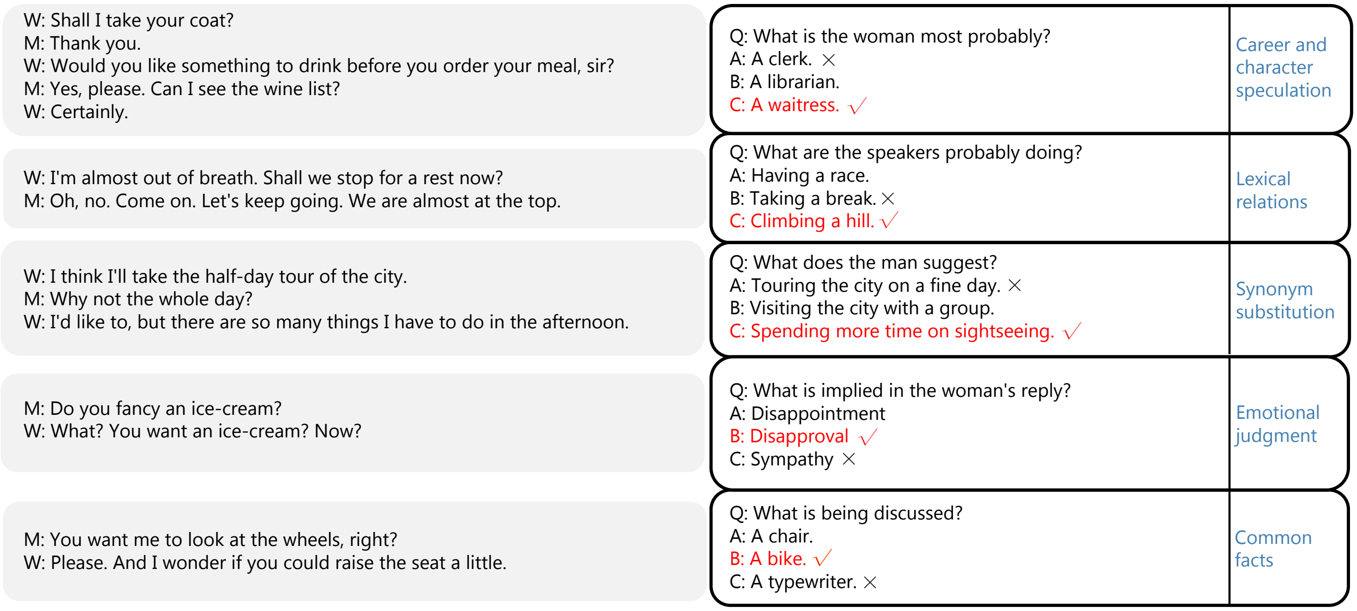
TABLE VIII
COMPARISON BETWEEN COMPLETE MODEL, MODEL WITHOUT KEY-TURNS-REFINEMENT, AND BASELINE.

Model	Dev	Test
ALBERT$_{base}$	67.40	67.31
ALBERT$_{complete}$(ALBERT$_{base}$+KPR)	69.32	68.71
-PR	67.94	67.66

G. Prediction Analysis

We analyzed the predictions from both of our system and the baseline on the test set of Ubuntu and DREAM datasets to understand how our model solves the error cases made by the baseline model, respectively.

Table IX shows an example from the Ubuntu test set. The response selected by PoDS is highly related to the topic flow of the utterances, from problem-solving to expressing the thanks, which indicates that our model is better at modeling the fluency of the dialogue, in other words, capturing the long-term relevance of the response and overall dialogue topic flow. For further exploration, we conduct an analysis by measuring the model performance on different context length that varies in different numbers of the utterances. Figure 7 shows that



W: Shall I take your coat?
M: Thank you.
W: Would you like something to drink before you order your meal, sir?
M: Yes, please. Can I see the wine list?
W: Certainly.

Q: What is the woman most probably?
A: A clerk. ×
B: A librarian.
C: A waitress. ✓

Career and character speculation

W: I'm almost out of breath. Shall we stop for a rest now?
M: Oh, no. Come on. Let's keep going. We are almost at the top.

Q: What are the speakers probably doing?
A: Having a race.
B: Taking a break. ×
C: Climbing a hill. ✓

Lexical relations

W: I think I'll take the half-day tour of the city.
M: Why not the whole day?
W: I'd like to, but there are so many things I have to do in the afternoon.

Q: What does the man suggest?
A: Touring the city on a fine day. ×
B: Visiting the city with a group.
C: Spending more time on sightseeing. ✓

Synonym substitution

M: Do you fancy an ice-cream?
W: What? You want an ice-cream? Now?

Q: What is implied in the woman's reply?
A: Disappointment
B: Disapproval ✓
C: Sympathy ×

Emotional judgment

M: You want me to look at the wheels, right?
W: Please. And I wonder if you could raise the seat a little.

Q: What is being discussed?
A: A chair.
B: A bike. ✓
C: A typewriter. ×

Common facts

Fig. 8. Commonsense problems from the test set of DREAM where the baseline fails but is successfully solved by our model.

PoDS performs robustly and significantly than the baseline, especially for long contexts with more than 8 utterances.

For the more readable DREAM dataset that involves commonsense knowledge, we collected 92 examples that the baseline failed to answer while our model succeeded.

- 44.6%: matching or summary cases which can be solved by simple text matching, extraction and searching.
- 25.0%: logic consistency that involves complex reasoning.
- 6.5%: arithmetic problems that involve mathematical calculation.
- 23.9%: commonsense problems that include career and character speculation, synonymous substitution, lexical relations, emotional judgment based on modal particles, common facts, etc.

We observe that 23.9% commonsense problems have been well solved by our model equipped with commonsense knowledge injection, as examples shown in Figure 8. Actually, the logic problems (25.0%) also rely on commonsense, such as synonym substitution, antonym, and modal particle, to construct logic chains.

VII. CONCLUSION

In this work, we proposed a pivot-oriented deep selection model using BERT as the encoder with pivot-aware contextualized attention mechanisms for the multi-turn response selection task. We first pick out some of the turns from the dialogue directly related to the candidate response or question as pivot utterances. Then, the relevant knowledge items are picked out and encoded with PrLM. The dialogue context is refined with pivot utterances and external knowledge items for better language representation, which is employed for the matching candidate response. The procedures of our method are highly explainable and reflect a primary idea of cascading different models to get better language representation. Experimental results on four benchmark datasets show that our proposed model outperforms baseline models, achieving new state-of-the-art performance for multi-turn response selection. Case studies show that our selection strategies and the extra knowledge injection have certain effectiveness for improving the model performance. In future work, we will investigate how to model the topic flow and logical consistency across multi-turn conversations to improve selection performance.

REFERENCES

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4762–4779, 2019. doi: 10.18653/v1/P19-1470.

[2] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)*, pages 632–642, 2015.

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

[4] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1724–1734, 2014.

[5] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3684–3690, 2019.

[6] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.130.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[8] Maxwell Forbes and Yejin Choi. Verb Physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2017)*, pages 266–276, 2017. doi: 10.18653/v1/P17-1025.

[9] Zhenxin Fu, Shaobo Cui, Mingyue Shang, Feng Ji, Dongyan Zhao, Haiqing Chen, and Rui Yan. Context-to-session matching: Utilizing whole session for response selection in information-seeking dialogue systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1605–1613, 2020.

[10] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895, 2019.

[11] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 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*, pages 2321–2324, 2019.

[12] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. Utterance-to-utterance interactive matching network for multi-turn response selection in retrieval-based chatbots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:369–379, 2019.

[13] Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. MMM: Multi-stage multi-task learning for multi-choice reading comprehension. *arXiv preprint arXiv:1910.00458*, 2019.

[14] Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*, 2015.

[15] Rakesh Khobragade, Heaven Patel, Anand Namdev, Anish Mishra, and Pushpak Bhattacharyya. Machine translation evaluation using bi-directional entailment. *arXiv preprint arXiv:1911.00681*, 2019.

[16] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 329–339, 2016. doi: 10.18653/v1/D16-1032.

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

[18] Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 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, 2017.

[19] Lu Li, Chenliang Li, and Donghong Ji. Deep context modeling for multi-turn response selection in dialogue systems. *Information Processing & Management*, 58(1): 102415, 2020.

[20] Zuchao Li, Chaoyu Guan, Hai Zhao, Rui Wang, Kevin Parnow, and Zhuosheng Zhang. Memory network for linguistic structure parsing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2743–2755, 2020.

[21] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2019)*, pages 2822–2832, 2019.

[22] Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.741.

[23] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*, 2019.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[25] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, 2002. doi: 10.3115/1118108.1118117.

[26] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 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 (SIGDIAL 2015), pages 285–294, 2015. doi: 10.18653/v1/W15-4640.

[27] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65, 2017.

[28] Todor Mihaylov and Anette Frank. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2018)*, pages 821–832, 2018. doi: 10.18653/v1/P18-1076.

[29] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, 2018.

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

[31] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 2383–2392, 2016.

[32] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3973–3983, 2019.

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

[34] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 4444–4451, 2017.

[35] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264.

[36] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *ACL*, pages 1–11, 2019.

[37] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM 2019)*, pages 267–275, 2019.

[38] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, 2017.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in neural information processing systems*, pages 5998–6008, 2017.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NIPS 2017)*, pages 5998–6008, 2017.

[41] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-SRNN: modeling the recursive matching structure with spatial RNN. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2922–2928, 2016.

[42] Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451, 2016.

[43] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. Domain adaptive training bert for response selection. *arXiv preprint arXiv:1908.04812*, 2019.

[44] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 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, 2017.

[45] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, 2019.

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

[47] Rui Yan, Yiping Song, and Hua Wu. 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*, pages 55–64. ACM, 2016.

[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Gen-

eralized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

[49] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 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, 2019.

[50] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 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 2019)*, pages 111–120, 2019. doi: 10.18653/v1/D19-1011.

[51] Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. DCMN+: Dual co-matching network for multi-choice reading comprehension. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*, 2020.

[52] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

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

[54] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, and Hai Zhao. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 3740—-3752, 2018.

[55] Zhuosheng Zhang, Hai Zhao, Kangwei Ling, Jiangtong Li, Shexia He, and Guohong Fu. Effective subword segmentation for text comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(11):1664–1674, 2019. doi: 10.1109/TASLP.2019.2922537.

[56] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*, 2020.

[57] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. SG-Net: Syntax-guided machine reading comprehension. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[58] Zhuosheng Zhang, Hai Zhao, and Rui Wang. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*, 2020.

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

[60] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP)*, pages 372–381, 2016.

[61] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 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, 2018.

[62] Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. 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, 2018.

[63] Pengfei Zhu, Hai Zhao, and Xiaoguang Li. Dual multi-head co-attention for multi-choice reading comprehension. *arXiv preprint arXiv:2001.09415*, 2020.

Zhuosheng Zhang received his Bachelor's degree in internet of things from Wuhan University in 2016, his M.S. degree in computer science from Shanghai Jiao Tong University in 2020. He is working towards his Ph.D. degree in computer science with the Center for Brain-like Computing and Machine Intelligence of Shanghai Jiao Tong University. He was an internship research fellow at NICT from 2019-2020. His research interests include natural language processing, machine reading comprehension, dialogue Systems, and machine translation.

Junlong Li is an undergraduate student in the IEEE honor class of Shanghai Jiao Tong University. He majors in computer science and is expected to get his Bachelor's degree in 2022. His research interests include natural language processing, dialogue systems, and machine reading comprehension.

Hai Zhao received the BEng degree in sensor and instrument engineering, and the MPhil degree in control theory and engineering from Yanshan University in 1999 and 2000, respectively, and the PhD degree in computer science from Shanghai Jiao Tong University, China in 2005. He is currently a full professor at department of computer science and engineering, Shanghai Jiao Tong University after he joined the university in 2009. He was a research fellow at the City University of Hong Kong from 2006 to 2009, a visiting scholar in Microsoft Research Asia in 2011, a visiting expert in NICT, Japan in 2012. He is an ACM professional member, and served as area co-chair in ACL 2017 on Tagging, Chunking, Syntax and Parsing, (senior) area chairs in ACL 2018, 2019 on Phonology, Morphology and Word Segmentation. His research interests include natural language processing and related machine learning, data mining and artificial intelligence.