Advances in Multi-turn Dialogue Comprehension: A Survey

Zhuosheng Zhang and **Hai Zhao***

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

Abstract

Training machines to understand natural language and interact with humans is an elusive and essential task in the field of artificial intelligence. In recent years, a diversity of dialogue systems has been designed with the rapid development of deep learning researches, especially the recent pre-trained language models. Among these studies, the fundamental yet challenging part is dialogue comprehension whose role is to teach the machines to read and comprehend the dialogue context before responding. In this paper, we review the previous methods from the perspective of dialogue modeling. We summarize the characteristics and challenges of dialogue comprehension in contrast to plain-text reading comprehension. Then, we discuss three typical patterns of dialogue modeling that are widely-used in dialogue comprehension tasks such as response selection and conversation question-answering, as well as dialogue-related language modeling techniques to enhance PrLMs in dialogue scenarios. Finally, we highlight the technical advances in recent years and point out the lessons we can learn from the empirical analysis and the prospects towards a new frontier of researches.

1 Introduction

Language as a means of communication tool is a bridge of people understanding each other, which is also the natural interface between human beings and machines. However, building an intelligent dialogue system that can communicate naturally and meaningfully with humans is a challenging problem towards high-level artificial intelligence, and has been drawing increasing interest from both academia and industry areas. To this end, a variety of tasks have been proposed such as response selection [Lowe *et al.*, 2015; Wu *et al.*, 2016; Zhang *et al.*, 2018], conversation-based question answering (QA) [Sun *et al.*, 2019; Reddy *et al.*, 2019; Choi *et al.*, 2018], decision making and question generation [Saeidi *et al.*, 2018].

A dialogue system usually consists of two main parts: understanding the dialogue history in natural language, and generating the response in natural language.[1] Existing studies either classify the dialogue systems into three types based on functionality: 1) task-oriented systems, 2) chat-oriented systems, 3) question answering systems [Zaib *et al.*, 2020], or categorized into 1) generative, 2) retrieval-based, 3) hybrid models according to the responding form [Wu *et al.*, 2019; Cai *et al.*, 2019; Weston *et al.*, 2018].

With the development of deep learning methods especially the recent pre-trained language models [Devlin *et al.*, 2019; Liu *et al.*, 2019; Yang *et al.*, 2019; Lan *et al.*, 2020; Clark *et al.*, 2020], traditional natural language processing (NLP) tasks, including dialogue-related tasks, have been undergoing a fast transformation, where those tasks tend to be crossed and unified in form [Zhang *et al.*, 2020b]. Among dialogue tasks, the fundamental yet challenging type is dialogue comprehension: given context, the system is required to reply or answer questions. The reply can be derived from retrieval or generation. The later question answering is known as machine reading comprehension [Sun *et al.*, 2019; Reddy *et al.*, 2019; Choi *et al.*, 2018]. Among the studies, the basic technique is dialogue modeling which focuses on how to encode the dialogue context effectively and efficiently to solve the tasks.

Technically, early studies concerning dialogue comprehension mainly focus on the matching mechanisms between the pairwise sequence of dialogue context and candidate response or question [Wu *et al.*, 2016; Zhang *et al.*, 2018; Huang *et al.*, 2019a]. Recently, inspired by the impressive performance of PrLMs, the mainstream is employing PrLMs to handle the whole pairwise texts as a linear sequence of successive tokens and implicitly capture the contextualized representations of those tokens through self-attention [Qu *et al.*, 2019; Liu *et al.*, 2020; Gu *et al.*, 2020a; Xu *et al.*, 2021a]. The word embeddings derived by these language models are pre-trained

*Corresponding author. We thank Deng Cai, Libo Qin, Yi Xu, Siru Ouyang and Junlong Li for their valuable feedback on this manuscript. 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.

[1] Although the response can be retrieved from a candidate list in retrieval-based dialogue systems, we call the process of proving the response as generating response for simplicity.

on large corpora and are then utilized as either distributed word embeddings [Peters *et al.*, 2018] or fine-tuned according to the specific task needs [Devlin *et al.*, 2019]. Besides employing PrLMs for fine-tuning, there also emerges interests in designing dialogue-motivated self-supervised tasks for pre-training.

Inspired by the recent advances above, in this survey, we review the previous studies of dialogue comprehension in the perspective of modeling the dialogue tasks as a two-stage Encoder-Decoder framework inspired by the advance of PrLMs and machine reading comprehension [Zhang *et al.*, 2020b; Zhang *et al.*, 2021b], in which way we bridge the gap between the dialogue modeling and comprehension, and hopefully benefit the future researches with the cutting-edge PrLMs. In detail, we will discuss both sides of architecture designs and the pre-training strategies. We summarize the technical advances in recent years and highlight the lessons we can learn from the empirical analysis and the prospects towards a new frontier of researches.

2 Characteristics

Compared with plain-text reading comprehension like SQuAD [Rajpurkar *et al.*, 2016], a multi-turn conversation is interactive, which involves multiple speakers, intentions, topics, thus the utterances are full of transitions.

1) The transition of speakers in conversations is in a random order, breaking the continuity as that in common non-dialogue texts due to the presence of crossing dependencies which are commonplace in a multi-party chat.

2) There may be multiple dialogue topics happening simultaneously within one dialogue history and topic drift is common and hard to detect in spoken conversations. Therefore, the multi-party dialogue appears discourse dependency relations between non-adjacent utterances, which leads up to a complex discourse structure.

3) Dialogue is colloquial, and it takes fewer efforts to speak than to write, resulting in the dialogue context rich in component ellipsis and information redundancy. However, during a conversation, the speakers cannot retract what has been said, which easily leads to self-contradiction, requiring more context, especially clarifications to fully understand the dialogue.

4) The importance of each utterance towards the expected reply is different, which makes the utterances contribute to the final response in dramatic diversity. Therefore, the order of utterance influences the dialogue modeling. In general, the latest utterances would be more critical [Zhang *et al.*, 2018; Zhang *et al.*, 2021a].

3 Methodology

3.1 Problem Formulation

In this survey, we take two typical dialogue comprehension tasks, i.e., response selection [Lowe *et al.*, 2015; Wu *et al.*, 2016; Zhang *et al.*, 2018; Cui *et al.*, 2020] and conversation-based QA [Sun *et al.*, 2019; Reddy *et al.*, 2019; Choi *et al.*, 2018], as examples to show the general technical patterns to gain insights, which would also hopefully facilitate other dialogue-related tasks.

Suppose that we have a dataset $D = \{(\mathrm{C}_i, \mathrm{X}_i; \mathrm{Y}_i)\}_{i=1}^{N}$, where $\mathrm{C}_i = \{u_{i,1}, ..., u_{i,n_i}\}$ represents the dialogue context with $\{u_{i,k}\}_{k=1}^{n_i}$ as utterances. X_i is a task-specific paired input, which can be either the candidate response R for response selection, or the question Q for conversation-based QA. Y_i denotes the model prediction.

Response Selection involves the pairwise input with R as a candidate response. The goal is to learn a discriminator $g(\cdot, \cdot)$ from D, and at the inference phase, given the context C and response R, we use the discriminator to calculate $Y = g(C, R)$ as their matching score.

Conversation-based QA aims to answer questions given the dialogue context. Let Q denotes the question Q. The goal is to learn a discriminator $g(C, Q)$ from D to extract the answer span from the context or select the right option from a candidate answer set.

Unified Encoding In the input encoding perspective, since both of the tasks share the paired inputs of either $\{C; R\}$ or $\{C; Q\}$, we simplify the formulation by focusing the response selection task, i.e., replacing R with Q can directly transform into the QA task.

3.2 Framework

As shown in Figure 1, the methods of dialogue modeling can be categorized into three patterns: 1) concatenated matching; 2) separate interaction; and 3) PrLM-based interaction.

Concatenated Matching The early methods [Kadlec *et al.*, 2015] treated the dialogue context as a whole by concatenating all previous utterances and last utterance as the context representation and then computed the matching degree score based on the context representation to encode candidate response [Lowe *et al.*, 2015]:

$$\begin{aligned}
\mathrm{EC} &= \mathrm{Encoder}(C); \\
\mathrm{ER} &= \mathrm{Encoder}(R); \\
\mathrm{Y} &= \mathrm{Decoder}(\mathrm{EC}; \mathrm{ER});
\end{aligned} \tag{1}$$

where Encoder is used to encode the raw texts into contextualized representations. Decoder is the module that transforms the contextualized representations to model predictions (Y), which depends on the tasks. For response selection, it can be the attention-based module that calculate the matching score between EC and ER.

Separate Interaction With the bloom of attention-based pairwise matching mechanisms, researches soon find it effective by calculating different levels of interactions between the dialogue context and response. The major research topic is how to improve the semantic matching between the dialogue context and candidate response. For example, Zhou *et al.* [2016] performed context-response matching with a multi-view model on both word level and utterance level. Wu *et al.* [2016] improved the leveraging of utterances relationship and contextual information by matching a response with each utterance in the context. Those methods can be unified

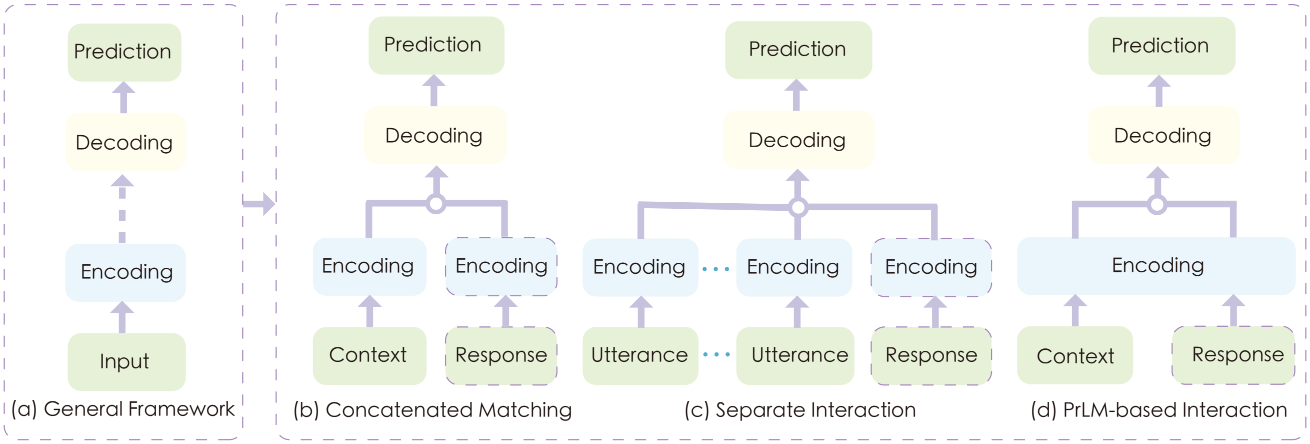

Figure 1: Dialogue Modeling framework. The dispensable parts are marked in dashed lines.

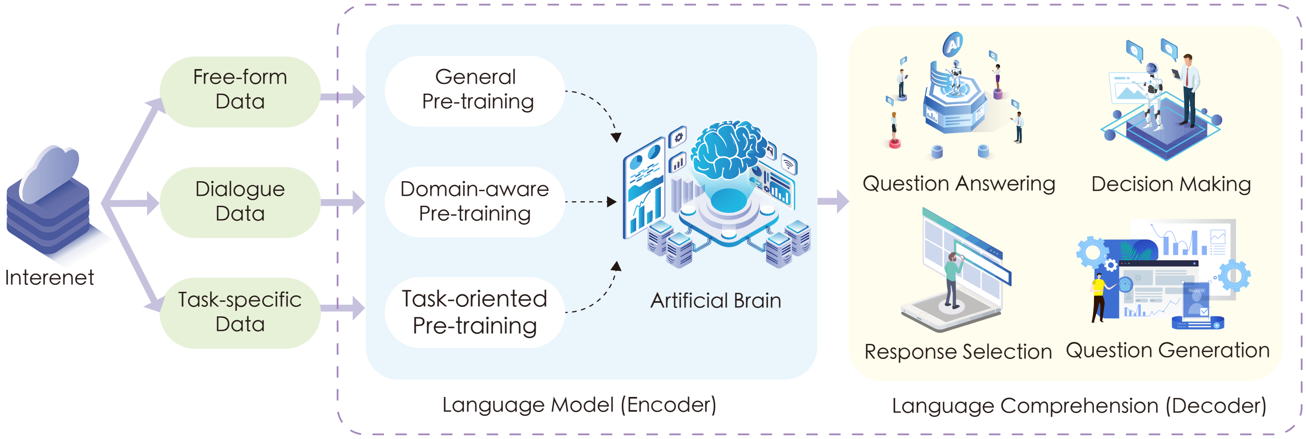

Figure 2: Dialogue-related Language Modeling.

by the view similar to the above concatenated matching:

$$\begin{aligned}
&\mathrm{EU}_i = \mathrm{Encoder}(u_i);\\
&\mathrm{ER} = \mathrm{Encoder}(R);\\
&\mathrm{I} = \mathrm{ATT}([\mathrm{EU}_1, \ldots, \mathrm{EU}_n]; \mathrm{ER});\\
&\mathrm{Y} = \mathrm{Decoder}(\mathrm{I});
\end{aligned} \tag{2}$$

where ATT denotes the attention-based interactions, which can be pairwise attention, self attention, or the combinations.

PrLM-based Interaction PrLMs handle the whole input text as a linear sequence of successive tokens and implicitly capture the contextualized representations of those tokens through self-attention [Devlin *et al.*, 2019]. Given the context C and response R, we concatenate all utterances in the context and the response candidate as a single consecutive token sequence with special tokens separating them, and then encode the text sequence by a PrLM:

$$\begin{aligned}
&\mathrm{EC} = \mathrm{Encoder}(\texttt{[CLS]}\,C\,\texttt{[SEP]}\,R\,\texttt{[SEP]});\\
&\mathrm{Y} = \mathrm{Decoder}(\mathrm{EC});
\end{aligned} \tag{3}$$

where [CLS] and [SEP] are special tokens.

3.3 Dialogue-related Language Modeling

Although the PrLMs demonstrate superior performance due to their strong representation ability from self-supervised pre-training, it is still challenging to effectively adapt task-related knowledge during the detailed task-specific training which is usually in a way of fine-tuning [Gururangan *et al.*, 2020]. Generally, those PrLMs handle the whole input text as a linear sequence of successive tokens and implicitly capture the contextualized representations of those tokens through self-attention. Such fine-tuning paradigm of exploiting PrLMs would be suboptimal to model dialogue task which holds exclusive text features that plain text for PrLM training may hardly embody.

Besides, pre-training on general corpora has critical limitations if task datasets are highly domain-specific [Whang *et al.*, 2019], which cannot be sufficiently and accurately covered by the learned universal language representation. Thus, some researchers have tried to let the task-specific factors involved in pre-training in advance by further pre-training PrLMs with general objectives like MLM and NSP on in-domain texts. For example, BioBERT [Lee *et al.*,

Models	Ubuntu			Douban						E-commerce		
	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
Single-turn models with concatenated matching												
CNN	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647	0.328	0.515	0.792
LSTM	0.638	0.784	0.949	0.485	0.537	0.320	0.187	0.343	0.720	0.365	0.536	0.828
BiLSTM	0.630	0.780	0.944	0.479	0.514	0.313	0.184	0.330	0.716	0.365	0.536	0.825
MV-LSTM	0.653	0.804	0.946	0.498	0.538	0.348	0.202	0.351	0.710	0.412	0.591	0.857
Match-LSTM	0.653	0.799	0.944	0.500	0.537	0.345	0.202	0.348	0.720	0.410	0.590	0.858
Multi-turn matching network with separate interaction												
Multi-View	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729	0.421	0.601	0.861
DL2R	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705	0.399	0.571	0.842
SMN	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	0.526	0.727	0.933
MRFN	0.786	0.886	0.976	0.571	0.617	0.448	0.276	0.435	0.783	-	-	-
IMN	0.794	0.889	0.974	0.570	0.615	0.433	0.262	0.452	0.789	0.621	0.797	0.964
IoI	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
MSN	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
G-MSN	0.812	0.911	0.987	0.599	0.645	0.476	0.308	0.468	0.826	0.613	0.786	0.964
PrLM-based methods for fine-tuning												
BERT	0.808	0.897	0.975	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
BERT-SS-DA	0.813	0.901	0.977	0.602	0.643	0.458	0.280	0.491	0.843	0.648	0.843	0.980
TADAM	0.821	0.906	0.978	0.594	0.633	0.453	0.282	0.472	0.828	0.660	0.834	0.975
PoDS	0.828	0.912	0.981	0.598	0.636	0.460	0.287	0.468	0.845	0.633	0.810	0.967
ELECTRA	0.845	0.919	0.979	0.599	0.643	0.471	0.287	0.474	0.831	0.607	0.813	0.960
MDFN	0.866	0.932	0.984	0.624	0.663	0.498	0.325	0.511	0.855	0.639	0.829	0.971
Dialogue-related language modeling												
BERT	0.851	0.924	0.984	-	-	-	-	-	-	-	-	-
BERT-VFT	0.858	0.931	0.985	-	-	-	-	-	-	-	-	-
SA-BERT	0.855	0.928	0.983	0.619	0.659	0.496	0.313	0.481	0.847	0.704	0.879	0.985
PoDS	0.856	0.929	0.985	0.599	0.637	0.460	0.287	0.469	0.839	0.671	0.842	0.973
DCM	0.868	0.936	0.987	0.611	0.649	-	0.294	0.498	0.842	0.685	0.864	0.982
UMS$_{\text{BERT}}$	0.875	0.942	0.988	0.625	0.664	0.499	0.318	0.482	0.858	0.762	0.905	0.986
BERT-SL	0.884	0.946	0.990	-	-	-	-	-	-	0.776	0.919	0.991
ELECTRA	0.861	0.932	0.985	0.612	0.655	0.480	0.301	0.499	0.836	0.673	0.835	0.974
UMS$_{\text{ELECTRA}}$	0.875	0.941	0.988	0.623	0.663	0.492	0.307	0.501	0.851	0.707	0.853	0.974

Table 1: Results on Ubuntu, Douban, and E-commerce datasets.

2020], SciBERT [Beltagy *et al.*, 2019], and Clinical-BERT [Huang *et al.*, 2019b] pre-train BERT further on texts of Biomedicine, Science and Clinical-Medicine respectively. DialoGPT [Zhang *et al.*, 2020a] pre-trains GPT on a large in-domain dialogue corpus, Reddit. Gururangan *et al.* [2020] further demonstrate the effectiveness of task-specific pre-training by concluding two sub-classes: Domain Adaptive Pre-training and Task-Adaptive Pre-training.

Despite the success of the above previous studies simply putting task-independent and in-domain task-related corpora together for pre-training, the guideline of task-specific training objective is not well exploited. It is obvious that texts of different domains and forms have different emphases. As our work lays emphasis on dialogue comprehension tasks, which is more complex than other forms of texts like sentence-pairs or essays, the corresponding training objective should be very carefully designed to fit the important elements of dialogues. As shown in Figure 2, there are three kinds of dialogue-related language modeling strategies, namely, *general-purpose pre-training*, *domain-aware pre-training*, and *task-oriented pre-training*, among which are self-supervised methods that do not require additional annotation and can be easily applied into existing approaches.[2]

General-purpose Pre-training As the standard pre-training procedure, PrLMs are pre-trained on large-scale domain-free texts and then used for fine-tuning according to the specific task needs. There are token-level and sentence-level objectives used in the general-purpose pre-training. BERT [Devlin *et al.*, 2019] adopts Masked Language Modeling (MLM) as its pre-training objective. It first masks out some tokens from the input sentences and then trains the model to predict them by the rest of the tokens. There are derivatives of MLM like Permuted Language Modeling (PLM) in XLNet [Yang *et al.*, 2019] and Sequence-to-Sequence MLM (Seq2Seq MLM) in MASS [Song *et al.*, 2019] and T5 [Raffel *et al.*, 2019]. Next Sentence Prediction (NSP) is another widely used pre-training objective, which

[2] Though general-purpose pre-training is not our major focus, we describe it here as the basic knowledge for completeness.

Model	**R$_4$@1**	**R$_4$@2**	**MRR**
TF-IDF	0.279	0.536	0.542
Dual LSTM	0.260	0.491	0.743
SMN	0.299	0.585	0.595
DAM	0.241	0.465	0.518
GPT-2	0.332	0.602	0.584
GPT-2-FT	0.392	0.670	0.629
BERT	0.648	0.847	0.795
RoBERTa	0.713	0.892	0.836
RoBERTa + OCN	0.867	0.958	0.926
ALBERT	0.847	0.962	0.916
GRN-v2	0.915	0.983	0.954
ELECTRA	0.900	0.979	0.946
MDFN	0.916	0.984	0.956
ELECTRA + DAPO	0.916	0.988	0.956

Table 2: Results on MuTual dataset. The upper and lower blocks present the models w/o and w/ PrLMs, respectively.

Model	**Accuracy**
Stanford Attentive Reader	39.8
Gated-Attention Reader	41.3
Word Matching	42.0
Sliding Window (SW)	42.5
Distance-Based Sliding Window	44.6
+ Dialogue Structure and ConceptNet Embedding	50.1
Co-Matching	45.5
Finetuned Transformer LM	55.5
+ Speaker Embedding	57.4
EER + FT	57.7
BERT-Large + WAE	69.0
RoBERTa-Large + MMM	88.9
ALBERT-xxlarge + DUMA	90.4
+ Multi-Task Learning	91.8

Table 3: Results (%) on DREAM dataset. The upper and lower blocks present the models w/o and w/ PrLMs, respectively.

trains the model to distinguish whether two input sentences are continuous segments from the training corpus. Sentence Order Prediction (SOP) is one of the replacements of NSP. It requires models to tell whether two consecutive sentences are swapped or not and is first used in ALBERT [Lan *et al.*, 2020]. Replaced Token Detection (RTD) is also used by recent PrLMs like ELECTRA [Clark *et al.*, 2020] with a similar idea used in Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014], which requires models to predict whether a token is replaced given its surrounding context.

Domain-aware Pre-training The original PrLMs are trained on a large text corpus to learn general language representations. To incorporate specific in-domain knowledge, adaptation on in-domain corpora, also known as domain-aware pre-training, is designed, which directly employs the original PrLMs as mentioned in the general-purpose paragraph above, using the dialogue-domain corpus [Whang *et al.*, 2020; Wu *et al.*, 2020]. The most widely-used PrLM for domain-adaption in the dialogue field is BERT [Devlin *et al.*, 2019], whose pre-training is based on two loss functions: (1) a next sentence prediction (NSP) loss, and (2) a masked language model (MLM) loss. Although NSP has been shown trivial in RoBERTa [Liu *et al.*, 2019] during general-purpose pre-training, it yields surprising gains in dialogue scenarios [Li *et al.*, 2020b]. The most plausible reason is that dialogue emphasizes the relevance between dialogue context and the subsequent response, which shares a similar goal with NSP.

Task-oriented Pre-training In contrast to the plain-text modeling as the focus of the PrLMs, dialogue texts involve multiple speakers and reflect special characteristics such as topic transitions and structure dependencies between distant utterances. Inspired by such phenomenon, recent studies are pondering the dialogue-specific training objectives to model dialogue-related features. Prior works have indicated that the order information would be important in the text representation, and the well-known next-sentence-prediction [Devlin *et al.*, 2019] and sentence-order-prediction [Lan *et al.*, 2020] can be viewed as special cases of order prediction. Especially in the dialogue scenario, predicting the word order of utterance, as well as the utterance order in the context, has shown effectiveness in the dialogue modeling task [Kumar *et al.*, 2020; Gu *et al.*, 2020b], where the utterance order information is well restored from shuffled dialogue context. Li *et al.* [2021] designed a variant of NSP called next utterance prediction as a pre-training scheme to adapt BERT to accommodate the inherent context continuity underlying the multi-turn dialogue. Whang *et al.* [2021] proposed various utterance manipulation strategies including utterance insertion, deletion, and search to maintain dialog coherence. Xu *et al.* [2021a] introduced four self-supervised tasks including next session prediction, utterance restoration, incoherence detection, and consistency discrimination, and jointly trained the PLM-based response selection model with these auxiliary tasks in a multi-task manner, to capture a better local optimum and produce better dialogue-aware features, such as coherence and consistency.

3.4 Empirical Analysis

Tables 1, 2 and 3 present the benchmark results on five typical dialogue comprehension tasks, including three response selection tasks, Ubuntu [Lowe *et al.*, 2015], Douban [Wu *et al.*, 2016], ECD [Zhang *et al.*, 2018], Mutual [Cui *et al.*, 2020] and one conversation-based QA task, DREAM [Sun *et al.*, 2019], from which we summarize the following observations:[3]

1) In the early stage without PrLMs, separate interaction commonly achieves better performance than the simple concatenated matching, verifying the effectiveness of attention-based pairwise matching.

2) Generally, the previous models based on multi-turn matching networks (separate interaction) perform worse than simple PrLMs-based ones, illustrating the power of contextualized representations in context-sensitive dialogue modeling.

[3] The evaluation results are collected from published literature [Zhang *et al.*, 2021a; Whang *et al.*, 2021; Xu *et al.*, 2021a; Lin *et al.*, 2020; Lowe *et al.*, 2015; Liu *et al.*, 2021b; Liu *et al.*, 2021a; Li *et al.*, 2020b; Sun *et al.*, 2019; Wan, 2020]

3) Compared with general-purpose PrLMs, dialogue-aware pre-training (e.g., BERT-VFT, SA-BERT, PoDS) can further improve the results by a large margin. In addition, task-oriented pre-training (e.g., DCM, UMS, BERT-SL) even shows superiority among the pre-training techniques.

4) Empirically, for the concerned dialogue comprehension tasks, retrieval-based or discriminative methods commonly show better performance than generative models such as GPT.

5) Among the models, G-MSN, BERT-SS-DA, ELECTRA+DAPO show that training/pre-training data construction, especially negative sampling is a critical influence factor to the model performance. In addition, SA-BERT and MDFN indicate that modeling the speaker information is also effective for dialogue modeling.

4 Trends and Prospects

The recent mainstream work of dialogue comprehension commonly adopts the PrLMs as an encoder to represent the dialogue contexts coarsely, dealing with the pairwise dialogue context and candidate response or question as a whole [Qu *et al.*, 2019; Gu *et al.*, 2020a; Li *et al.*, 2020a]. However, how to effectively adapt PrLMs with dialogue texts still remains a challenge. For example, multi-party multi-turn dialogue modeling launches new challenges of speaker role transition and complex discourse structure. In addition, grounding the dialogue with adequate background knowledge, and interacting with people with the ability of multilingual and multimodal conversation also deserves further exploration.

4.1 Dialogue Context Decoupling

Recent widely-used PrLM-based models deal with the whole dialogue,[4] which results in entangled information that originally belongs to different parts and is not optimal for dialogue modeling. Sequence decoupling is a strategy to tackle this problem by explicitly separating the context into different parts and further constructing the relationships between those parts to yield more fine-grained representations. One possible solution is splitting the context into different continuous topic blocks [Xu *et al.*, 2021b; Lu *et al.*, 2020]. However, there existing topic crossing, which would hinder the segmentation effect. Another scheme is to employ a masking mechanism inside self-attention network [Liu *et al.*, 2021a], to limit the focus of each word only on the related ones, such as those from the same utterance, or the same speaker, to model the local dependencies, in complement with the global contextualized representation from PrLMs.

Training machines to understand dialogue has been shown much more challenging than the common MRC as every utterance in dialogue has an additional property of speaker role, which breaks the continuity in common non-dialogue texts due to the presence of discourse dependencies which are commonplace in dialogue history [Li *et al.*, 2020a]. Gu *et al.* [2020a] proposed Speaker-Aware BERT for two-party dialogue tasks by organizing utterances according to *spoken-from speaker* and *spoken-to speaker* and adding a speaker embedding at token representation stage. Recent studies show that explicitly modeling discourse segmentation and graph-like dependencies [Ouyang *et al.*, 2020] would be effective for improving dialogue comprehension.

4.2 Background Knowledge Grounding

There are various kinds of background knowledge that can be grounded in dialogue modeling, including commonsense items from knowledge graphs to strengthen reasoning ability [Zhang *et al.*, 2021a], persona-based attributes such as speaker identity, dialogue topic, speaker sentiments, to enrich the dialogue context [Olabiyi *et al.*, 2019], scenario information to provide the dialogue background [Ouyang *et al.*, 2020], etc.

4.3 Dialogue-aware Language Modeling

Recent studies have indicated that dialogue-related language modeling can enhance dialogue comprehension substantially [Gu *et al.*, 2020a; Li *et al.*, 2021; Zhang *et al.*, 2021a; Whang *et al.*, 2021; Xu *et al.*, 2021a]. However, these methods rely on the dialogue-style corpus for the pre-training, which is not always available in general application scenarios. Given the massive free-form and domain-free data from the internet, how to simulate the conversation, e.g., in an adversarial way, with the general-purpose and general-domain data is a promising research direction. Besides transferring from general-purpose to dialogue-aware modeling, multi-domain adaption is another important topic which is effective to reduce the annotation cost and achieve robust and scalable dialogue systems [Qin *et al.*, 2020b].

4.4 Reliable Data Construction

Most prior works train the dialogue comprehension models with training data constructed by a simple heuristic. They treated human-written responses as positive examples and randomly sampled responses from other dialogue contexts as equally-bad negative examples, i.e., inappropriate responses [Lowe *et al.*, 2015; Wu *et al.*, 2016; Zhang *et al.*, 2018]. As discussed in Section 3.4, data construction is also critical to the model capacity. The randomly sampled negative responses are often too trivial that making the model lack the strength to handle strong distractors for dialogue comprehension. To train a more effective and reliable model, there is an emerging interest in mining better training data [Lin *et al.*, 2020; Li *et al.*, 2020b; Su *et al.*, 2020].

4.5 Multilingual and Multimodal Dialogue

As a natural interface of human and machine interaction, a dialogue system would beneficial for people in different language backgrounds to communicate with each other. Besides the natural language texts, visual and audio sources are also effective carriers are can be incorporated with texts for comprehensive and immersed conversations. With the rapid development of multilingual and multimodal researches [Qin *et al.*, 2020a; Firdaus *et al.*, 2021], building an intelligent dialogue system is not elusive in the future.

[4] Because PrLMs are interaction-based methods, thus encoding the context as a whole achieves better performance than encoding the utterances individually.

5 Conclusion

In this survey, we conduct a comprehensive overview of the recent advances in multi-turn dialogue comprehension, including task background, characteristics, methodology, dialogue-related language modeling, empirical analysis. We summarize the existing dialogue modeling methods and dialogue-related language modeling techniques into three general patterns, respectively. According to the empirical analysis on the typical dialogue comprehension tasks, we demonstrate our observations and lessons from different stages of dialogue comprehension studies. Despite the latest achievements, we highlight the recent trends and suggest several possible future research directions.

References

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics.

[Cai *et al.*, 2019] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. Skeleton-to-response: Dialogue generation guided by retrieval memory. 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 1219–1228, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[Choi *et al.*, 2018] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, 2018. Association for Computational Linguistics.

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

[Cui *et al.*, 2020] 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, Online, 2020. Association for Computational Linguistics.

[Devlin *et al.*, 2019] 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, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[Firdaus *et al.*, 2021] Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. Aspect-aware response generation for multimodal dialogue system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33, 2021.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[Gu *et al.*, 2020a] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM, 2020.

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

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

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

[Huang *et al.*, 2019b] Kexin Huang, Jaan Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*, 2019.

[Kadlec *et al.*, 2015] Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *NIPS Workshop*, 2015.

[Kumar *et al.*, 2020] Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. Deep attentive ranking networks for learning to order sentences. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8115–8122. AAAI Press, 2020.

[Lan *et al.*, 2020] 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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Lee *et al.*, 2020] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020.

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

[Li *et al.*, 2020b] Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. Task-specific Objectives of Pre-trained Language Models for Dialogue Adaptation. *arXiv: 2009.04984*, 2020.

[Li *et al.*, 2021] Lu Li, Chenliang Li, and Donghong Ji. Deep context modeling for multi-turn response selection in dialogue systems. *IPM*, 2021.

[Lin *et al.*, 2020] 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, Online, 2020. Association for Computational Linguistics.

[Liu *et al.*, 2019] 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: 1907.11692*, 2019.

[Liu *et al.*, 2020] Chuang Liu, Deyi Xiong, Yuxiang Jia, Hongying Zan, and Changjian Hu. Hisbert for conversational reading comprehension. In *IALP*, 2020.

[Liu *et al.*, 2021a] Longxiang Liu, Zhuosheng Zhang, , Hai Zhao, Xi Zhou, and Xiang Zhou. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue. In *AAAI*, 2021.

[Liu *et al.*, 2021b] Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. A graph reasoning network for multi-turn response selection via customized pre-training. In *AAAI*, 2021.

[Lowe *et al.*, 2015] Ryan Lowe, Nissan Pow, Iulian 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*, pages 285–294, Prague, Czech Republic, 2015. Association for Computational Linguistics.

[Lu *et al.*, 2020] Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. Improving contextual language models for response retrieval in multi-turn conversation. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1805–1808. ACM, 2020.

[Olabiyi *et al.*, 2019] Oluwatobi Olabiyi, Anish Khazane, Alan Salimov, and Erik Mueller. An adversarial learning framework for a persona-based multi-turn dialogue model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 1–10, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[Ouyang *et al.*, 2020] Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. Dialogue graph modeling for conversational machine reading. *arXiv:2012.14827*, 2020.

[Peters *et al.*, 2018] 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, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[Qin *et al.*, 2020a] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org, 2020.

[Qin *et al.*, 2020b] Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354, 2020.

[Qu *et al.*, 2019] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. BERT with history answer embedding for conversational question answering. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1133–1136. ACM, 2019.

[Raffel *et al.*, 2019] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv: 1910.10683*, 2019.

[Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics.

[Reddy *et al.*, 2019] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[Saeidi *et al.*, 2018] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium, 2018. Association for Computational Linguistics.

[Song *et al.*, 2019] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR, 2019.

[Su *et al.*, 2020] Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. Dialogue response selection with hierarchical curriculum learning. *arXiv preprint arXiv:2012.14756*, 2020.

[Sun *et al.*, 2019] 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.

[Wan, 2020] Hui Wan. Multi-task learning with multi-head attention for multi-choice reading comprehension. *arXiv:2003.04992*, 2020.

[Weston *et al.*, 2018] Jason Weston, Emily Dinan, and Alexander Miller. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium, 2018. Association for Computational Linguistics.

[Whang *et al.*, 2019] T. Whang, Dongyub Lee, C. Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. An effective domain adaptive post-training method for bert in response selection. In *INTERSPEECH*, 2019.

[Whang *et al.*, 2020] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. An effective domain adaptive post-training method for bert in response selection. *INTERSPEECH*, 2020.

[Whang *et al.*, 2021] Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *AAAI*, 2021.

[Wu *et al.*, 2016] Yu Wu, Wei Wu, Ming Zhou, and Zhoujun Li. Sequential Match Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. *ACL*, 2016.

[Wu *et al.*, 2019] Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7281–7288. AAAI Press, 2019.

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

[Xu *et al.*, 2021a] Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *AAAI*, 2021.

[Xu *et al.*, 2021b] Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue modeling. In *AAAI*, 2021.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.

[Zaib *et al.*, 2020] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *ACSW*, 2020.

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

[Zhang *et al.*, 2020a] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng

Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, 2020. Association for Computational Linguistics.

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

[Zhang *et al.*, 2021a] Zhuosheng Zhang, Junlong Li, and Hai Zhao. Multi-turn dialogue reading comprehension with pivot turns and knowledge. *TASLP*, 2021.

[Zhang *et al.*, 2021b] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. Retrospective reader for machine reading comprehension. In *AAAI*, 2021.

[Zhou *et al.*, 2016] 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*, pages 372–381, Austin, Texas, 2016. Association for Computational Linguistics.