A Survey on Response Selection for Retrieval-based Dialogues

Chongyang Tao[1] , **Jiazhan Feng**[2] , **Rui Yan**[2] , **Wei Wu**[3] and **Daxin Jiang**[1*]

[1]Microsoft Corporation
[2]Peking University
[3]Meituan Corporation

{chongyang.tao, djiang}@microsoft.com, {fengjiazhan, ruiyan}@pku.edu.cn, wuwei@meituan.com

Abstract

Building an intelligent dialogue system capable of naturally and coherently conversing with humans has been a long-standing goal of artificial intelligence. In the past decade, with the development of machine/deep learning technology and the explosive growth of available conversation data in social media, numerous neural models have been developed for context-response matching tasks in retrieval-based dialogue systems, with more fluent and informative responses compared with generative models. This paper presents a comprehensive survey of recent advances in response selection for retrieval-based dialogues. In particular, we first formulate the problem of response selection and review state-of-the-art context-response matching models categorized by their architecture. Then we summarize some recent advances on the research of response selection, including incorporation with extra knowledge and exploration on more effective model learning. Finally, we highlight the challenges which are not yet well addressed in this task and present future research directions.

1 Introduction

Building a smart dialogue system that can converse with humans naturally and meaningfully has long been an attractive but challenging task in artificial intelligence. Early rule-based dialogue systems, such as Eliza, although helpful in improving machine intelligence, could only respond in a limited space. Recently, the flourish of social networking services has accumulated a great number of conversation data among humans on the Web, and thus encourages researchers to investigate data-driven approaches to building open-domain dialogue systems. Existing studies can be generally categorized into generation-based methods [Sankar *et al.*, 2019] or retrieval-based methods [Lowe *et al.*, 2015; Whang *et al.*, 2020]. The former directly synthesize a response via natural language generation techniques, and the latter retrieves a number of response candidates from a pre-built index, and then selects an appropriate one as a response.

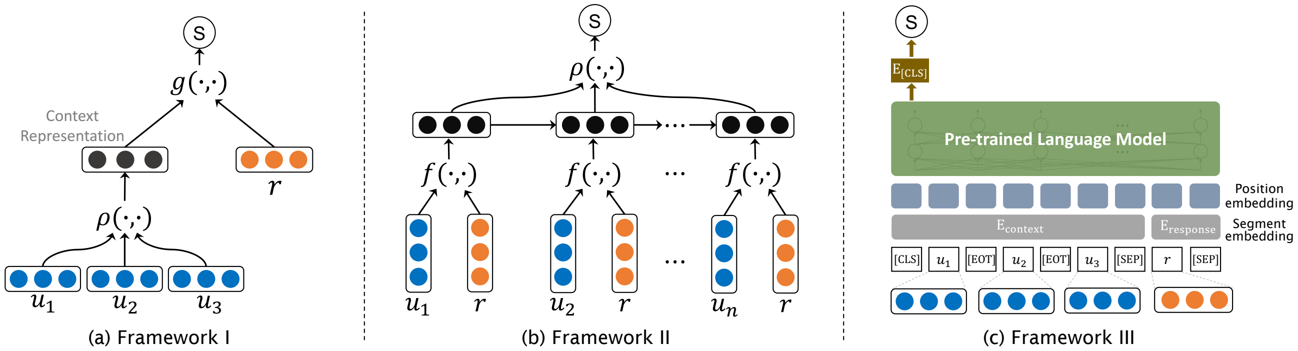
Among the effort, retrieval-based methods re-use the existing human conversations and select a response for new input from a bunch of candidates. They are often superior to the generation-based counterparts on response fluency and informativeness, are easy to evaluate, and have powered some real products such as the social bot XiaoIce from Microsoft. Moreover, the achievements on information retrieval technologies, such as the research on learning to rank and learning to match methodologies during the evolution of modern search engines, and the advances in neural representation learning and pre-trained methods [Devlin *et al.*, 2019] also lays a solid technical foundation for retrieval-based dialogue systems. Besides, AAAI also organizes Dialog System Technology Challenges (DSTC)[1] for evaluating the performance of response selection for retrieval-based dialogues annually, attracting a large number of researchers and greatly promoting the development of retrieval-based conversation models.

Although such systems borrow a lot from the design of search engines, the task of response selection raises new challenges for the community of information retrieval: 1) Conversation data is often in a hierarchical structure and renders more complicated semantics; 2) Conversational responses are often expressed in short and informal forms with semantics context-dependent; 3) In addition to semantic relevance, the logical consistency and coherence of response should also be considered; 4) There is a lack of high-quality training data and useful meta-signals, like the click-through data in Web search, that can effectively aid the learning of response selection models. Consequently, response retrieval for open-domain dialogues is emerging as a hot and challenging topic in the interdisciplinary research of human-computer interaction and information retrieval. Here, we provide a unified view to neural models for response selection.

The paper starts with a brief introduction to the architecture of the response selection model for retrieval-based dialogue systems. Then, we present three frameworks including representation-based Models, interaction-based Models, and PLM-based Models that subsume most of the existing models as special cases and unveil the common behind the different structures. On top of a summary of benchmarks for response selection, we report a thorough comparison among representative models of the frameworks. Then we extend the response

[*] Corresponding author.

[1] http://workshop.colips.org/dstc7/workshop.html



(a) Framework I

(b) Framework II

(c) Framework III

Figure 1: Three context-response matching frameworks: (a) representation-based framework; (b) interaction-based framework; (c) PLM-based framework.

selection task to dialogues with additional knowledge since knowledge is an awareness and understanding of the input text and its surrounding dialogue context. We introduce various response ranking models dealing with different categories of knowledge including topic words, visual information, background documents, dialogue acts, and emotions. After that, we summarize some recent exploration on more effective learning methods for the response selection task. The paper ends up with discussions on some open challenges on the research of retrieval-based dialogues.

2 Response Selection Models

In this section, we first formulate the problem of response selection for retrieval-based dialogue and then review existing context-response matching models categorized by their architecture. Finally, we compare three frameworks.

2.1 Task Formalization

Given a sequence of utterance $\mathcal{C} = \{u_1, u_2, \ldots, u_n\}$ as conversation history, where $n \geq 1$[2] and $\{u_i\}_{i=1}^{n}$ are arrayed in a temporal order, retrieval-based dialogue models selects a proper response r from a bunch of candidates $\{r_i\}_{i=1}^{m}$ which are often retrieved from an index of existing human conversations. Since index building and pre-retrieval methods have been well studied in information retrieval area. Thus, the core problem of this research lies in building a context-response matching model $s(\mathcal{C}, r)$ that measures how likely a candidate r_i to be a proper response after $\mathcal{C}$. The learning of $s(\cdot, \cdot)$ needs supervision, and is often performed with a set of triples $\{(y_i, \mathcal{C}_i, r_i)\}_{i=1}^{N}$ where y_i is a (binary) label indicating the matching degree between $\mathcal{C}_i$ and r_i.

Conventional context-response matching models for response selection generally fall in two frameworks, namely *representation-based matching framework* and *interaction-based matching framework*. Particularly, the former performs matching based on sentence embedding while the latter performs matching based on context-response interaction. More recently, researchers begin to explore *PLM-based matching framework* that builds the context-response matching models on the basis of the pre-trained language models (PLMs) due to their strong representation and understanding capability. Next, We will illustrate the architecture of each framework, and summarize some representative models in each framework.

2.2 Framework I: Representation-based Models

The framework I usually follows a representation-matching paradigm and consists of a representation layer and a matching layer, as shown in Figure 1 (a). In the representation layer, the context message c and the candidate response r are individually represented as vectors by a representation function $\phi(\cdot)$. Then an aggregation function $\rho(\cdot, \cdot)$ is applied to fuse the representations of all utterances in the context into a context-level vector[3]. Finally, the matching layer uses a matching function $g(\cdot, \cdot)$ to calculate the final matching score from the two vectors. The implementation of the framework includes the definition of the representation function, aggregation function, and matching function, and different matching models can be constructed according to different functions.

For sentence representation, hand-crafted features can be used, such as sentence length [Wang *et al.*, 2017], the number of common words [Wang *et al.*, 2017], TD-IDF [Kang *et al.*, 2014], topic features [Wu *et al.*, 2018a], and dependency tree [Wang *et al.*, 2015], etc. In recent years, with the widespread rise of neural representation technology, researchers usually employ neural-based methods to represent the context or response candidate, such as pooling [Yan *et al.*, 2018], convolutional neural network (CNN) [Hu *et al.*, 2014; Yan *et al.*, 2018; Wu *et al.*, 2018a], recurrent neural network (RNN) [Lowe *et al.*, 2015; Yan *et al.*, 2016], or self-attention network [Humeau *et al.*, 2020]. The aggregation function $\rho(\cdot, \cdot)$ can also be defined as Pooling, CNN or RNN. The similarity function $g(\cdot, \cdot)$ can be defined as euclidean distance [Yan *et al.*, 2018], dot function [Humeau *et al.*, 2020], bi-linear function [Lowe *et al.*, 2015], multi-layer perceptron (MLP) [Hu *et al.*, 2014] and neural tensor function [Wu *et al.*, 2018a]. Table 1 summarizes the details of existing models under the representation-based matching framework.

[2] The research of retrieval-based dialogues starts from a single-turn assumption where $n = 1$, but now focuses on a more natural multi-turn assumption where $n > 1$.

[3] Especially, the scenario is known as the single-turn conversation when only one utterance is kept in the context. The aggregation function $\rho(\cdot, \cdot)$ is equivalent with an unit function in the single-turn scenario.

Representation-based Models

Model	Representation Function $\phi(\cdot)$	Aggregation Function $\rho(\cdot)$	Matching Function $g(\cdot)$
ARC-I [Hu *et al.*, 2014]	CNN	-	MLP
DeepMATCH [Wang *et al.*, 2015]	Dependency tree	-	Nonlinear function
Dual-LSTM [Lowe *et al.*, 2015]	LSTM	-	Bilinear function
DocChat [Yan *et al.*, 2018]	CNN+Attention Pooling	-	Euclidean distance
TACNTN [Wu *et al.*, 2018a]	CNN	-	Neural tensor function
DL2R [Yan *et al.*, 2016]	BiLSTM+CNN	Identity	MLP
Multi-View [Zhou *et al.*, 2016]	Word2vec/CNN	GRU	Bilinear function
TADAM [Xu *et al.*, 2020]	Transformer	GRU	MLP

Interaction-based Models

Model	Encoding Function $\psi(\cdot)$	Interaction Function $f(\cdot,\cdot) \Rightarrow \eta(\cdot)$	Aggregation Function $\rho(\cdot)$
DeepMatch$_{topic}$ [Lu and Li, 2013]	Topic feature	Similarity-based ⇒MLP	-
ARC-II [Hu *et al.*, 2014]	CNN	Similarity-based ⇒ CNN	-
KEHNN [Wu *et al.*, 2018c]	Word2vec & BiGRU	Similarity-based ⇒ CNN	-
CSRAN [Tay *et al.*, 2018]	BiLSTM	Attention-based⇒BiLSTM+Pooling	-
ESIM [Chen and Wang, 2019]	BiLSTM	Attention-based ⇒ BiLSTM+Pooling	-
Poly-encoders [Humeau *et al.*, 2020]	BERT	Attention-based ⇒ Pooling	-
SMN [Wu *et al.*, 2017]	Word2vec/GRU	Similarity-based ⇒ CNN	GRU
DAM [Zhou *et al.*, 2018b]	Transformer/Cross-attention	Similarity-based ⇒ CNN	CNN
DUA [Zhang *et al.*, 2018b]	GRU+attention	Similarity-based ⇒ CNN	GRU
MRFN [Tao *et al.*, 2019a]	Word2vec/CNN/RNN/Transformer	Attention-based ⇒ GRU	GRU
IoI [Tao *et al.*, 2019b]	Transformer	Attention-based ⇒ LSTM	LSTM
MSN [Yuan *et al.*, 2019]	BiLSTM	Attention-based ⇒ BiLSTM	BiLSTM

Table 1: A summary of conventional context-response matching models based on representation-based framework and interaction-based framework. Gray rows mean that the models are single-turn settings.

2.3 Framework II: Interaction-based Models

Unlike the representation-based matching framework where the interaction between context and response occurs at the last stage, the interaction-based matching framework allows the context and response candidate to interact with each other at the beginning. As shown in Figure 1 (b), the framework usually follows a representation-matching-aggregation paradigm. Firstly, an encoding function $\psi(\cdot)$ (e.g., unit function[4], RNN, CNN, and self-attention networks, etc.) is employed to encode two input sentences, resulting in the representation matrices of the context and the response with each column of the matrix a word representation. Then an interaction function $f(\cdot)$ is used to calculate the interaction between the two representation matrices and an interaction feature matrix is obtained. After that, a feature extraction function $\eta(\cdot)$ converts the interaction matrix into a matching vector. Finally, an aggregation function $\rho(\cdot)$ is utilized to model the sequential relationship among a sequence of matching features in the context, and the matching score can be calculated from the aggregated feature through a nonlinear transformation.

There are two main types of definitions for the interaction function $f(\cdot)$ in the interaction-based single-turn matching framework. The first type of interaction function is *similarity-based approach* [Lu and Li, 2013; Wu *et al.*, 2017]. By calculating the similarity of each word pair between the context message and the response candidate, a similarity matrix is generated as an interactive representation; The second interaction function is *attention-based approach* [Tao *et al.*, 2019a; Yuan *et al.*, 2019]. In this type, the context message attends to the candidate response through the attention mechanism. For each word in the response, the attention weight is first calculated according to the similarity between the current word and each word of the context message, then the word vectors of the context are linearly combined with the weights as a new attention representation. Finally, the word vector of the response is merged with its corresponding attention representation to obtain an interaction-based representation. Each word in the response corresponds to such an interaction-based vector, and all these vectors constitute the final interaction matrix. The core idea of these two approaches is to perform word-level matching at first and then convert word-level matching features into sentence-level matching features. In addition to the above two common interaction methods, Hu *et al.* (2014) use a convolution operation to interact with each word vectors of the context and response, and then convert the interaction matrix into a final matching score through a convolutional neural network.

Table 1 summarizes the interaction-based matching model for both single-turn and multi-turn dialogues. Among these models, sequential matching network (SMN) [Wu *et al.*, 2017] is the most representative model. The model first encodes the utterance in the context and response candidate with RNNs and then lets each utterance in a context interacts with a response candidate based on a similarity-based approach. After that, the interaction matrix for each utterance-response pair is transformed into a matching vector with CNNs. The matching vectors are finally aggregated with an RNN as a matching score of the context and the response candidate. Following SMN, deep attention matching network (DAM) [Zhou *et al.*, 2018b] let each utterance interacts with a response candidate at different levels of representations obtained by a stacked self-attention and cross-attention module. Following the framework, Tao *et al.* (2019a) explores multiple granularities of

[4] This function means that each word in the sequence is independently encoded into word vectors.

representations for context-response matching, and Tao *et al.* (2019b) present a interaction-over-interaction network (IoI) that lets the context-utterance matching process go deep via iterative interactions. Besides, to alleviate the impact of noisy utterance on response selection, Yuan *et al.* (2019) propose a multi-hop selector network (MSN) to identify the relevant utterances which are further used for response matching.

2.4 Framework III: PLM-based Models

Recently, pre-trained language models (PLMs) [Devlin *et al.*, 2019] have shown impressive benefits for various downstream NLP tasks due to their strong capability of language representation and understanding, and some researchers try to apply them on response selection. By feeding the concatenation of all utterances in the context and the response candidates into a pre-trained multi-layer self-attention network (such as BERT), the representation, interaction, and aggregation operations can be performed through the attention mechanism in a unified model. The overall architecture of PLM-based matching framework is shown in Figure 1 (c). Particularly, Henderson *et al.* (2019) utilizes BERT to represent each utterance-response pair and aggregate these representations to calculate the matching score. Whang *et al.* (2020) treat the context as a long sequence and perform context-response matching with the BERT. Besides, the model also introduces the next utterance prediction and mask language modeling tasks borrowed from BERT during the post-training on dialogue corpus to incorporate in-domain knowledge for the matching model. Gu *et al.* (2020a) further consider incorporating speaker embeddings into BERT to promote the capability of context understanding in multi-turn dialogues. To further improve the training of PLM-based matching models, some researchers investigate various self-supervised training approaches along with the response selection task and lots of impressive results have been obtained [Xu *et al.*, 2021]. Wu *et al.* (2020) propose task-oriented dialogue BERT (TOD-BERT) trained on multiple human-human and multi-turn task-oriented datasets across over 60 domains and achieve impressive results on response selection tasks.

2.5 Comparison of Three Frameworks

In terms of efficacy, interaction-based models are generally better than representation-based models. This is because interaction-based models let a context message and a response interact at the very beginning, and thus matching information could be sufficiently preserved. Besides, PLM-based models conduct full interaction over the context and the response candidates and are pre-trained on large-scale corpora through many self-supervised tasks, therefore they usually significantly outperform the conventional interaction-based models and representation-based models on various benchmarks [Xu *et al.*, 2021; Whang *et al.*, 2020].

However, there is no free lunch. In terms of efficiency, PLM-based models and interaction-based models are more costly than models in representation-based models, due to the heavy computation of interaction. In particular, PLM-based models perform multi-layers of word-by-word interaction among the concatenated sequence of the whole context and the response, resulting in a large number of parameters. Moreover, for models in representation-based models, one can pre-compute the embedding of response candidates and store them in the index along with the text, which can further reduce the cost of representation-based models in online systems. To trade off the efficacy and efficacy, Humeau *et al.* (2020) present the Poly-encoder, a architecture with an additional learnt attention mechanism that represents more global features from which to perform self-attention, resulting in performance gains over Bi-encoders and large speed gains over PLM-based models.

3 Response Selection with Extra Knowledge

Knowledge is an awareness and understanding of the input message and its surrounding context. It can be obtained from various information sources, including but not limited to keywords, topics, linguistic features, knowledge bases, grounded texts and visual information. These sources provide information that can be used as extra information and then applied to enhance the process of response selection.

Document. A key step in response selection is measuring the matching degree between a context and a response candidate. Existing methods have achieved impressive performance on benchmarks, but responses are selected solely based on conversation history [Zhou *et al.*, 2018b; Yuan *et al.*, 2019]. Actually, different from a human being, who is capable of associating the dialogue with the background knowledge in her/his mind, a machine can merely capture limited information from the surface text of the query message. Consequently, it is difficult for a machine to understand the query fully, and to predict a proper response to make it more engaging. To bridge the gap of the knowledge between the human and the machine, researchers have begun to simulating this motivation by grounding dialogue agents with unstructured background documents [Ghazvininejad *et al.*, 2018; Zhang *et al.*, 2018a; Dinan *et al.*, 2019]. For example, Zhang *et al.* (2018a) build a persona-based conversation data set that employs the interlocutor's profile as the background knowledge; Zhou *et al.* (2018a) publish a data where conversations are grounded in articles about popular movies; Dinan *et al.* (2019) release another document-grounded data with Wiki articles covering a wide range of topics. Meanwhile, several retrieval-based knowledge-grounded dialogue models are proposed, such as document-grounded matching network (DGMN) [Zhao *et al.*, 2019] which lets the dialogue context and all knowledge entries interact with the response candidate respectively via the cross-attention mechanism. Gu *et al.* (2020b) further propose to pre-filter the context and the knowledge and then use the filtered context and knowledge to perform the matching with the response. Besides, with the help of gold knowledge index annotated by human wizards, Dinan *et al.* (2019) consider joint learning the knowledge selection and response matching in a multi-task manner or training a two-stage model.

Visual Information. Human conversation, on the other hand, is often grounded by more than one kind of perception. In addition to what they read (e.g., text), people also respond according to what they see (e.g., images or videos). Thus, many works consider multi-modal response selection for retrieval-based chatbots, such as image-grounded conversation [Huber *et al.*, 2018; Shuster *et al.*, 2020] or video-grounded conversation [Pasunuru and Bansal, 2018] where there exists a visual context besides a textual context, and a proper response is selected with both the textual conversation history and the visual content in the video taking into consideration. In the past few years, a lot of effort has been paid to bringing vision and dialogue together. Das *et al.* (2017) extend the scenario of VQA to a visual dialog task which requires a machine to answer a series of questions according to an image, and Mostafazadeh *et al.* (2017) propose a new task where a question is first asked according to an image and then a response is generated following the image and the question. Huber *et al.* (2018) propose an image-grounded conversational agent using visual sentiment, facial expression and scene features for emotional image-based dialogue. Pasunuru and Bansal (2018) assess the ability to execute dialogue given video of computer soccer games and present a new game-chat based video context, many-speaker dialogue task and dataset. More recently, Shuster *et al.* (2020) release the Image-Chat dataset which contains grounded dialogue involving open-ended discussion of a given image, and propose to handle multi-modal dialogue by fusing Transformer architectures for encoding dialogue history and responses and ResNet architectures for encoding images.

Topic Information. In context-response matching, a good candidate response tends to have the same topic as the context. Therefore, the topic can be used as the prior knowledge in the matching process. Based on this idea, many researchers propose to incorporate topic information into context-response matching. Wu *et al.* (2018a) introduce the topic information to enrich the semantics of the context and the response candidate, and consider extra matching channels between the context and the topic of candidate response, the response and the topic of the context, and the topic of the context and the response candidate. Wu *et al.* (2018c) propose a general matching framework that can integrate external knowledge (such as topic words, grammatical patterns, relations) into context-response matching model. The model utilizes external knowledge to enhance the representation of context messages and response candidates, and then matches the context and replies based on enhanced representations. Wang *et al.* (2020) treat response selection as a dynamic topic tracking task to match the topic between the response and relevant conversation context in multi-party conversation.

Dialogue Act. Dialogue acts (DA) are higher-level semantic abstraction associated to utterances in a conversation [Kumar *et al.*, 2018]. To build a dialogue system that can have coherent conversations with humans, several studies explore using dialogue act to guide the response selection in retrieval-based dialogue model [Kumar *et al.*, 2018; Kumar *et al.*, 2019; Yang *et al.*, 2020]. For examples, Kumar *et al.* (2018) propose a dialogue-act-driven hierarchical siamese model that uses the sequential dialogue act information for response matching; Yang *et al.* (2020) define and characterize different user intent types, and then propose an intent-aware neural ranking model for response retrieval which incorporates intent-aware utterance attention to derive the importance weighting scheme of different utterances to improve conversation history understanding. Kumar *et al.* (2019) combines the predicted dialogue acts of the context and the response with the context, and use the combined representation to select an appropriate response.

Emotion. Both the semantic meaning and the implicit feeling of the input message can help predict a more empathetic response candidate. To construct an empathetic retrieval-based conversation system, many studies consider incorporating emotional factors into context-response matching [Rashkin *et al.*, 2019; Qiu *et al.*, 2020; Zandie and Mahoor, 2020; Zhong *et al.*, 2020]. Rashkin *et al.* (2019) try a naive matching model on the proposed dataset by prepending the emotion labels predicted by a pre-trained classifier. Qiu *et al.* (2020) propose an emotion-aware transition network to model the emotion flow in a conversation for the context-response matching task and further design a unified model for emotion-controllable response selection. Zandie and Mahoor (2020) present a multi-head Transformer architecture that can use explicit contextual information on emotions, topic, and DA to respond to users' utterances with proper emotions. Zhong *et al.* (2020) release a multi-turn persona-based empathetic conversation dataset in two domains and meanwhile design an efficient BERT-based response selection model using multi-hop co-attention.

4 Learning of Response Selection Models

Existing research has made great efforts to build matching models with various neural architectures or use additional knowledge besides dialogue utterances. At the same time, how to better learn the matching model has also received more and more attention from researchers. Some of these works are devoted to learning a robust dialogue model from existing imperfect dialogue data sets (such as noisy dialogue utterances or inaccurate response labels) and others use cutting-edge machine learning technology to seek better models.

Learning From Noisy Data. Most existing models are simply learned by distinguishing human responses from some automatically constructed negative response candidates (e.g., by random sampling). Although this heuristic approach can avoid expensive and exhausting human labeling, it suffers from noise in training data, as many negative examples are actually false negatives[5]. To alleviate the problem, Wu *et al.* (2018b) leverage a Seq2Seq model as a weak annotator to assign a score for each response candidate of the dialogue and learn matching models through the scores. Feng *et al.* (2019) introduce the co-teaching framework for eliminating the effect of training noises. The learning approach maintains two matching models and lets them teach each other. More recently, Lin *et al.* (2020) attempt to diversify the training negative examples with an offline retrieval system and a pre-trained Seq2seq model. Zhang *et al.* (2019) propose an adversarial learning framework to enhance a retrieval-generation ensemble model which consists of a language-model-like generator, a ranker generator, and a ranker discriminator. This framework encourages two generators to generate responses that are scored higher by the discriminative ranker, while the discriminator downweighs adversarial samples and selects those responses that

[5] Responses sampled from other contexts may also be proper candidates for a given context.

are favored by the two generators. Su *et al.* (2020) propose a hierarchical curriculum learning framework that progressively strengthens the model's ability in identifying the mismatched information between the dialogue context and response.

Other Advanced Learning Methods. At the same time, many researchers have applied numerous machine learning methods to retrieval-based dialogue tasks such as transfer learning, incremental learning, multi-task learning, self-supervised learning, and so on. To solve the issue that current models may not be efficient enough for industrial applications, Qiu *et al.* (2018) employ transfer learning for context-aware question matching in information-seeking conversations in e-commerce. To address the problem that existing dialogue systems may break down when encountering unconsidered user needs, Wang *et al.* (2019) propose a novel incremental learning framework namely Incremental Dialogue System (IDS) to design task-oriented dialogue systems. Considering the low-data regime of most task-oriented dialogue tasks, Henderson *et al.* (2019) first pretrain the response selection model on large general-domain conversational corpora and then fine-tune the pretrained model for the target dialogue domain. In order to model the emotion in the ongoing dialogue, Qiu *et al.* (2020) propose an emotion-aware transition network to enhance dialogue response selection with emotional information in a multi-task way. Mesgar *et al.* (2020) also propose a multi-task learning approach for dialogue coherence assessment using dialogue act prediction as an auxiliary task, yielding more informative utterance representations for coherence assessment. Recently, to make better use of the potential training signals contained in the dialogue, and at the same time solve the incoherent and inconsistent problems between context and response, Xu *et al.* (2021) propose learning a context-response matching model with auxiliary self-supervised tasks designed for the dialogue data based on pre-trained language models.

5 Conclusion and Open Challenges

This paper reviews recent context-response matching models for retrieval-based dialogue systems categorized by their architecture, and summarize recent advances on the research of response selection models, including the application of pre-training techniques, incorporation with extra knowledge, and exploration on more effective model learning method. Through extensive efforts have been made and impressive results have been obtained on many benchmarks on response selection for retrieval-based dialogue systems, there are still several open challenges.

- **Do Dialogue Models Understand the Conversation History Effectively?** Multi-turn conversation modeling and understanding plays an important role in dialogue systems, either for retrieval-based methods or generation-based methods. While existing retrieval models have demonstrated the fair ability to selecting relevant responses, they still lack the ability to "understand" and process the dialog history to match coherent and appropriate responses. In particular, Sankar *et al.* (2019) find that both recurrent and transformer-based seq2seq models are not significantly affected even by drastic and unnatural modifications to the dialog history in multi-turn response generation. Besides, Li *et al.* (2019) demonstrate that the performance of context-response matching models is closely related to the degree of word overlap between context and the response candidates and the disturbance of local matching signals between context and response has a significant influence on the performance of a matching model. Moreover, in MuTual [Cui *et al.*, 2020], a recently released multi-turn dialogue reasoning benchmark, some representative multi-turn response selection model (such as SMN and DAM) achieve poor performance and they drop by more than 50 absolute R@1 points compared to their performance on the Ubuntu Corpus, indicating that existing response matching models cannot handle multi-turn dialogue reasoning problem well. All these studies indicate that current response selection models can not fully understand the dialogue but select the response based on surface lexical features. Therefore, more studies should be conducted to explore more effectively conversation understating and reasoning models.

- **Logical Consistency of Multi-turn Response Selection.** Existing context-response matching models pay much attention to capturing the semantic relevance between the context message and the response candidate, but usually neglect the logical consistency of a response candidate that is a long-standing issue faced by dialogue models [Wu *et al.*, 2017; Zhou *et al.*, 2016]. To reduce the problem of consistency in dialogue, Welleck *et al.* (2019) first construct a dataset, Dialogue NLI, which contains sentence pairs labeled as entailment, neutral, or contradiction, and then consider training a natural language inference model that measures the logical consistency to re-rank the response candidates. Although dialogue NLI models provide a solution to identify some inconsistencies in the selected responses, the relatively small training size makes it hard for the dialogue NLI models to generalize on other domains or topics. In the future, more effort should be paid to model the logical consistency in the context-response matching model.

- **Domain Shift in Multi-turn Response Selection.** For better model reproduction and comparison, existing dialogue models usually focus on a single domain or human conversation data from a fixed source (e.g. Douban, Twitter) for model learning and testing. However, the content of human dialogue will change with social development and language evolution. This kind of change causes a large deviation in the data distribution between online user input and the actual training data, resulting in the dialogue model can not accurately understand the user input and generate poor quality responses. Some researchers use the current large-scale pre-trained language model to fine-tune the static dialogue data to iteratively update the model, but the fine-tuning can not meet the conversations of emerging topics or audiences. More importantly, the currently available conversation data sets are far from covering all contents that can be involved in open-domain conversations. Therefore, future work may consider building a sustainable "evolution" dialogue system models that can learn and evolve themselves according to the continuously updated open-source dialogue data in various social platforms.

References

[Chen and Wang, 2019] Qian Chen and Wen Wang. Sequential matching model for end-to-end multi-turn response selection. In *ICASSP*, pages 7350–7354, 2019.

[Cui *et al.*, 2020] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In *ACL*, pages 1406–1416, 2020.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[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 *NAACL*, pages 4171–4186, 2019.

[Dinan *et al.*, 2019] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*, 2019.

[Feng *et al.*, 2019] Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems. In *ACL*, 2019.

[Ghazvininejad *et al.*, 2018] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wentau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *AAAI*, 2018.

[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 *CIKM*, 2020.

[Gu *et al.*, 2020b] Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Si Wei, and Xiaodan Zhu. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *EMNLP*, 2020.

[Henderson *et al.*, 2019] Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. Training neural response selection for task-oriented dialogue systems. In *ACL*, 2019.

[Hu *et al.*, 2014] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *NIPS*, 2014.

[Huber *et al.*, 2018] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. Emotional dialogue generation using image-grounded language models. In *CHI*, pages 1–12, 2018.

[Humeau *et al.*, 2020] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*, 2020.

[Kang *et al.*, 2014] Longbiao Kang, Baotian Hu, Xiangping Wu, Qingcai Chen, and Yan He. A short texts matching method using shallow features and deep features. In *NLPCC*, pages 150–159. 2014.

[Kumar *et al.*, 2018] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. Dialogue-act-driven conversation model : An experimental study. In *COLING*, 2018.

[Kumar *et al.*, 2019] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. A practical dialogue-act-driven conversation model for multi-turn response selection. In *EMNLP*, pages 1980–1989, 2019.

[Li *et al.*, 2019] Jia Li, Chongyang Tao, Nanyun Peng, Wei Wu, Dongyan Zhao, and Rui Yan. Evaluating and enhancing the robustness of retrieval-based dialogue systems with adversarial examples. In *NLPCC*, pages 142–154, 2019.

[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 *EMNLP*, pages 9220–9229, 2020.

[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 *SIGDIAL*, pages 285–294, 2015.

[Lu and Li, 2013] Zhengdong Lu and Hang Li. A deep architecture for matching short texts. In *NIPS*, 2013.

[Mesgar *et al.*, 2020] Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. Dialogue coherence assessment without explicit dialogue act labels. In *ACL*, July 2020.

[Mostafazadeh *et al.*, 2017] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP*, 2017.

[Pasunuru and Bansal, 2018] Ramakanth Pasunuru and Mohit Bansal. Game-based video-context dialogue. In *EMNLP*, pages 125–136, 2018.

[Qiu *et al.*, 2018] Minghui Qiu, Liu Yang, Feng Ji, Wei Zhou, Jun Huang, Haiqing Chen, Bruce Croft, and Wei Lin. Transfer learning for context-aware question matching in information-seeking conversations in E-commerce. In *ACL*, pages 208–213, 2018.

[Qiu *et al.*, 2020] Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. What if bots feel moods? In *SIGIR*, pages 1161–1170, 2020.

[Rashkin *et al.*, 2019] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381, 2019.

[Sankar *et al.*, 2019] Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *ACL*, pages 32–37, 2019.

[Shuster *et al.*, 2020] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image-chat: Engaging grounded conversations. In *ACL*, pages 2414–2429, 2020.

[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:2012.14756*, 2020.

[Tao *et al.*, 2019a] 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 *WSDM*, 2019.

[Tao *et al.*, 2019b] 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.

[Tay *et al.*, 2018] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. In *EMNLP*, pages 4492–4502, 2018.

[Wang *et al.*, 2015] Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. Syntax-based deep matching of short texts. In *AAAI*, 2015.

[Wang *et al.*, 2017] Ziliang Wang, Si Li, Guang Chen, and Zhiqing Lin. Deep and shallow features learning for short texts matching. In *PIC*, pages 51–55, 2017.

[Wang *et al.*, 2019] Weikang Wang, Jiajun Zhang, Qian Li, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. Incremental learning from scratch for task-oriented dialogue systems. In *ACL*, pages 3710–3720, 2019.

[Wang *et al.*, 2020] Weishi Wang, Shafiq Joty, and Steven CH Hoi. Response selection for multi-party conversations with dynamic topic tracking. In *EMNLP*, 2020.

[Welleck *et al.*, 2019] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *ACL*, pages 3731–3741, 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. In *Interspeech*, 2020.

[Wu *et al.*, 2017] 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 *ACL*, pages 496–505, 2017.

[Wu *et al.*, 2018a] Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing*, 316:251–261, 2018.

[Wu *et al.*, 2018b] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *ACL*, pages 420–425, 2018.

[Wu *et al.*, 2018c] Yu Wu, Wei Wu, Can Xu, and Zhoujun Li. Knowledge enhanced hybrid neural network for text matching. In *AAAI*, 2018.

[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 *EMNLP*, pages 917–929, 2020.

[Xu *et al.*, 2020] Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue modeling. *arXiv preprint arXiv:2009.12539*, 2020.

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

[Yan *et al.*, 2016] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, 2016.

[Yan *et al.*, 2018] Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, and Zhoujun Li. Response selection from unstructured documents for human-computer conversation systems. *Knowledge-Based Systems*, 142, 2018.

[Yang *et al.*, 2020] Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, and Haiqing Chen. IART: Intent-aware response ranking with transformers in information-seeking conversation systems. In *WWW*, pages 2592–2598, 2020.

[Yuan *et al.*, 2019] 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 *EMNLP*, 2019.

[Zandie and Mahoor, 2020] Rohola Zandie and Mohammad H Mahoor. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*, 2020.

[Zhang *et al.*, 2018a] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213, 2018.

[Zhang *et al.*, 2018b] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. Modeling multi-turn conversation with deep utterance aggregation. In *COLING*, pages 3740–3752, 2018.

[Zhang *et al.*, 2019] Jiayi Zhang, Chongyang Tao, Zhenjing Xu, Qiaojing Xie, Wei Chen, and Rui Yan. Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation. In *SIGIR*, 2019.

[Zhao *et al.*, 2019] Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. A document-grounded matching network for response selection in retrieval-based chatbots. In *IJCAI*, pages 5443–5449, 2019.

[Zhong *et al.*, 2020] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In *EMNLP*, 2020.

[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 *EMNLP*, pages 372–381, 2016.

[Zhou *et al.*, 2018a] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In *EMNLP*, pages 708–713, 2018.

[Zhou *et al.*, 2018b] 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 *ACL*, 2018.