ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746656

A MULTI DOMAIN KNOWLEDGE ENHANCED MATCHING NETWORK FOR RESPONSE SELECTION IN RETRIEVAL-BASED DIALOGUE SYSTEMS

Xiuyi Chen, Feilong Chen, Shuang Xu, Bo Xu

School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Science, Beijing, P.R.China
{chenxiuyi2017, chenfeilong2018, shuang.xu, xubo}@ia.ac.cn

ABSTRACT

Building a human-machine conversational agent is a core problem in Artificial Intelligence, where knowledge has to be integrated into the model effectively. In this paper, we propose a Multi Domain Knowledge Enhanced Matching Network (MDKEMN) to build retrieval-based dialogue systems that could leverage both explicit knowledge graph and implicit domain knowledge for response selection. Specifically, our MDKEMN leverages the self-attention mechanism of a single-stream Transformer to make deep interactions among the dialogue context, response candidate and external knowledge graph, and finally returns the matching degree of each context-response pair under the external knowledge. Furthermore, to leverage the implicit domain knowledge from all domains to improve the performance of each domain, we combine the multi-domain datasets for training and then finetune the pretrained model on each domain. Experimental results show (1) the effectiveness of both explicit and implicit knowledge incorporating and (2) the superiority of our approach over previous baselines on a Chinese multi-domain knowledge-driven dialogue dataset.

Index Terms— External Knowledge, Deep Matching, Response Selection, Retrieval-based Dialogue Systems

1. INTRODUCTION

Building a human-machine conversational agent is one of the most important and challenging tasks in artificial intelligent(AI). Thanks to the availability of large amounts of human dialogue data and the recent progress on deep learning-based approaches, past few years have witnessed the rapid development of conversational systems [1, 2, 3, 4, 5]. Those methods can be roughly divided into two categories: the generative methods and retrieval-based methods [6]. In this paper, we are interested in the retrieval-based approaches, which reply to human input by selecting a matched response from the prebuilt response candidates [7, 8, 9].

It is crucial for retrieval-based systems to measure the matching degree between the dialogue context and response candidates. Attention is drawn from single-turn context-response matching [10, 11] to multi-turn response selection. And various works improve retrieval-based systems with multi-granularity matching [12, 13], deep contexts-response interaction [14, 15] and new training strategy [16]. However, those methods, which highly rely on the information stored in training corpora, still suffer from the semantic gaps between dialogue contexts and responses [17] to obtain a better matching function. An important reason is that these systems do not explicitly use external knowledge and have limited implicit knowledge learned from the specific domain training corpora while human with general implicit background knowledge often uses external explicit materials to conduct conversations.

Considering the gap of knowledge between human and machine, previous approaches have been developed to incorporate external knowledge into matching networks for response selection in retrieval-based dialogue systems. Early works attempt to leverage the topic clues [18, 19], text-related entities [20], domain keyword descriptions [21] or relevant question-answer pairs [22] as the external knowledge to strengthen the text representation for response selection. However, those works tend to incorporate the external knowledge in a shallow fusion manner. Recent works [23, 22, 9] tend to incorporate the external knowledge in an early fusion way via the attention mechanism. However, most of them still first encode the text and knowledge information separately, and then blend these information via the attention mechanism. Currently, pretrained language models with its implicit parameter knowledge learned on large corpus have shown significant benefits for various downstream NLP tasks [24, 25], and some researchers have tried to apply them on the vanilla response selection task without external knowledge [26, 27, 28].

Different from them, we propose a Multi Domain Knowledge Enhanced Matching Network (MDKEMN) to build retrieval-based systems that could leverage both explicit knowledge graph and implicit domain knowledge for response selection. Specifically, for explicit knowledge, we leverage a single-stream Transformer [29] to make deep interactions among the dialogue context, response candidate and external knowledge graph after embedding; for implicit knowledge, we leverage the implicit domain knowledge from all domains to improve the performance of each domain, while most of previous works train their models on each domain dataset separately. Finally, we evaluate our approach on a Chinese multi-domain knowledge-driven dialogue dataset, i.e., KdConv [30], and experimental results show (1) the effectiveness of both explicit and implicit knowledge incorporating and (2) the superiority of our approach over previous baselines.

2. METHOD

2.1. Task Definition

Suppose that we have a dataset $\mathcal{D} = \left\{\left(K^i, c^i, r^i, y^i\right)\right\}_{i=1}^{N}$ with N samples, where the knowledge graph $K^i = \{K_1^i, \cdots, K_{l_K}^i\}$ consists of l_K triples[1], the dialogue context $c^i = \{c_1^i, \cdots, c_{l_c}^i\}$ and the

[1]Each triple $K_j^i = (h_j^i, p_j^i, t_j^i)$ consists of the head entity h_j^i, the relation or predicate p_j^i and the tail entity t_j^i. (Flying Higher, Release date, March 19, 2005) is an example.

Authorized licensed use limited to: IEEE Xplore. Downloaded on April 30,2022 at 13:15:24 UTC from IEEE Xplore. Restrictions apply.

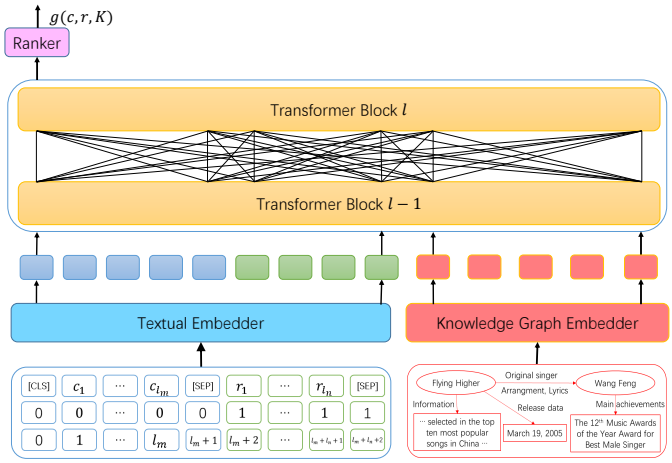

Fig. 1. The proposed MDKEMN architecture for retrieval-based dialogue systems. Our MDKEMN leverages a single-stream Transformer encoder to encode the dialogue context, response candidate and knowledge graph information together and allows deep interactions among those information via the multi-head self-attention, and finally returns the matching degree between context and response under the external knowledge.

response candidate $r^i = \{r_1^i, \cdots, r_{l_r}^i\}$ contains l_c and l_r tokens, respectively, and the label $y^i \in \{0, 1\}$ indicates whether this is a positive example. $y^i = 1$ means r^i is an appropriate response given c^i and K^i while $y^i = 0$ indicates the negative example. The task is to learn a matching model $g(\cdot, \cdot, \cdot)$ from the dataset $\mathcal{D}$, and thus for any context-response pair (c, r), $g(c, r, K)$ returns the matching degree between dialogue context c and response candidate r under the external knowledge K.[2]

2.2. Model Overview

In this paper, we propose our MDKEMN as the matching model $g(c, r, K)$ in retrieval-based dialogue systems and Fig. 1 shows the architecture of our model. In brief, MDKEMN is composed of four parts: the textual embedder, the knowledge graph embedder, the transformer-based interaction layer and the output layer. First, the two embedders are responsible for transforming the symbolic data, i.e., the dialogue context, the response candidates and the external knowledge graph, into the distributed representations, e.g., the word embedding [31]. And then, the transformer-based interaction layer leverages the multi-head self-attention mechanism to fully fuse those three distributed embedding representations and get the final vector representation. Finally, in the output layer, we compute the similarity between dialogue context and the response candidate under the external knowledge graph. The details of each components are described in the following sections.

2.3. Model Details

2.3.1. Textual Embedding

We first format the dialogue context c and response candidate r as a token sequence $x = \{[\text{CLS}], c_1, \ldots, c_{l_m}, [\text{SEP}], r_1, \ldots, r_{l_n}, [\text{SEP}]\}$, where [CLS] and [SEP] are special tokens to mark the beginning and end of a sentence. And then, the textual embedder converts each token x_i into a vector representation $\mathbf{x}_i$ as follows:

$$\mathbf{x}_i = \text{TE}(x_i) + \text{PE}(x_i) + \text{SE}(x_i) \in \mathbb{R}^d, \tag{1}$$

where d is the hidden size, $\text{TE}(\cdot)$, $\text{PE}(\cdot)$ and $\text{SE}(\cdot)$ are token, position and segment embeddings, and the input representation for each token is the sum of those embeddings.

2.3.2. Knowledge Graph Embedding

For each knowledge triple $K_j = (h_j, p_j, t_j)$, we use the knowledge graph embedder to obtain the representation $\mathbf{k}_j$ as follows:

$$\begin{aligned}\mathbf{k}_j &= \mathbf{W_k} \cdot \hat{\mathbf{k}}_j + \mathbf{b_k} \in \mathbb{R}^d \\ \hat{\mathbf{k}}_j &= \frac{1}{|h_j| + |p_j| + |t_j|} \sum_{K_{j,i} \in K_j} \sum_{e \in K_{j,i}} \text{KE}(e) \in \mathbb{R}^{\hat{d}},\end{aligned} \tag{2}$$

where $|\cdot|$ denotes the token number, $\text{KE}(\cdot)$ is the Knowledge Embedder to convert each knowledge triple into a vector $\hat{\mathbf{k}}_j$ in the $\hat{d}$-th dimension embedding space via averaging all embeddings of the tokens e in this triple, and trainable parameters $\mathbf{W_k} \in \mathbb{R}^{d \times \hat{d}}$ and $\mathbf{b_k} \in \mathbb{R}^d$ are used to project $\hat{\mathbf{k}}_j$ into $\mathbf{k}_j$ in the text embedding space.

2.3.3. Deep Interaction

In this paper, we leverage a single-stream Transformer encoder to encode dialogue context, response candidate and knowledge graph information together and thus allow deep interactions among those information in an early fusion manner via the multi-head self attention mechanism. Specifically, we first pack the textual embeddings and knowledge embeddings into an augmented embedding representation $\mathbf{H}^0 = [\mathbf{x}_1, \cdots, \mathbf{x}_{l_m+l_n+3}, \mathbf{k}_1, \cdots, \mathbf{k}_{l_K}]$ and then encode them into multiple levels of contextual representations $\mathbf{H}^l = [\mathbf{h}_1^l, \ldots, \mathbf{h}_{l_H}^l]$ using L-stacked Transformer blocks, where $l_H = l_m + l_n + 3 + l_K$ and the l-th Transformer block is denoted as $\mathbf{H}^l = \text{Transformer}(\mathbf{H}^{l-1}), l \in [1, L]$. Inside each Transformer block, the previous layer's output $\mathbf{H}^{l-1} \in \mathbb{R}^{l_H \times d}$ is aggregated using the multi-head self-attention [29]:

$$\begin{aligned}\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_\mathbf{Q}^l, \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_\mathbf{K}^l, \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_\mathbf{V}^l \\ \mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend,} \\ -\infty, & \text{prevent from attending ,} \end{cases} \\ \mathbf{A}^l = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}\end{aligned} \tag{3}$$

where $\mathbf{W}_\mathbf{Q}^l, \mathbf{W}_\mathbf{K}^l, \mathbf{W}_\mathbf{V}^l \in \mathbb{R}^{d \times d_k}$ are learnable weights for computing the queries, keys, and values respectively, and $\mathbf{M} \in \mathbb{R}^{l_H \times l_H}$ is the self-attention mask that determines whether tokens from two layers can attend each other. Then $\mathbf{A}^l$ is passed into a feedforward layer to compute $\mathbf{H}^l$ for the next layer.

2.3.4. Output Layer

The output layer calculates the similarity between dialogue context and the response candidate under the external knowledge graph. And we define the score function $g(c, r, K)$ as following:

$$g(c, r, K) = \text{sigmoid}\left(\mathbf{W}_o\mathbf{H}_{[\text{CLS}]}^L + \mathbf{b}_o\right) \in \mathbb{R}^1, \tag{4}$$

where $\mathbf{H}_{[\text{CLS}]}^L \in \mathbb{R}^d$ is the representation for [CLS] token from the Transformer encoder and $\mathbf{W}_o \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_o \in \mathbb{R}^1$ are trainable parameters.

[2] We omit the superscript of each sample (K^i, c^i, r^i, y^i) for brevity.

Authorized licensed use limited to: IEEE Xplore. Downloaded on April 30,2022 at 13:15:24 UTC from IEEE Xplore. Restrictions apply.

Table 1. Statistics of the KdConv dataset.

Domain	Film	Music	Travel	Total
Train Dialogues	1200	1200	1200	3600
Dev Dialogues	150	150	150	450
Test Dialogues	150	150	150	450
Total Utterances	36618	24885	24093	85596
Avg. # utters per dialogue	24.4	16.6	16.1	19.0
Avg. # tokens per utter	13.3	12.9	14.5	13.5
Total triples of KG	11875	5747	5287	22909
Avg. # triples per dialogue	16.8	10.4	10.0	10.1
Avg. # tokens per triple	25.8	29.7	31.0	28.3

2.4. Multi Domain Training

The success of pretrained language models suggests that the pretrained parameters could implicitly store knowledge from a large corpus [32, 25, 33, 34, 35]. Inspired by this discovery, we attempt to mine the implicit information from multi domain to improve the performance of each domain.

Specifically, we combine the multi-domain datasets for training and then finetune the pretrained model on each domain. And in both multi-domain training and single-domain finetuning stages, we optimize our models via the negative log likelihood loss function. Let Θ denote the parameters, the objective function can be formulated as:

$$\mathcal{L}(D, \Theta) = -\sum_{i=1}^{N} \left(y^i \log g^i + \left(1 - y^i\right) \log \left(1 - g^i\right) \right) \quad (5)$$

3. EXPERIMENTS

3.1. Dataset

We evaluate our approach on KdConv [30] and its statistics are shown in the Table 1. KdConv contains 4.5K dialogues from three domains (i.e., film, music and travel) and each dialogue is associated with a related explicit knowledge graph. In this paper, we use the retrieval setting where there are 10 response candidates, including the ground truth response.

3.2. Models for Comparison

DAM [14] is a transformer encoder-based model, which leverages self-attention and cross-attention to calculate the matching score.
IOI [15] performs deep matching between the utterances and responses through multiple interaction block chains.
IMN [36] is the interactive matching network to perform the global and bidirectional interactions between the context and response.
BERT [24] is a vanilla model finetuned to the response selection task without external explicit knowledge.
BERT+KVMN [30] enhances the finetuned BERT with a Key-Value Memory Network [37] to store and read out the external knowledge.
MDKEMN is our single-stream transformer-based model that leverages the Implicit Domain Knowledge (IDK) and Explicit Knowledge Graph (EKG). Removing the implicit and explicit knowledge, our model degenerates into the vanilla finetuned BERT.

3.3. Implementation Details

We implement our models with PyTorch[3] and use the parameters of BERT[4] to initialize our textual embedder and deep interaction layer. For knowledge graph embedder, we first use the Jieba Chinese word segmenter[5] for tokenization to obtain a knowledge token dictionary, and then initialize the $\mathrm{KE}(\cdot)$ operator using a pretrained 200-dimensional word embeddings[6]. For tokens not appearing in the pretrained embeddings, we assign them random embeddings sampled from a standard normal distribution $\mathcal{N}(0, 1)$.

We set the max length of the textual sequence x to 192 and the max length of each knowledge triple to 100. And the hidden size d is 768 and the knowledge embedding size $\hat{d}$ is 200. Then we use the Adam optimizer [38] with 0.1 warmup proportion to train all the models up to 8 epochs on two GPUs (TITAN Xp). The base learning rate is $2e$-5 for travel domain and $5e$-5 for other settings, and the batch size is set to 32.

Table 2. Automatic evaluation on the KdConv dataset. Note that we report models with "†" using results from the published paper [30].

Models	$\mathbf{R}_{10}$@1 ↑	$\mathbf{R}_{10}$@3 ↑	$\mathbf{R}_{10}$@5 ↑	MRR ↑	MAP ↓
Film					
DAM	38.68	72.88	88.91	0.587	2.710
IOI	29.82	61.04	80.03	0.502	3.385
IMN	51.54	82.77	93.57	0.688	2.163
BERT†	65.36	91.79	N/A	N/A	N/A
BERT+KVMN†	65.67	91.79	N/A	N/A	N/A
MDKEMN(Ours)	**71.32**	**93.14**	**98.39**	**0.827**	**1.545**
Music					
DAM	31.37	65.77	84.41	0.525	3.084
IOI	35.09	70.73	87.37	0.560	2.855
IMN	44.59	79.05	92.18	0.639	2.383
BERT†	55.64	86.90	N/A	N/A	N/A
BERT+KVMN†	56.08	86.87	N/A	N/A	N/A
MDKEMN(Ours)	**63.53**	**91.06**	**97.65**	**0.778**	**1.718**
Travel					
DAM	31.69	61.28	80.17	0.513	3.346
IOI	40.22	69.06	85.09	0.584	2.891
IMN	41.02	69.53	84.11	0.590	2.886
BERT†	45.25	71.87	N/A	N/A	N/A
BERT+KVMN†	46.64	73.98	N/A	N/A	N/A
MDKEMN(Ours)	**55.05**	**80.40**	**91.40**	**0.699**	**2.261**

3.4. Evaluation

We measure the performance of the matching models with the widely-used retrieval metrics [14, 36, 30]. We calculated the recall of the true positive replies among the n best-matched responses from 10 available candidates, denoted as $\mathbf{R}_{10}@n$. Here, we use $n \in \{1, 3, 5\}$. Moreover, we also adopt rank-aware evaluation metrics: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) of the human response.

3.5. Main Results

As Table 2 shows, our MDKEMN outperforms all baselines significantly on three domains of the KdConv dataset (t-test, p-value < 0.01) by achieving the highest scores in all automatic metrics. We see models without external knowledge perform poorly on this task, and BERT is a very strong baseline with general implicit knowledge learned from a large scale corpus. By leveraging both the explicit knowledge graph and implicit domain knowledge for response selection, our MDKEMN learns a much better matching function for retrieval-based dialogue systems.

[3]https://github.com/pytorch/pytorch
[4]https://github.com/ymcui/Chinese-BERT-wwm
[5]https://github.com/fxsjy/jieba
[6]https://ai.tencent.com/ailab/nlp/en/embedding.html

Authorized licensed use limited to: IEEE Xplore. Downloaded on April 30,2022 at 13:15:24 UTC from IEEE Xplore. Restrictions apply.

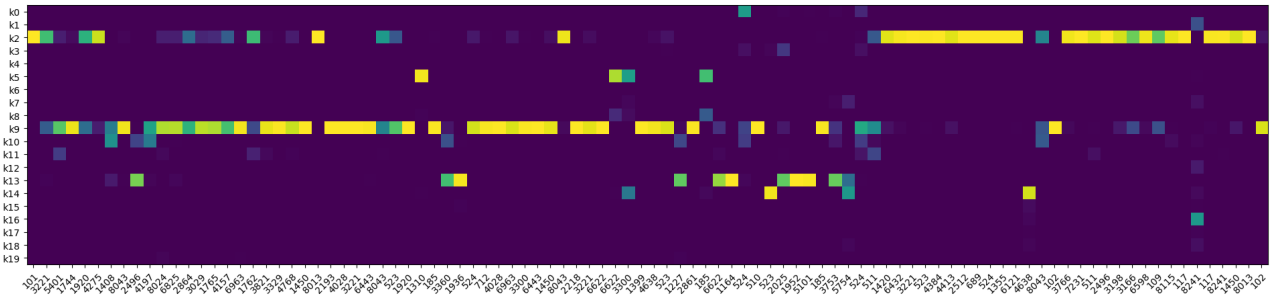

Fig. 2. Attention weight visualization in our MDKEMN. K0-K19 in the y axis indicates there are 20 triples in the external knowledge graph, and the numbers in the x axis belongs to the token sequence x, which can be translated into Chinese dialogue context and response according to the BERT-wwm-ext vocabulary[4]. In this example, the dialogue context focuses on the information of a film, while the response is mainly about the production cost. The K2 and K9 triples are related to the dialogue response and context, respectively.

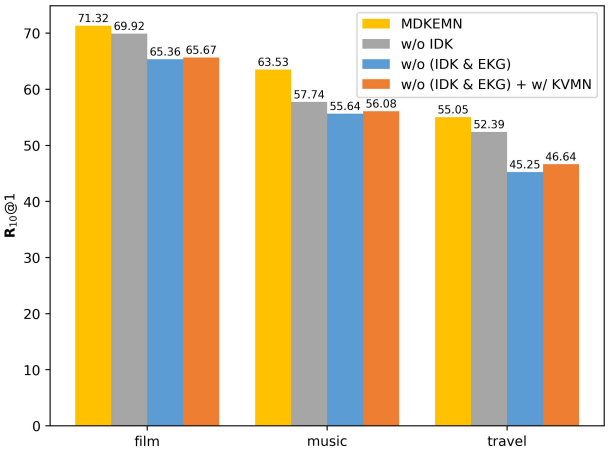

Fig. 3. Ablation Study on KdConv according to $\mathbf{R}_{10}@1$.

3.6. Ablation Study

We further conduct an ablation study to understand the effectiveness of both explicit and implicit knowledge incorporating in Figure 3. As our MDKEMN mainly contains two key features: the Implicit Domain Knowledge (IDK) and Explicit Knowledge Graph (EKG), and our model degenerates into the vanilla finetuned BERT after removing these two components. First, we see that there is shared knowledge among different domains and implicit domain knowledge mined from all domains could help to improve the performance of each domain. Moreover, the domain with insufficient training examples benefits from this kind of implicit knowledge more. For example, the $\mathbf{R}_{10}@1$ improvement on the music domain is more than 4 absolute points, whereas there is less than 2 point improvement on $\mathbf{R}_{10}@1$ on the film domain. Second, without the external knowledge graph, it is difficult to overcome the semantic gaps between dialogue contexts and response candidates and there is performance drop for response selection. Although adding KVMN helps, the improvement is not significant due to its relative shallow structure compared with BERT. On the other hand, our MDKEMN could leverage the external knowledge graph better and get the big improvement because our MDKEMN allows the deep interactions among the dialogue context, response candidate and external knowledge graph via the multi-head self-attention mechanism of the single-stream Transformer.

3.7. Attention Visualization

To interpret our MDKEMN, we visualize the attention weights between the text representation $\mathbf{H}^L_{[0:l_m+l_n+2]}$ and knowledge graph representation $\mathbf{H}^L_{[l_m+l_n+3:]}$ in Fig. 2. In this example, there is a topic change from the information of a film (e.g., the director and the actor) to the production cost, which causes the semantic gaps between dialogue context and response. In the knowledge graph, there exists a relation between K9 and K2, which are related to the dialogue context and response, respectively. We observe that our MDKEMN successfully changes its attention from the K9 information "...在美国上映。影片主要讲述..."(...released in the United States. The film is mainly about...) to the right part K2 "制片成本"(the product cost) in the first column, though the dialogue context is more related to the K9. Therefore, we believe that our MDKEMN leverages both explicit knowledge graph and implicit domain knowledge to overcome the semantic gaps and finally gets the right response "没错。当时斥资$15,000,000呢"(That's right. It costs $15 million).

4. CONCLUSION

In this paper, we propose a Multi Domain Knowledge Enhanced Matching Network (MDKEMN) to build retrieval-based dialogue systems that could leverage both explicit knowledge graph and implicit domain knowledge for response selection. Specifically, our MDKEMN leverages a single-stream Transformer to encode the dialogue context, response candidate and external knowledge graph together and blend these information via the multi-head self-attention mechanism and finally returns the matching degree of each context-response pair under the external knowledge. Furthermore, to leverage the implicit domain knowledge from all domains to improve the performance of each domain, we combine the multi-domain datasets for training and then finetune the pretrained model on each domain. We evaluate our approach on the KdConv dataset and show (1) the effectiveness of both explicit and implicit knowledge incorporating and (2) the superiority of our approach over previous baselines.

5. ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China under Grant No.2018YFB1005104, the Key Research Program of the Chinese Academy of Sciences under Grant No.ZDBS-SSW-JSC006 and Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDA27030300.

Authorized licensed use limited to: IEEE Xplore. Downloaded on April 30,2022 at 13:15:24 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang, "A survey on dialogue systems: Recent advances and new frontiers," *SIGKDD*, 2017.

[2] Jianfeng Gao, Michel Galley, and Lihong Li, "Neural approaches to conversational AI," in *ACL*, 2018.

[3] Xiuyi Chen, Jiaming Xu, and Bo Xu, "A working memory model for task-oriented dialog response generation," in *ACL*, 2019.

[4] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao, "Challenges in Building Intelligent Open-Domain Dialog Systems," *ACM Trans. Inf. Syst.*, 2020.

[5] Jinjie Ni, Tom Young, Vlad Pandelea, et al., "Recent advances in deep learning-based dialogue systems," *CoRR*, 2021.

[6] Yiping Song, Cheng-Te Li, Jian-Yun Nie, et al., "An ensemble of retrieval-based and generation-based human-computer conversation systems," in *IJCAI*, 2018.

[7] Zongcheng Ji, Zhengdong Lu, and Hang Li, "An information retrieval approach to short text conversation," *CoRR*, 2014.

[8] Rui Yan and Dongyan Zhao, "Coupled context modeling for deep chit-chat: Towards conversations between human and computer," in *SIGKDD*, 2018.

[9] Yajing Sun, Yue Hu, Luxi Xing, Jing Yu, and Yuqiang Xie, "History-adaption Knowledge Incorporation Mechanism for Multi-turn Dialogue System," in *AAAI*, 2020.

[10] Zongcheng Ji, Zhengdong Lu, and Hang Li, "An information retrieval approach to short text conversation," *arXiv preprint arXiv:1408.6988*, 2014.

[11] Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu, "Syntax-based deep matching of short texts," in *IJCAI*, 2015.

[12] Xiangyang Zhou, Daxiang Dong, et al., "Multi-view response selection for human-computer conversation," in *EMNLP*, 2016.

[13] Yu Wu, Wei Wu, et al., "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots," in *ACL*, 2017.

[14] Xiangyang Zhou, Lu Li, Daxiang Dong, et al., "Multi-turn response selection for chatbots with deep attention matching network," in *ACL*, 2018.

[15] Chongyang Tao, Wei Wu, Can Xu, et al., "One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues," in *ACL*, 2019.

[16] Chunyuan Yuan, Wei Zhou, Mingming Li, et al., "Multi-hop selector network for multi-turn response selection in retrieval-based chatbots," in *EMNLP-IJCNLP*, 2019.

[17] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen, "Convolutional neural network architectures for matching natural language sentences," in *NeurIPS*, 2014.

[18] Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou, "Response selection with topic clues for retrieval-based chatbots," *Neurocomputing*, 2018.

[19] Yu Wu, Wei Wu, Can Xu, and Zhoujun Li, "Knowledge enhanced hybrid neural network for text matching," *AAAI*, 2018.

[20] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang, "Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm," in *IJCNN*, 2017.

[21] Debanjan Chaudhuri, Agustinus Kristiadi, Jens Lehmann, and Asja Fischer, "Improving response selection in multi-turn dialogue systems by incorporating domain knowledge," in *CoNLL*, 2018.

[22] L. Yang, M. Qiu, C. Qu, et al., "Response ranking with deep matching networks and external knowledge in information-seeking conversation systems," in *SIGIR*, 2018.

[23] Jatin Ganhotra, Siva Sankalp Patel, and Kshitij Fadnis, "Knowledge-incorporating esim models for response selection in retrieval-based dialog systems," *DSTC-7 workshop at AAAI*, 2019.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[25] Rogers Anna, Kovaleva Olga, and Rumshisky Anna, "A primer in bertology: What we know about how bert works," *TACL*, 2020.

[26] Jesse Vig and Kalai Ramea, "Comparison of transfer-learning approaches for response selection in multi-turn conversations," in *Workshop on DSTC7*, 2019.

[27] Taesun Whang, Dongyub Lee, Chanhee Lee, et al., "An effective domain adaptive post-training method for bert in response selection.," in *INTERSPEECH*, 2020.

[28] Jia-Chen Gu, Tianda Li, Quan Liu, et al., "Speaker-aware bert for multi-turn response selection in retrieval-based chatbots," in *CIKM*, 2020.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," in *NeurIPS*, 2017.

[30] H. Zhou, Chujie Zheng, Kaili Huang, et al., "Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation," in *ACL*, 2020.

[31] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, 2003.

[32] Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov, "Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model," in *ICLR*, 2020.

[33] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, et al., "Probing pretrained language models for lexical semantics," in *EMNLP*, 2020.

[34] Jiahua Dong, Yang Cong, et al., "What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation," in *CVPR*, 2020.

[35] Alon Talmor, Oyvind Tafjord, Peter Clark, et al., "Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge," in *NeurIPS*, 2020.

[36] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu, "Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots," in *CIKM*, 2019.

[37] Alexander Miller, Adam Fisch, et al., "Key-value memory networks for directly reading documents," in *EMNLP*, 2016.

[38] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

Authorized licensed use limited to: IEEE Xplore. Downloaded on April 30,2022 at 13:15:24 UTC from IEEE Xplore. Restrictions apply.