UniRetriever: Multi-task Candidates Selection for Various Context-Adaptive Conversational Retrieval

Hongru Wang[♡▽], **Boyang Xue**[♡▽], **Baohang Zhou**[♠], **Rui Wang**[♣],
Fei Mi[◇], **Weichao Wang**[◇], **Yasheng Wang**[◇], **Kam-fai Wong**[♡▽]
[♡] MoE Key Laboratory of High Confidence Software Technologies, China
[▽] The Chinese University of Hong Kong [♠] Nankai University
[♣] Harbin Institute of Technology, Shenzhen [◇] Huawei Noah's Ark Lab
{hrwang, byxue}@se.cuhk.edu.hk, zhoubaohang@dbis.nankai.edu.cn, ruiwangnlp@outlook.com

Abstract

Conversational retrieval refers to an information retrieval system that operates in an iterative and interactive manner, requiring the retrieval of various external resources, such as persona, knowledge, and even response, to effectively engage with the user and successfully complete the dialogue. However, most previous work trained independent retrievers for each specific resource, resulting in sub-optimal performance and low efficiency. Thus, we propose a multi-task framework function as a universal retriever for three dominant retrieval tasks during the conversation: persona selection, knowledge selection, and response selection. To this end, we design a dual-encoder architecture consisting of a context-adaptive dialogue encoder and a candidate encoder, aiming to attention to the relevant context from the long dialogue and retrieve suitable candidates by simply a dot product. Furthermore, we introduce two loss constraints to capture the subtle relationship between dialogue context and different candidates by regarding historically selected candidates as hard negatives. Extensive experiments and analysis establish state-of-the-art retrieval quality both within and outside its training domain, revealing the promising potential and generalization capability of our model to serve as a universal retriever for different candidate selection tasks simultaneously.

Keywords: Conversational Retrieval, Knowledge Selection, Response Selection, Multi-task Framework

1. Introduction

Information Retrieval (IR) refers to the task of retrieving the most relevant candidates (e.g. top *n*) from a large corpus (a.k.a, candidates pool) for a given query, which receives a rapid proliferation of interest and attention in both academia and industry (Izacard and Grave, 2021a; Lu et al., 2022b; Chen et al., 2022). These retrieved evidence serve as additional semantic signals to provide important information, guiding the generation of the final answer in many downstream natural language tasks such as question answering (Karpukhin et al., 2020; Izacard and Grave, 2021b; Zhuang et al., 2022), machine reading comprehension (Khattab and Zaharia, 2020; Ren et al., 2021; Qu et al., 2021) and also dialogue systems (Dinan et al., 2018; Zhou et al., 2020; Xu et al., 2022; Wang et al., 2021, 2024). Benefiting from more accurate retrieved results, it is observed and well-acknowledged that the performance of these downstream tasks can be further improved (Dinan et al., 2018; Shuster et al., 2022; Xu et al., 2022) with the more powerful retriever such as DPR (Karpukhin et al., 2020), and RocketQA (Qu et al., 2021; Ren et al., 2021).

Unlike traditional Information Retrieval (IR) systems, Conversational Retrieval (CR) is an embodiment of an iterative and interactive IR system that has two distinct characteristics. On the one hand, the dialogue context is much longer than the question (a.k.a, the query) in the question-answering field (Karpukhin et al., 2020; Izacard and Grave, 2021b) because dialogue can last for many sessions (Qu et al., 2020), as seen in the Multi-Session Chat dataset (Xu et al., 2022), proposing additional challenges and difficulty in locating relevant contextual information and modeling the relationship between the query and the candidate. Despite the query rewriting (Wu et al., 2022) is a possible way to tackle the lengthy input, it necessitates large amounts of labeled data and still requires locating the relevant contextual turns once the length of multi-session dialogue exceeds the maximum input limit of the language models (Devlin et al., 2018; Zhuang et al., 2022; Touvron et al., 2023). Furthermore, in the ongoing conversation, multiple turns may revolve around a common topic but draw upon various external candidates. The subtle differences between these candidates play a key role in determining the order relationship between them and the current context.

On the other hand, various external candidates[1], such as persona and knowledge sources, need to

[1] To avoid ambiguity, we use "candidates" to represent all required sources in dialogue such as "persona", "knowledge" and others.

LREC-COLING 2024, pages 17074–17086
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

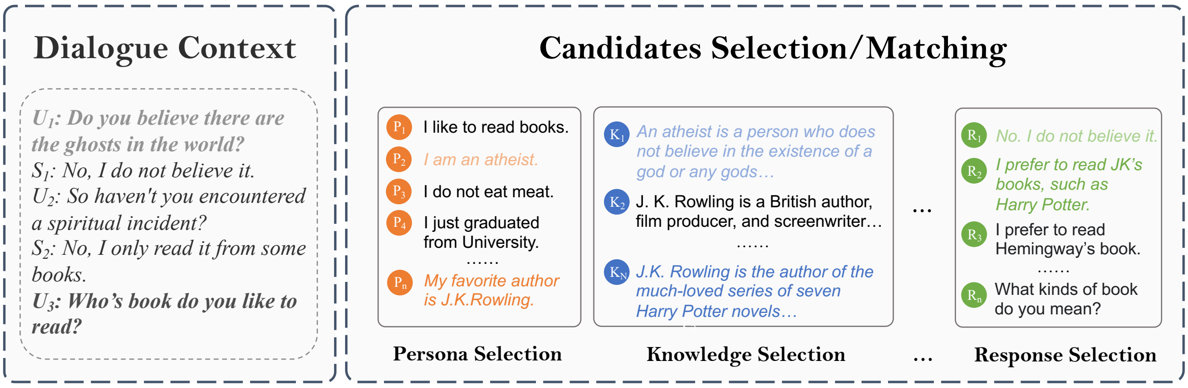
Dialogue Context
U1: Do you believe there are the ghosts in the world?
S1: No, I do not believe it.
U2: So haven't you encountered a spiritual incident?
S2: No, I only read it from some books.
U3: Who's book do you like to read?
Candidates Selection/Matching
P1 I like to read books.
P2 I am an atheist.
P3 I do not eat meat.
P4 I just graduated from University.
……
Pn My favorite author is J.K.Rowling.
K1 An atheist is a person who does not believe in the existence of a god or any gods…
K2 J. K. Rowling is a British author, film producer, and screenwriter…
……
KN J.K. Rowling is the author of the much-loved series of seven Harry Potter novels…
…
R1 No. I do not believe it.
R2 I prefer to read JK's books, such as Harry Potter.
R3 I prefer to read Hemingway's book.
……
Rn What kinds of book do you mean?
Persona Selection
Knowledge Selection
…
Response Selection

Figure 1: Different candidates selection tasks in a dialogue system: persona selection, knowledge selection, and response selection task. According to u_3 in the dialogue context, it is obvious to select p_n, k_n, and r_2 as target persona, knowledge, and response for the next turn respectively, while the p_2, k_1, and r_1 are historical selected persona, knowledge and response for historical turn u_1.

be retrieved to engage the users and complete the dialogue goal during the interactions (Wang et al., 2023c, 2024). For example, a human-like dialogue system not only needs to retrieve suitable persona descriptions (Xu et al., 2022; Liu et al., 2022) to maintain a consistent personality but also external knowledge such as Wikipedia[2] to answer the user's query (Dinan et al., 2018). These various candidates serve as crucial plugins to enhance the quality of responses, ensuring they are personalized, informative, coherent, and encompassing other vital features, depending on the particular source employed (Wang et al., 2023a,b). However, previous works usually train an independent retriever for each source, resulting in sub-optimal performance and inefficient computing (Dinan et al., 2018; Xu et al., 2022).

In response to these problems, we propose the **Universal C**onversational **R**etrieval (a.k.a. UniversalCR), a multi-task framework to advocate for a unifying view of three dominant candidates selection tasks in the conversation, including *persona selection*, *knowledge selection* and *response selection*. As shown in Figure 1, to respond to the final turn U_3, the dialogue system needs to select relevant persona P_n and knowledge K_n when generating the response similar to R_2. Besides that, the response selection task has always been a hot research topic in the retrieval-based dialogue systems (Hua et al., 2020; Gu et al., 2019, 2021a). It is observed that the persona P_n, knowledge K_n, and the ground-truth response R_2 share similar semantics, which motivates us to consider approaching these candidate selection tasks using a multi-task approach. Specifically, we first design a context-adaptive dialogue encoder to dynamically select related dialogue histories according to the query (i.e. the last utterance in the dialogue), while discarding noisy and unrelated utterances from lengthy dialogues. Then, we utilize historically selected candidates (P_2, K_1 and R_1 in the Figure 1 for current turn) during conversation to propose *historical contrastive loss*, while regarding historically selected candidates as **semi-hard negatives** and randomly sampled candidates as **easy negatives**. In addition, we utilize *pairwise similarity loss* to rank different pairs of candidates and dialogue context, inspired by previous negative sample mining (Xuan et al., 2020; Zhou et al., 2022b). To summarize, this work makes the following contributions:

- We propose a universal conversational retrieval framework, unifying three dominant candidate selection tasks: *persona selection*, *knowledge selection*, and *response selection*, in one framework while keeping the bottleneck layer as a single dot-product with a fixed size to achieve the balance of effectiveness and efficiency.

- We design one context-adaptive encoder and two carefully crafted loss constraints to address lengthy dialogue and capture subtle differences across various candidates respectively.

- We conduct extensive experiments to demonstrate the superiority of our proposed framework on six datasets in both supervised and unsupervised settings. Besides that, we offer an in-depth analysis of various candidate pool sizes and different context processing methods. These findings suggest a promising path toward building a robust and universal dialogue retrieval framework.

[2] https://www.wikipedia.org/

2. Related Work

Conversational Retrieval. Information Retrieval (IR) has been investigated and used in many applications such as web search and digital libraries, aiming to retrieve a ranked list of relevant documents in response to the query, while conversational retrieval is one embodied IR system. Most previous works conduct conversational retrieval in the context of conversational question answering, following the retrieval-then-rank framework in the traditional IR systems (Qu et al., 2020; Wu et al., 2021; Hu et al., 2022; Dai et al., 2022). However, they always are fine-tuned for a specific type of resource and thus limited in application and generalization (Yu et al., 2021; Kim and Kim, 2022). For example, Wu et al. (2021) introduces a knowledge identification model with an auxiliary task that predicts previously used knowledge to capture the history of dialogue-document connections. Instead, we directly regard previously used candidates[3] as the hard negative sample to model the relationships among different candidates. Additionally, many researchers have proposed different methods to fine-tune dual encoder retrievers (Guu et al., 2020; Karpukhin et al., 2020; Lin et al., 2021; Kim and Kim, 2022), such as coupling both coarse retrieval and fine reranking features to facilitate the final retrieval performance (Kumar and Callan, 2020; Zhang et al., 2022).

Recently, due to the exceptional performance of large language model (LLM) on various downstream tasks (Bang et al., 2023), there are some attempts which directly prompt LLMs to function as a retriever or re-ranker (Zhu et al., 2024; Wang et al., 2024). For example, Sun et al. (2023) evaluate the performance of LLMs as a re-ranker, and a very recent work formulate conversational retrieval as relevance score prediction task, which is optimized with knowledge source selection and response generation in a multi-task manner using a single LLM (Wang et al., 2024). It is worthy noting retrievers, rather than re-rankers, are typically applied to thousands of documents or queries, posing inefficiency and affordability challenges when using LLMs (no matter using fine-tuning or prompting). Additionally, LLMs are not optimal solutions in certain cases (Ma et al., 2023; Wang et al., 2024).

Pre-trained LMs for Dialogue. Large language models such as ToDBERT (Wu et al., 2020) and DialoGPT (Zhang et al., 2020) have shown impressive open-ended capabilities in both understanding and generation tasks after in-domain pertraining or fine-tuning. Zhang et al. (2021) propose a novel contextual dialogue encoder (i.e. DialogueBERT) with five well-designed pre-training tasks including response selection. Besides that, there are also some works that design specific architectures and embeddings to effectively exploit the semantic information in dialogues (Qu et al., 2019; Gu et al., 2021b). Our work also is in line with these previous works, by training a universal and special retrieval for dialogue, specifically long ones, to retrieve various external candidates needed to engage users and complete the dialogue goal.

3. Method

In this section, we will introduce the three modules of our proposed framework one by one: context-adaptive encoder, candidate encoder, and the final training objectives, starting with a formal problem definition. The whole framework is illustrated in Figure 2.

3.1. Problem Definitions

Given a long dialogue $\mathcal{D}_T = \{(u_i, s_i)_{i=0}^T\}$, u_i and s_i are i_{th} user utterance and system utterance respectively, for current dialogue context $\mathcal{D}_{context}$ consisting of $\mathcal{D}_{T-1}$ and u_t, the model is required to retrieve appropriate persona p (persona selection), knowledge k (knowledge selection), and sometimes even s_t itself (response selection) from a corresponding candidate pool to accomplish the dialogue and engage the users,

$$c_* = \arg\max_{c \in C} \text{sim}(q, c) \tag{1}$$

where C is the candidate pool, q is the query consisting of (D_{T-1}, u_t), c is a candidate from the pool which is composed of a sequence of words, and sim(q, c) is a similarity function[4] that measures the similarity between the query and candidate. The objective of this function is to find the candidate c in the pool C that has the highest similarity score with the query q. Here c can either be persona, knowledge, or response.

3.2. Context-adaptive Encoder

To locate the relevant contextual information in the lengthy dialogues, we choose to process individual utterances (u_i or s_i) through the encoder instead of combining them all into a single input, which often exceeds the maximum input limit or introduces unnecessary noises. In detail, we feed the utterance into the encoder with special indicator tokens [CLS] and [USR] or [SYS] at the beginning to obtain its representation [5]:

[3] It could be anything, here we denote persona, knowledge, and response.

[4] Without other statements, we adapt a simple dot product as the similarity function by default.

[5] To exploit discourse-level coherence among utterances (Gu et al., 2021b), we can add the 2-d position embedding based on the order of utterance in dialogue

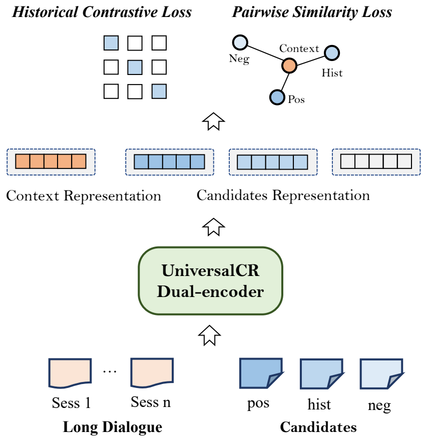

Figure 2: The proposed Universal Conversational Retrieval based on Dual-encoder architecture, with the goal of optimizing *historical contrastive loss* and *pairwise similarity loss* thanks to the introduction of historical candidates.

$$h_i = \textbf{Enc}(utter) \tag{2}$$

where $utter \in \{u_i, s_i\}$ and $h_i \in \{h_i^u, h_i^s\}$ accordingly. Considering that the dialogue may span multiple sessions (Xu et al., 2022)[6], we first divide the given long dialogue into the previous sessions D_{prev} and the current session D_{curr}. Inspired by lots of previous work which proved that the last utterance h_t^u plays a key role in retrieving relevant resource (Xu et al., 2022), here we regard it as the query to retrieve relevant utterance in the previous session:

$$H_{prev} = TopK_{h_j \in D_{prev}}(sim(h_t^u, h_j)) \tag{3}$$

Thus we can easily filter unrelated and redundant utterances in the previous session[7]. Then we concatenate retrieved H_{prev} with H_{curr} from the current session $[H_{prev}; H_{curr}]$ to form H_{hist} as key and value to learn contextualized embeddings of multi-turn dialogues, here H_{curr} consist of utterances from the current session except the last one.

$$h_{hist} = \textbf{Attn}(u_t, H_{hist}, H_{hist}) \tag{4}$$

In order to control the query and dialogue history contribution in the final representation, we add a gate after the attention block,

$$\begin{aligned} h_d &= \lambda * h_{hist} + (1-\lambda) * h_t^u \\ \lambda &= \sigma(\mathbf{w} * [h_{hist}; h_t^u]) \end{aligned} \tag{5}$$

where **w**, σ, h_d indicate a learnable parameter, sigmoid function and the final context representation respectively.

3.3. Candidates Encoder

To unify different candidate selection tasks into one framework, we design unique tokens to indicate each task. Specifically, we use [PERSONA] for persona selection, [KNOWLEDGE] for knowledge selection, and [RESPONSE] for response selection. Thus the model can easily recognize each task and perform the corresponding retrieval seamlessly. Then we feed it to the same encoder described in Eq.2

$$c_i = \textbf{Enc}(cand) \tag{6}$$

Where $cand$ consists of [CLS] [CANDIDATES] $w_1, w_2, ..., w_n$ in which [CANDIDATES] can be replaced by any candidate selection task indicator token described above. Thus, the framework can be easily extended to other candidate selection tasks and activated by specific candidate tokens without the necessity of training separate retrievers for each individual candidate selection task.

3.4. Training Objectives

Negative sample mining is vital to effectively train the dense retrieval model (Xiong et al., 2020). Previous work empirically showed the negatives ranked around the positives are generally more informative and less likely to be false negatives, requiring more attention (Zhou et al., 2022b). Building on these previous findings, we introduce a novel approach by considering the historically selected candidates as semi-hard negatives, which share closer semantics to the positives (as showed in Figure 1) than random negatives, to design two distinct loss objectives, namely *historical contrastive learning* and *pairwise similarity loss*.

Historical Contrastive Learning. Instead of predicting historically selected candidates (Wu et al., 2021), we take advantage of similar but different semantics between historical candidates and current ones by regarding the former as ***semi-hard negative samples***. For example, for the dialogue context in Figure 1, the P_n is the positive persona for the current turn, P_2 is the semi-hard negative, and other personas such as P_1 is the randomly sampled easy negative. With semi-hard negative samples in the batch, the model is optimized by:

$$\mathcal{L}_{hist} = \frac{e^{sim(h_d^i, c_i^+)}}{\sum_{j \in \mathcal{B}} e^{sim(h_d^i, c_j^+)} + e^{sim(h_d^i, c_i^-)}} \tag{7}$$

(discourse level) and order of word in utterance (utterance level).

[6] If not, we can simply regard one turn $[u_i, s_i]$ as one session unit.

[7] It is well noted that the retriever becomes more accurate since the encoder gets optimized as the training continues.

where *sim* is a similarity function, $\mathcal{B}$ is a mini-batch of examples, c_i^+ and c_i^- are positive candidates and semi-hard negative candidates for i_{th} dialogue context h_d^i. Once there are no semi-hard negatives or the semi-hard negative is the same as the positive, we directly use randomly sampled easy negative as c_i^-. In this way, the loss function simplifies to a conventional contrastive loss. By including historical negative candidates in the training data, the model is forced to learn to identify the subtle and key information required for the current turn instead of useless or redundant ones.

Pair-wise Similarity Loss. Instead of considering each context-candidate pair in isolation, we can alternatively focus on pairwise comparisons (Zhuang et al., 2022) to improve the accuracy of our ranking. To achieve this goal, we trained our model using a modified pairwise circle loss function (Sun et al., 2020). This loss function has a unified formula that can be used for two fundamental paradigms in deep feature learning: learning with class-level labels and learning with pairwise labels. The original loss function is shown below:

$$\mathcal{L}_{pair} = log[1 + \sum_{i=1}^{K} \sum_{j=1}^{L} e^{\gamma(s_n^j - s_p^i)}] \tag{8}$$

where γ is a scale factor s_n and s_p stands for two pairs where $s_n < s_p$. Specifically, here we have three different pairs as shown in Figure 2: *(context, positive candidate)*, *(context, historical candidate)* and *(context, negative candidate)*. As such, we can have the following preference ranking $r_{pos} > r_{hist} > r_{neg}$. Then the modified loss objective becomes:

$$\begin{aligned} \mathcal{L}_{pair} = log[1 + \sum_{i=1}^{K} e^{\gamma(s_{neg}^i - s_{hist}^i)} \\ + \sum_{j=1}^{L} e^{\gamma(s_{hist}^j - s_{pos}^j)}] \end{aligned} \tag{9}$$

Where K and L denote the number of similarity scores, s_{neg} denotes the similarity between context and the negative candidate, and so on[8]. Thus, the partial-order relationship between dialogue context and different candidates can be modeled and captured, which is the key point of a re-ranking module (Zhang et al., 2022).

Lastly, we combine these two losses at the same time to form the final training objective.

$$\mathcal{L} = \mathcal{L}_{hist} + \mathcal{L}_{pair} \tag{10}$$

Datasets	#Train	#Dev	#Test	#All
DuLeMon	28,243	1,993	2,036	30,202
KBP	4,788	589	584	5,961
Dusinc	2,565	319	359	3,243
KiDial	21795	2813	2580	27188
Diamante	29,758	2,548	2,556	34,862
KdConv	26,038	3,759	3,968	33,765
All	113187	12021	12083	137291

Table 1: The data statistics of used dialogue datasets, including persona-grounded dialogue dataset, knowledge-grounded dialogue dataset and also some conventional chit-chat dataset for response selection.

4. Experiments

In this section, we first introduce our six used datasets for persona selection, knowledge selection, and response selection tasks, and then present baselines and our main experimental results.

4.1. Datasets

DuLeMon. (Xu et al., 2022) a latest open-domain dialogue dataset with long-term persona memory in which a response is grounded on persona information that occurred in historical sessions, leading to better dialogue engagingness. ***persona selection***

Knowledge Behind Persona (aka, KBP). a dialogue dataset, in which the response is grounded on both the persona and its corresponding implicit knowledge (Wang et al., 2023a). We utilize this data to evaluate the out-of-domain and zero-shot performance of our model. ***persona selection***

KiDial-S. another collected knowledge-grounded dialogue set following Dai et al. (2022), which automatically turning knowledge documents into simulated multi-turn dialogues (Wang et al., 2023d). ***knowledge selection***

DuSinc. (Zhou et al., 2022a) an open-domain human-human dialogue dataset, where a participant can access the service to get the information needed for dialogue responses. ***knowledge selection***

Diamante. (Lu et al., 2022a), a chit-chat dataset by asking annotators to select or amend the model-generated candidate responses. Since the dataset contains human-generated responses and model-generated responses, we regard the former as positive samples, the latter as hard negative samples, and random responses as negative. ***response selection***

KdConv. (Zhou et al., 2020) a multi-domain dialogue dataset towards the multi-turn knowledge-driven conversation. Since it grounds the response to knowledge graphs, we do not consider it for knowledge selection. ***response selection***

[8] We found add $r_{pos} > r_{neg}$ order relationship in Eq. 9 explicitly can not bring significant improvement.

Model	Persona Sel.			Knowledge Sel.			Response Sel.		
	R@1	R@5	MRR	R@1	R@5	MRR	R@1	R@5	MRR
BM25	0.06	0.38	9.49	0.35	1.45	28.37	0.59	1.35	30.72
DPR	7.38	17.62	-	29.18	57.91	-	11.85	36.35	-
MultiCPR	10.70	17.27	-	41.45	58.81	-	9.65	19.15	-
RocketQAv1	21.39	52.02	34.23	37.86	65.91	42.86	21.46	71.20	41.92
SentenceBERT	18.99	47.64	32.91	43.57	86.59	61.13	35.45	73.59	51.86
Bi-Encoder	26.79	56.13	40.46	80.08	98.88	86.14	52.93	90.73	68.69
Poly-Encoder 16	26.26	55.76	40.08	79.11	99.11	85.61	51.17	**91.51**	67.83
Poly-Encoder 32	25.73	55.76	39.80	78.76	98.91	85.46	50.67	90.88	67.49
RocketQAv2	21.87	50.80	34.27	34.62	61.64	39.71	21.87	70.23	42.31
UniversalCR$_{single}$	28.43	55.17	40.99	85.65	98.72	90.69	57.20	85.68	69.29
UniversalCR$_{full}$	**28.73**	**55.99**	**41.47**	**86.67**	**99.22**	**91.45**	**59.94**	87.09	**71.73**

Table 2: The performance of our proposed model and baselines on dataset DuLemon (Xu et al., 2022), KiDial, and Diamante (Lu et al., 2022a), correspond to persona selection, knowledge selection, and response selection. UniversalCR$_{sigle}$ simply fine-tune our model on each dataset instead of all in UniversalCR$_{full}$.

4.2. Baselines

To better evaluate the performance of our proposed method, we conducted extensive experiments with different baselines, including both sparse and dense retrieval models: (1) **BM25** (Robertson and Zaragoza, 2009) (2) **DPR** (Karpukhin et al., 2020) (3) **SentenceBERT** (Reimers and Gurevych, 2019), (4) **PolyEncoder** (Humeau et al., 2019), (5) **Bi-Encoder**, (6) **RocketQAv1** (Qu et al., 2021) and **RocketQAv2** (Ren et al., 2021) [9], (7) **MultiCPR** (Long et al., 2022), (8) **UniversalCR**$_{single}$, and (9) **UniversalCR**$_{full}$. For the former 8 models, we train them independently for each candidate selection task, and we only adopt multi-task learning for **UniversalCR**$_{full}$. We chose these baselines since they mostly implemented typical dual-encoder architecture while adapting different interaction strategies: late interaction such as PolyEncoder, hard negative mining such as RocketQAv2, and knowledge distillation such as MultiCPR, without fancy and complicated architecture, resulting in more efficient computing. In addition, there are some methods that take advantage of more than one strategy such as RockerQAv2. We emphasize that almost all of these strategies can be plugged into our framework. We left this in our future work.

4.3. Implementation Details

We utilize LERT-base (Cui et al., 2022) as the backbone of our base version and all other baselines[10] for a fair comparison, the latest Chinese pre-trained language model that is trained on three types of linguistic features along with the original MLM pretraining task, bringing significant improvement over other variants (Cui et al., 2021). We use AdamW as our optimizer and we set the initial learning rate as 5e-5 with a linear decay. In particular, we use sequences of 64 tokens for each utterance and 512 for each candidate, we set the minimum window size as 5 and the training batch size as 64, and train our models for 5 epochs using the combination of DuLeMon, KiDial, and Diamante datasets. For the evaluation, we consider three retrieval metrics: R@1, R@5, and MRR following Humeau et al. (2019). We set the size of the candidate pool as 64 during all evaluations without other statements.

4.4. Main Result

With the setups described above, we fine-tuned the model on the datasets, and report the results in Table 2. It is exciting to see that UniversalCR$_{single}$ achieves better performance than all other baselines, particularly on knowledge selection (R@1 $\uparrow 5.57\%$) and response selection tasks (R@1 $\uparrow 4.27\%$), revealing the effectiveness of our framework. We also found that our method can not achieve consistent improvement at R@5, this is due to there being only one hard negative in the candidate pool for each query. We observed that the efficacy of our framework is positively correlated with the presence of more such difficult negative samples. In addition, we attribute the large improvement in knowledge and response selection tasks to the benefits of $\mathcal{L}_{pair}$ and $\mathcal{L}_{hist}$ since we model the more subtle order relationship between dialogue context and different candidates, which suit the design of corresponding datasets[11] and also downstream applications. Moreover, the

[9] https://github.com/PaddlePaddle/RocketQA

[10] https://github.com/ymcui/LERT

[11] since the knowledge and response selection is relatively sensitive with historical candidates compared with persona selection. For example, the candidates from

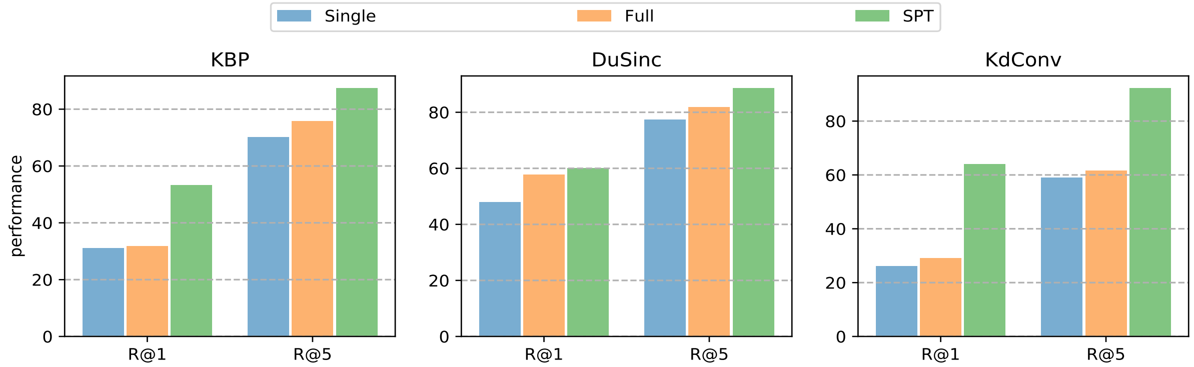

Figure 3: The zero-shot performance of UnifiedD$_{single}$ and Unified$_{full}$, and the supervised fine-tuning result of UnifiedD$_{full}$ on three ***New*** different datasets: Knowledge Behined Persona (persona selection), DuSinc (knowledge selection), and KdConv (response selection).

Model	P.R@1	K.R@1	R.R@1
UniversalCR$_{full}$	28.73	86.67	59.94
– *context enc.*	21.10	83.64	53.68
– $\mathcal{L}_{pair}$	28.52	85.27	56.65
– $\mathcal{L}_{hist}$	20.14	43.64	34.12

Table 3: Ablation Study. The - *context enc.* stands for considering all utterances in dialogue history by using a mean representation, and - $\mathcal{L}_{pair}$ and - $\mathcal{L}_{hist}$ means removing the corresponding loss constraint.

performance of UniversalCR$_{full}$ is comparable to or even better than UniversalCR$_{single}$, which was already a highly effective model. These findings suggest a promising path towards building robust and omnipotent dialogue retrieval systems which learn to perform diversity retrieval tasks to complete the dialogue successfully.

5. Analysis

In this section, we conduct detailed analysis experiments to demonstrate the superiority of our proposed method in various aspects: ablation study, zero-shot performance, and robustness.

5.1. Ablation Study

To test the effectiveness of different loss objectives, context-adaptive encoder, and multi-task learning, we conduct an ablation study by removing specific modules respectively and report the results on Table 3. The performance of all candidate selection tasks drops when removing the context-adaptive encoder (*PS: 28.73* → *21.10; KS: 86.67* → *83.64; RS: 59.94* → *53.68*) and drops to worst performance after removing $\mathcal{L}_{hist}$. We also find the model converges much faster with the help of $\mathcal{L}_{hist}$. Unsurprisingly, we observe a reduction when $\mathcal{L}_{pair}$ is removed. Thus we conclude that the combination of $\mathcal{L}_{pair}$ and $\mathcal{L}_{hist}$ leads to better performance and faster optimization while $\mathcal{L}_{pair}$ can well capture the order relationship between context and different candidates since it is a major driver of the performance gains achieved in knowledge selection (*56.65* → *59.94*) and response selection task (*85.27* → *86.67*) compared with persona selection task (*28.52* → *28.73*).

5.2. Zero-shot Performance

To investigate the zero-shot performance of our proposed model, we use another three datasets to evaluate under two settings: zero-shot (aka, Zero-shot) and supervised fine-tuning (aka, SPT)[12]. The result can be found in Figure 3. Notably, we observe that UniversalCR$_{full}$ achieves higher performance on **all** three datasets than UniversalCR$_{single}$, especially at the DuSinc dataset which is even comparable with SPT method [13], although there still is a gap between SPT and Zero-shot setting. The disparity between the Zero-shot and SPT settings exhibits a considerable magnitude in the KdConv dataset no matter R@1 or R@5, which we attribute to the distinct design employed in our response selection dialogue dataset, namely Diamante (Lu et al., 2022a), during the main experiment. Due to the involvement of human annotators in the process of selecting or amending model-generated candidate responses in a Diamante task, the approach to this task differs from the conventional response selection task, where negative responses are randomly selected. As a result, the transfer of knowl-

the sample of KiDial likely come from the same document while sharing similar semantics.

[12] We keep the setting as same as the main experiment.

[13] We surprisingly find that the zero-shot performance of UniversalCR$_{full}$ on DuSinc is even higher than the SPT of other baselines, for example, 47.57 R@1/64 of UniversalCR$_{full}$ v.s. 40.33 R@1/64 of Bi-encoder

Model	P.R@1	K.R@1	R.R@1
full concatenation	31.41	89.34	65.34
context enc.	28.73	86.67	59.94

Table 4: The performance of different ways to process the dialogue history in which the full concatenation can be viewed as our theoretical upper bound.

edge from the diamante task to the conventional response selection task is limited without any finetuning. However, we argue the design of Diamante is much better and more realistic in practice with the development of LLMs (Touvron et al., 2023).

5.3. The Effects of Previous Session

In addition, we compare the performance of our proposed framework under the different choices of K in Eq. 3 to investigate the influence of the previous session. We set the K as [1,2,3,4] to retrieve K most related utterance from the previous session and we also conduct an experiment in which we do not use any information from the previous session. To make a fair comparison, we evaluate the performance by loading the parameters from the latest three checkpoints and report the results in Figure 4. First of all, when comparing the performance of models that incorporate historical information with those that do not, it has been observed that the task of knowledge selection is particularly vulnerable and exhibits the greatest decrease in performance. This is hypothesized to stem from the fact that persona selection and response selection are comparatively more dependent on recent expressions, while knowledge selection differs in this regard. As conversations typically center around a specific topic, the inclusion of historical information can notably aid the model in effectively filtering out irrelevant information. Secondly, The model's performance is observed to degrade when the value of K is either too small or too large in general. This is in agreement with the notion that an excessively small value of K may result in important information being overlooked, while an overly large value may lead to the inclusion of noise data. Again, the knowledge selection task is more sensitive than the persona and response selection task with the choice of K. Due to these findings, we set the K as 3 to get the best average performance during the main experiment.

Besides that, we examined the effectiveness of a conventional dialogue processing method by directly concatenating all utterances together. However, we argue that this approach is impractical and inefficient for long dialogues in our setting. To provide a theoretical upper bound for our proposed method, we compared the performance of the two methods in different candidate selection tasks. The results, as presented in Table 4, indicate that the gap in performance between the two methods is much larger in the response selection task compared to the other two tasks. We suggest that this may be due to the historical turns in the context providing additional information about how humans respond, enabling a better understanding of the response selection task. Overall, our findings suggest that the proposed method is more effective for long dialogues and can achieve better performance with the assistance of utterances from the previous session, even there is still a gap with the theoretical upper bound.

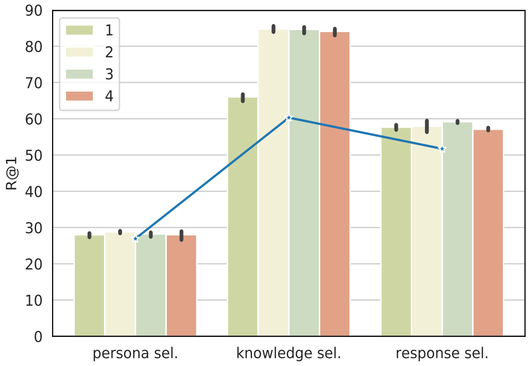

Figure 4: The Performance of UnifiedD with different k or without any utterance from the previous session. Blue line denotes the performance of Unified$_{full}$ without using any information from previous session. Here we report the R@1 metric.

5.4. The Effects of Different Sizes of Candidate Pool

For the retrieval task, the size of the candidate pool is very important considering the efficiency of the retrieval model. We compared the performance of UniversalCR$_{full}$ under different sizes of the candidate pool (from 256 to 2) with a bi-encoder which is commonly used in large-scale retrieval tasks in Figure 5. Surprisingly, we find our model can achieve par with the bi-encoder even when the size is over 128, and our model additionally demonstrates undeniable and consistent improvement when the size is relatively small (less than 64) in all three tasks. These findings suggest that our model has promising potential to serve as both a retrieval and re-ranker model simultaneously, thanks to the introduction of pairwise similarity loss.

6. Conclusion

In this paper, we present a novel universal conversational retrieval framework, which can be applied

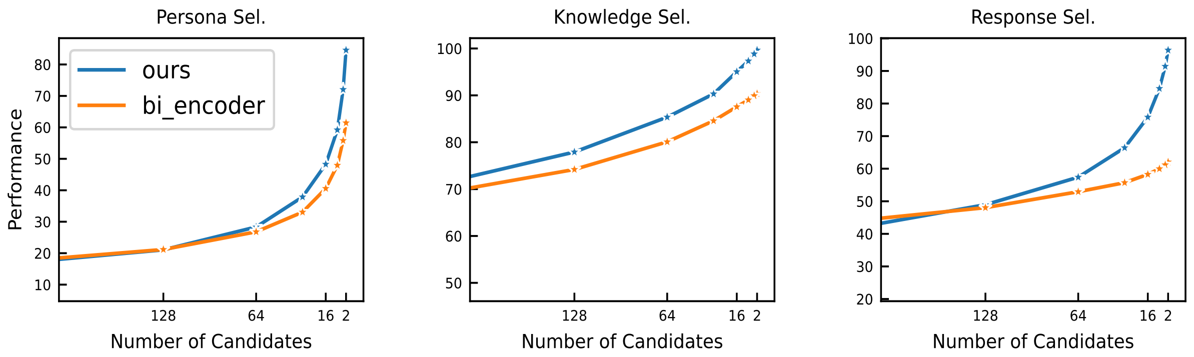

Figure 5: The Performance of UniversalCR$_{full}$ on different selection tasks: persona selection, knowledge selection, and response selection, with the number of candidates ranging from 256 to 2.

to retrieve diverse external resources to complete the conversation successfully at the same time. We conduct extensive experiments on three major candidate selection tasks, including persona selection, knowledge selection, and response selection tasks. The experimental results suggest the effectiveness and potential of our framework to be a robust and omnipotent conversational retrieval system. In addition, we also found the framework demonstrates strong zero-shot performance and robustness serving as a re-ranker and a retriever simultaneously. We left other more complicated architectural improvements e.g. interactions between two encoder towers in future work.

7. Acknowledgement

We would like to express our heartfelt gratitude to all anonymous reviewers for their insightful comments and suggestions. This research work is partially supported by CUHK direct grant No. 4055209 and CUHK Knowledge Transfer Project Fund No. KPF23GWP20.

8. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Wanxiang Che, Shijin Wang, and Ting Liu. 2022. Lert: A linguistically-motivated pre-trained language model.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Mike Green, Qazi Rashid, and Kelvin Guu. 2022. Dialog inpainting: Turning documents to dialogs. In *International Conference on Machine Learning (ICML)*. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1845–1854, Hong Kong, China. Association for Computational Linguistics.

Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021a. Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In *SIGIR '21: The 44th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pages 565–574. ACM.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021b. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14, pages 12911–12919.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Songbo Hu, Ivan Vulić, Fangyu Liu, and Anna Korhonen. 2022. Reranking overgenerated responses for end-to-end task-oriented dialogue systems.

Kai Hua, Zhiyuan Feng, Chongyang Tao, Rui Yan, and Lu Zhang. 2020. Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 525–534. ACM.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert.

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10278–10287, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3971–3980, Online. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. ACM.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *SIGIR*, pages 3046–3056. ACM.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2022a. Towards boosting the open-domain chatbot with human feedback.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, and Haifeng Wang. 2022b. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. Bert with history answer embedding for conversational question answering. In *Proceedings of the*

42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023a. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore. Association for Computational Linguistics.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems.

Hongru Wang, Huimin Wang, Lingzhi Wang, Minda Hu, Rui Wang, Boyang Xue, Hongyuan Lu, Fei Mi, and Kam-Fai Wong. 2023b. Tpe: Towards better compositional reasoning over conceptual tools with multi-persona collaboration.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023c. A survey of the evolution of language model-based dialogue systems.

Rui Wang, Jianzhu Bao, Fei Mi, Yi Chen, Hongru Wang, Yasheng Wang, Yitong Li, Lifeng Shang, Kam-Fai Wong, and Ruifeng Xu. 2023d. Retrieval-free knowledge injection through multi-document traversal for dialogue models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6619, Toronto, Canada. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. 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. Association for Computational Linguistics.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 126–142. Springer.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 829–838, New York, NY, USA. Association for Computing Machinery.

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022. Hlatr: Enhance multi-stage text retrieval with hybrid list aware transformer reranking.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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. Association for Computational Linguistics.

Zhenyu Zhang, Tao Guo, and Meng Chen. 2021. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, page 3647–3651, New York, NY, USA. Association for Computing Machinery.

Han Zhou, Xinchao Xu, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Siqi Bao, Fan Wang, and Haifeng Wang. 2022a. Link the world: Improving open-domain conversation with dynamic spatiotemporal-aware knowledge.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-rong Wen, and Nan Duan. 2022b. SimANS: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 548–559, Abu Dhabi, UAE. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2022. Rankt5: Fine-tuning t5 for text ranking with ranking losses.

9. Language Resource References

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents.

Hongru Wang, Wai-Chung Kwan, Min Li, Zimo Zhou, and Kam-Fai Wong. 2024. Kddres: A multi-level knowledge-driven dialogue dataset for restaurant towards customized dialogue system. *Computer Speech Language*, 87:101637.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.