Learning to Match Representations is Better for End-to-End Task-Oriented Dialog System

Wanshi Xu, Xuxin Cheng, Zhihong Zhu, Zhanpeng Chen, Yuexian Zou*
ADSPLAB, School of ECE, Peking University, China
{xwanshi, chengxx, zhihongzhu, troychen927}@stu.pku.edu.cn
zouyx@pku.edu.cn

Abstract

Due to the rapid development with pre-trained language models, fully end-to-end Task-Oriented Dialogue (TOD) systems exhibit superior performance. How to achieve the ability to retrieve entities in cross-domain large-scale databases efficiently is a key issue. Most existing end-to-end Task-Oriented Dialogue systems suffer from the following problems: The ability to handle erroneous but easily confused entities needs to be improved; Matching information between contexts and entities is not captured, leading to weak modeling of domain-invariant and interpretable features, making it difficult to generalize to unseen domains. In this paper, we propose a method for knowledge retrieval driven by matching representations. The approach consists of a matching signal extractor for extracting matching representations between contexts and entities that have generic conceptual features and hence domain invariant properties, and an Attribute Filter for filtering irrelevant information to facilitate the re-selection of entities. Experiments on three standard benchmarks at the dialogue level and on large knowledge bases show that our retriever performs knowledge retrieval more efficiently than existing approaches.

1 Introduction

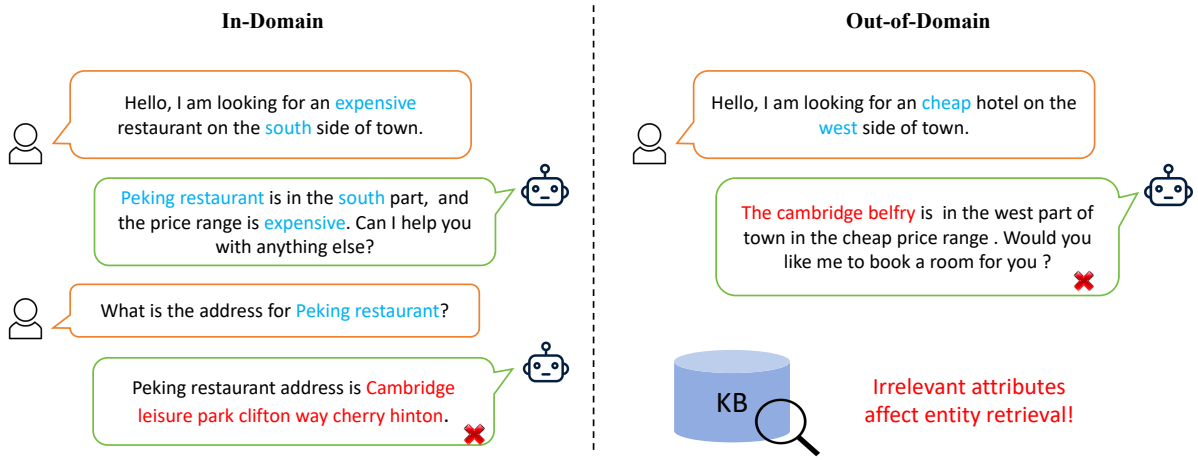
Task-oriented dialogue systems (Zhang et al., 2020) are designed to help users complete certain specific tasks, such as table booking, hotel booking, ticket booking, and online shopping. Traditional task-oriented dialogue systems are developed through dialogue state tracking (Kenton and Toutanova, 2019; Wu et al., 2019a), dialogue strategies (Takanobu et al., 2019) and natural language generation (Wen et al.). These modules require annotations to train and need to generate belief states to query the database to generate response. In contrast, fully end-to-end task-oriented dialog system directly encodes KB and uses a neural network to query the KB in a differentiable manner, and it can directly generate system response given only dialogue history and the corresponding KB.

Although the end-to-end paradigm gains increasing attention, retrieving the correct knowledge from external databases becomes a key limiting factor for the performance of such models due to the lack of belief states as supervisory signals. Existing end-to-end TOD systems can be divided into two categories based on the relationship between knowledge retrieval and corresponding generation. The first class of approaches typically integrates the processes of knowledge retrieval and response generation and trains them under the supervision of reference responses (Madotto et al., 2018; Qin et al., 2020; Raghu et al., 2021; Xie et al., 2022; Wu et al., 2022; Madotto et al., 2020; Huang et al., 2022).The second class of approaches decouples knowledge retrieval from the corresponding generation and explicitly extracts the supervisory signals from the response generation to improve the retrieval process, which alleviates the above problem to some extent. Q-TOD (Tian et al., 2022) extracts the essential information from the dialogue context into a query, which is further employed to retrieve relevant knowledge records for response generation.

However, although these methods are able to provide supervisory signals for retrieval, they only use rough dot product interactions to compute the similarity when filtering the candidate entities for generation, which is vulnerable to receiving interference from irrelevant attributes and has poor domain generalization, as shown in Fig.1. Whereas the information extracted by response generation originally comes from the candidate entities, the effectiveness of this supervisory signal is therefore closely related to the accuracy of the candidate entities' input to the response generator.

*Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 10409–10419



Retrieved Entities

	name	food	address	postcode	area	pricerange
1	frankie and bennys	italian	cambridge leisure park clifton way cherry hinton	cb17dy	south	expensive
2	**peking restaurant**	**chinese**	**10 homerton street city centre**	**cb28nx**	**south**	**expensive**
3	taj tandoori	indian	64 cherry hinton road cherry hinton	cb17aa	south	expensive

Figure 1: Visualisation of the impact of irrelevant attributes on entity selection and entity retrieval errors for different domain extensions. Attributes colored in grey indicate attributes with low context relevance scores. Wrong answers are caused by the fact that the retriever views every attribute in the entity equally. In the in-domain case, the retrieval of the entity is interfered by irrelevant attributes, while in the out-of-domain case, the entity will directly have errors at the entity level.

In this paper, we propose a knowledge retrieval method that is sensitive to matching attributes to improve generalisation and interpretability by capturing matching information. Matching information can be used to filter mismatched parts of entities to improve entity selection. We first filter a batch of candidate entities by calculating the rough similarity between the context and the entity, and then calculate the match representation between the context and the entity and compute its match score with each attribute in the entity to obtain the relevance of the attribute to the context. Based on this relevance score, we filter out irrelevant attributes and recalculate the top-K entities used to generate the response. The top-K entities are then collocated in a certain form and fed into the generator. We compare our system with other systems on three benchmark datasets(Eric et al., 2017; Wen et al., 2017; Eric et al., 2020). Our main contributions can be summarised as:

(i) We extract fine-grained matching information between contexts and entities, not just coarse-grained interactions to judge relevance;

(ii) We discover that attributes can also facilitate the selection of entities, by filtering irrelevant attributes first to select entities;

(iii) The experimental results show that our system achieves state-of-the-art performance both at the dialog-level and on the total knowledge database.

2 Related Work

2.1 End-to-End Task-Oriented Dialog

The development of end-to-end trainable methods to generate responses in conjunction with external knowledge bases has received increasing attention relative to traditional modular approaches in task-oriented dialogue systems. Some works encode the knowledge base (KB) with memory networks, and KB records are selected using attention weights between dialogue context and memory cells (Qin et al., 2019; Wu et al., 2019b; Raghu et al., 2021). Some work has explored the use of tandems of knowledge bases and dialogue contexts, which are used as input to pre-trained language models (Xie et al., 2022; Rony et al., 2022). Additionally, the knowledge base is stored in model parameters for implicit retrieval during response generation (Madotto et al., 2020; Huang et al., 2022). However, these methods generally blend entity retrieval and response generation during response generation, which leads to sub-optimal retrieval performance when large-scale knowledge bases are

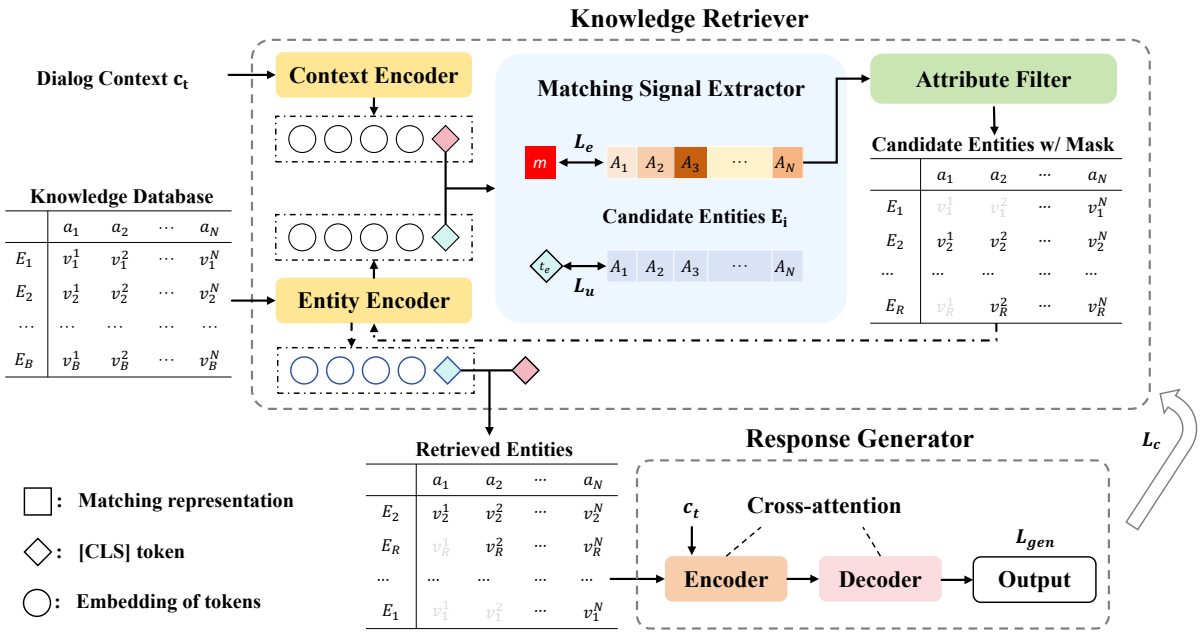


Figure 2: An overview of our end-to-end task-oriented dialogue system, which consists of a matching signal extractor an attribute filter, and a response generator. The match signal extractor is used to extract matching information, homogeneous entity global information, and highlight local matching information.

provided.

The second class of approaches decouples knowledge retrieval from the corresponding generation and explicitly extracts the supervisory signals from the response generation to improve the retrieval process, which alleviates the above problem to some extent. Q-TOD (Tian et al., 2022) extracts the essential information from the dialogue context into a query, which is further employed to retrieve relevant knowledge records for response generation. MAKER (Wan et al., 2023) introduces a multi-grained retrieval with both entity and attribute selection. Shi et al. (2023) proposed a dual feedback network to obtain the supervision signal of the corresponding generator.Shen et al. (2023) propose the application of maximal marginal likelihood to train a perceptive retriever by utilizing signals from response generation for supervision.

2.2 Dense Retriever

Because of the excellent performance in efficiency and effectiveness, dense retrieval has been widely used in first-stage retrieval that efficiently recalls candidate documents from the large corpus (Karpukhin et al., 2020).

As deep neural networks have achieved promising results in various NLP tasks, they have also been explored for information retrieval applications. One of the mainstream approaches to information retrieval is to build retrievers using dual-encoder architecture (Yih et al., 2011). Gillick et al. (2019) employs the dual encoder architecture for separately encoding mentions and entities into high-dimensional vectors for entity retrieval. However, through this coarse interaction metric calculation, fine-grained information about entities in the modeling external database is often ignored. In order to learn entity representations that can match different mentions, Liu et al. (2023) proposes a multi-view augmented distillation framework, where entities are divided into multiple views, while, at the same time, a global view is retained to prevent the spread of uniform information. In document retrieval, BERM (Xu et al., 2023) improves the generalisation of dense retrieval by capturing the matching signals, splitting a single paragraph into multiple units, and proposing two unit-level requirements to represent training constraints. Motivated by these works, we can conclude that the homogeneous representation of global information and the identification of locally relevant matches of an entity are two key factors in entity retrieval.

3 Method

As shown in Fig. 2, our system first retrieves an initial rough batch of candidate entities from the knowledge base. Then, the matching signal extractor performs global information decentralization and local information sharpening on the retrieved candidate entities to extract the relevant informa-

tion of their attributes. Then, the entity enters the attribute filter to eliminate irrelevant attributes. After filtering out the irrelevant attributes, the entities are reordered to select the top-K candidate entities. These candidate entities along with the dialogue context are then fed into the generator model to generate replies.

3.1 Problem Definition

Given a dialog $\mathcal{D} = \{U_1, R_1, ..., U_T, R_T\}$ of T turns, where U_t and R_t are the t-th turn user utterance and system response, respectively. We use C_t to represent the dialog context of the t-th turn, where $C_t = \{U_1, R_1, ..., U_{t-1}, R_{t-1}, U_t\}$. An external knowledge base (KB) is provided in the form of a set of entities, i.e., $\mathcal{K} = \{E_1, E_2, ..., E_B\}$, where each entity E_i is composed of N attribute-value pairs, i.e., $E_i = \{a^1, v_i^1, ..., a^N, v_i^N\}$. The goal of an end-to-end task-oriented dialogue system model is to learn a mapping that takes the dialogue context C_t and knowledge base $\mathcal{K}$ as input and generates an information response R_t.

$$R_t = f(C_t, \mathcal{K}) \tag{1}$$

3.2 Knowledge Retriever for Entities

Dual encoders are the most commonly used architecture for large-scale retrieval, so we model the entity selector as a dual encoder architecture, where one encoder, Enc_c, is used to encode the dialogue context, and the other encoder, Enc_e, is used to encode each entity in the knowledge base, both of which are dense vectors. To encode an entity, we concatenate the attribute-value pairs of this entity into a sequence and pass it to Enc_e. The selection score $s_{t,i}$ for entity E_i is defined as the dot product between the context vector and the entity vector as:

$$s_{t,i} = \mathrm{Enc}_c\left(C_t\right)^T \mathrm{Enc}_e\left(E_i\right) \tag{2}$$

Then, the top-K candidate entities are obtained by:

$$\mathcal{E}_t = \mathrm{Top}\, K\left(s_{t,i}\right) = \{E_1, \ldots, E_K\} \tag{3}$$

We implement Enc_c and Enc_e with a pre-trained language model and allow them to share weights, where the final '[CLS]' token representation is used as the encoder output. We follow Shi et al. (2023) to initialize the pre-trained model.

3.3 Matching Signal Processing

We process the match signals of candidate entities by first feeding them into the Match Signal Extractor to homogenize the entity global information and capture local relevance information, and then feeding them into the Attribute Filter to filter irrelevant attributes. Finally, we re-encode the entities after masking irrelevant attributes to compute a score that is used to select the entities that are finally used to generate responses. **Global Information Decentralization**: This process involves distributing or dispersing the global, or overall, information content of an entity across its attributes. It aims to ensure that the entity representation includes a comprehensive summary of all its attributes without bias towards any single aspect. This step is crucial for capturing the full semantic context of the entity. **Local Information Sharpening**: After decentralizing the global information, this process focuses on highlighting and refining the local, or specific, information within the entity's attributes. It sharpens the distinctions between different attributes and their relevance to the context, thus enhancing the ability to identify and match the most contextually relevant aspects of the entity.

Matching Signal Extractor Firstly, entity representations unify and summarise the semantics of each attribute within an entity, resulting in a comprehensive and refined representation of the entity.

After encoding the entity E_i using the encoder described in Section 3.2, we obtain the hidden state $Z = \mathrm{Enc}_e\left(E_i\right)$, and we use t_e, t_c to denote the "[CLS]" token of $\mathrm{Enc}_e\left(E_i\right)$ and $\mathrm{Enc}_c\left(C_t\right)$. The embeddings A of attributes tokens in an entity can be obtained from Z as the segmentation of different attributes:

$$A = \{A_1, A_2, \ldots, A_N\},$$

where A_i is the embedding of the corresponding attribute and it is the average pooling of the embeddings of tokens in the attribute. In order for the entity representation to carry the information of each attribute uniformly, the relationship between $\mathrm{Enc}_e\left(E_i\right)$ and A is described by the loss function as:

$$\mathcal{L}_{\mathrm{u}} = D_{KL}\left[b \| \,\mathrm{sim}\left(t_e, A\right)\right], \tag{4}$$

where $D_{KL}[\cdot\|\cdot]$ is KL-divergence loss, $b = \left[\frac{1}{N}, \ldots, \frac{1}{N}\right]$ is a uniform distribution with equal values and $\mathrm{sim}\left(t_e, A\right) = \{\mathrm{dot}\left(t_e, A_i\right) \mid A_i \in A\}$ is a distribution to represent the semantic similarity between t_e and $A_i \in A$, $\mathrm{dot}(\cdot,\cdot)$ is dot product.

Next, we extract matching units m within the entity that capture the contextual information. By balancing the semantic representation of attribute

units in a fair manner, the element-wise multiplication between t_e and t_c amplifies similarity parts of the tensor, namely the semantic representation of basic matching units, as demonstrated in (Xu et al., 2023). Therefore, we introduce this matching mechanism as:

$$m = GELU\left(t_e \odot t_c\right) \tag{5}$$

where $\odot$ denotes element-wise multiplication operator and GELU represents the activation function (Hendrycks and Gimpel, 2016)for introducing stochasticity. Pseudo-labeling is employed for supervision by constructing an N-dimensional 0-1 vector based on the presence or absence of each attribute in the context. For the pair (c_t, E_i), y_i in attribute label list $Y = \{y_1, ..., y_N\}$ for matching attribute is that if A_i appear in the C_t, $y_i = 1$, otherwise, $y_i = 0$. During training, we optimize the semantic distance between m and each attribute, namely the attributes scores s_a using cross-entropy loss to identify the corresponding matching attributes:

$$\mathcal{L}_e = -\sum_{i=1}^{n} y_i \log\left(\mathrm{dot}\left(m, A_i\right)\right) \tag{6}$$

where $A_i \in A, y_i \in Y, s_a = (\mathrm{dot}\left(m, A_i\right))$. m is only used as the constraint in training but has important implications for inference. It is because that m is the combination of text representations (t_c and t_e). The optimization for m is training the text encoder to output the text representation that is suitable for matching to improve the generalization ability.

The matching signal extractor enables the model to implicitly aggregate the semantics of each attribute within the entity into the entity representation while ensuring the semantic orthogonality of the attributes. In the dot product between the context and the entity representation, the semantic information of the basic matching units is preserved, while the semantic information of other units is masked.

Attribute Filtering and Entity Reselection After obtaining attribute scores for each entity, we select the retrieved entities by pruning attributes with importance scores greater than a predefined threshold t. Specifically, we mask irrelevant attributes to obtain a new entity set with irrelevant attributes masked out. We then re-retrieve these entities using the retriever to obtain entity scores $\hat{s}_t$ and rank them. Finally, we select the top-r entities $\hat{\mathcal{E}}_t$ and feed them into the generator to generate the response. In this way, the score of the re-selected entity no longer receives the influence of extraneous attributes, and the extraneous attributes are masked before being fed into the generator.

3.4 Response Generator

Following (Wan et al., 2023), we employ a modified sequence-to-sequence structure for the response generator to facilitate direct interaction between dialog context and retrieved entities.

Generator Encoder Each entity $\hat{E}_i$ in $\hat{\mathcal{E}}_t$ is first concatenated with dialog context C_t and encoded into a sequence of vector representations $H_{t,i}$:

$$H_{t,i} = \mathrm{Enc}_g\left(\left[C_t; \hat{E}_i\right]\right) \tag{7}$$

where Enc_g represents the encoder of the response generator. Then, the representations of all retrieved entities are concatenated into H_t :

$$H_t = [H_{t,1}; \ldots; H_{t,K}] \tag{8}$$

Generator Decoder Taking H_t as input, the generator decoder Dec_g produces the system response token by token. During this process, the decoder not only attends to the previously generated tokens through self-attention but also attends to the dialogue context and retrieved entities by cross-attention, we find KB-related tokens in the response and regard the cross-attention scores from these tokens to each retrieved entity as the knowledge to distill:

$$\mathcal{L}_c = \mathcal{D}_{KL}(\hat{s}_t || c_t). \tag{9}$$

where c_t Represents cross attention score.

The probability distribution for each response token in R_t is defined as:

$$P\left(R_{t,i}\right) = \mathrm{Dec}_g\left(R_{t,i} \mid R_{t,<i}, H_t\right) \tag{10}$$

We train the response generator by the standard cross-entropy loss as:

$$\mathcal{L}_{gen} = \sum_{i=1}^{|R_t|} -\log P(R_{t,i}), \tag{11}$$

where $|R_t|$ denotes the length of R_t.

Lastly, the overall loss of the system is the sum of $\mathcal{L}_u$, $\mathcal{L}_e$, and response generation loss $\mathcal{L}_{\text{gen}}$:

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_e + \mathcal{L}_{\text{gen}} + \mathcal{L}_c \tag{12}$$

Model	MWOZ		SMD		CamRest	
	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1
DF-Net (Qin et al., 2020)	9.40	35.10	14.40	62.70	-	-
GPT-2+KE (Madotto et al., 2020)	15.05	39.58	17.35	59.78	18.00	54.85
EER (He et al., 2020b)	13.60§	35.60§	17.20§	59.00§	19.20§	65.70§
FG2Seq (He et al., 2020a)	14.60§	36.50§	16.80§	61.10§	20.20§	66.40§
CDNET (Raghu et al., 2021)	11.90	38.70	17.80	62.90	21.80	68.60
GraphMemDialog (Wu et al., 2022)	14.90	40.20	18.80	64.50	22.30	64.40
ECO (Huang et al., 2022)	12.61	40.87	-	-	18.42	71.56
DialoKG (Rony et al., 2022)	12.60	43.50	20.00	65.90	23.40	**75.60**
UnifiedSKG (T5-Base) (Xie et al., 2022)	-	-	17.41	66.45	-	-
UnifiedSKG (T5-Large) (Xie et al., 2022)	13.69*	46.04*	17.27	65.85	20.31*	71.03*
Q-TOD (T5-Base) (Tian et al., 2022)	-	-	20.14	68.22	-	-
Q-TOD (T5-Large) (Tian et al., 2022)	17.62	50.61	21.33	71.11	23.75	74.22
DF-TOD (T5-Base) (Shi et al., 2023)	**18.26**	52.52	24.12	69.36	25.85	72.83
DF-TOD (T5-Large) (Shi et al., 2023)	18.48	53.17	25.10	71.58	26.00	74.04
MK-TOD (T5-Base) (Shen et al., 2023)	17.33	51.86	24.77	67.86	**26.76**	73.60
MK-TOD (T5-Large) (Shen et al., 2023)	17.55	52.97	25.43	73.31	**26.20**	71.72
MAKER (T5-Base) (Wan et al., 2023)	17.23	**53.68**	24.79	69.79	25.04	73.09
MAKER (T5-Large) (Wan et al., 2023)	18.77	54.72	25.91	71.30	25.53	74.36
Ours (T5-Base)	17.63	53.65	**25.06**	**71.79**	25.54	**73.69**
Ours (T5-Large)	**18.84**	**54.80**	**26.91**	**73.39**	25.73	74.86

Table 1: Overall results of end-to-end TOD systems with dialog-level knowledge bases on MWOZ, SMD, and CamRest. The best scores are highlighted in bold, and the second-best scores are underlined. †, ‡, §, ∗ indicates that the results are cited from (Qin et al., 2019), (Qin et al., 2020), (Raghu et al., 2021), and (Tian et al., 2022), respectively.

Model	MWOZ		CamRest	
	BLEU	Entity F1	BLEU	Entity F1
FG2Seq	10.74	33.68	19.20	59.35
CDNET	10.90	31.40	16.50	63.60
Q-TOD	16.67	47.13	21.44	63.88
MK-TOD (T5-Base)	17.56	50.09	26.85	**73.51**
MK-TOD (T5-Large)	17.40	53.26	27.82	71.98
DK-TOD (T5-Base)	**17.61**	51.61	**27.39**	70.74
DK-TOD (T5-Large)	18.36	52.96	26.61	73.58
MAKER (T5-Base)	16.25	50.87	26.19	72.09
MAKER (T5-Large)	18.23	52.12	25.34	72.43
Ours (T5-Base)	16.76	**51.87**	27.09	72.47
Ours (T5-Large)	**18.77**	**53.82**	**27.88**	**73.93**

Table 2: Overall results of end-to-end TOD systems with the total knowledge base on MWOZ and CamRest, respectively. The best scores are highlighted in bold, and the second-best scores are underlined.

4 Experiments

4.1 Datasets

We evaluate our system on three multi-turn task-oriented dialogue datasets: MultiWOZ 2.1 (MWOZ) (Eric et al., 2020), Stanford Multi-Domain (SMD) (Eric et al., 2017), and CamRest (Wen et al., 2017). Each dialog in these datasets is associated with a condensed knowledge base, which contains all the entities that meet the user goal of this dialog. For MWOZ, each condensed knowledge base contains 7 entities. For SMD and CamRest, the size of condensed knowledge bases is not fixed: it ranges from 0 to 8 with a mean of 5.95 for SMD and from 0 to 57 with a mean of 1.93 for CamRest. We follow the same partitions as previous work (Raghu et al., 2021).

BLEU (Papineni et al., 2002) and Entity F1 (Eric et al., 2017) are used as the evaluation metrics. BLEU measures the fluency of a generated response based on its n-gram overlaps with the gold response. Entity F1 measures whether the generated response contains correct knowledge by micro-averaging the precision and recall scores of attribute values in the generated response.

4.2 Implementation Details

We employ BERT (Devlin et al., 2019) as the encoder of our entity selector and attribute selector and employ T5 (Raffel et al., 2020) to implement the response generator. All these models are fine-tuned using AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 64. We train these models for 15k gradient steps with a linear decay learning rate of 10^{-4}. We conduct all experiments on a single 24G NVIDIA RTX 3090 GPU and select the best checkpoint based on model per-

formance on the validation set.

4.3 Results

We conduct separate experiments on the dialogue-level dense knowledge database and the total knowledge database. First, we show the overall performance of the evaluated systems when a subsidiary knowledge base is provided for each dialogue. Then, we replace the database with the total knowledge base that includes all entities of the dataset for comparison.

The results on the dialog-level database are shown in Table 1. We find that our system achieves state-of-the-art (SOTA) performance on several datasets when using T5-Large as the generator model. Specifically, on MWOZ, our system outperforms the previous SOTA (i.e., MAKER) in terms of BLEU and Entity F1 by 0.07 and 0.08 points, respectively. On SMD, compared to the previous SOTA (i.e., MAKER), BLEU improves by 1 point and Entity F1 improves by 0.08 points. On CamRest, however, our system is slightly inferior relative to MK-TOD. This is due to the fact that many of the dialogues in CamRest contain very small knowledge bases of only 1-2 entities, in which case improving the retrieval of entities does not improve the performance metrics of the dialogues much. By comparing the experimental results we can find that our results do not show much improvement relative to MK-TOD, DK-TOD, and MAKER, the reason behind this may be that most of these three models focus on constructing a mechanism for distilling knowledge from generation to retrieval, while our model focuses more on improving the accuracy of retrieved entities.

Most of the previous baselines have been performed with each dialogue corresponding to a condensed knowledge base. However, constructing an exclusive and relevant database for each dialogue in a real scenario is difficult, and training on a small and precise database leads to poor scalability of the model. In the future, it is more likely that the system will face large knowledge bases across domains. Therefore, we collected the entities of all the dialogues in the original dataset, constructed a total knowledge base, and implemented several recognized E2E TOD systems on MWOZ and CamRest, respectively, and examined the performance of these systems, with results shown in Table 2.

We compared it only to those systems that also implemented an overarching database. We found that the advantage of our system over the other benchmark systems is more pronounced when using the full knowledge base. Comparing the results in Table 1 and Table 2, we notice that our system has a greater improvement in experimental results on the total database compared to other systems. For example, on MWOZ, our system improves 0.41 and 0.56 points on BLEU and Entity F1, respectively; on CamRest, our system improves 0.06 and 0.35 points on BLEU and Entity F1, respectively. This may be because on large-scale databases, we filter attributes by extracting fine-grained matching information between entities and contexts to reselect entities, which is somewhat generic. These observations validate the superiority of our system when applied to large-scale knowledge bases and the feasibility of applying it to real-world scenarios. In addition, we also find that our model is better able to gain advantages in both metrics relative to other models, regardless of the type of knowledge database.

5 Analysis

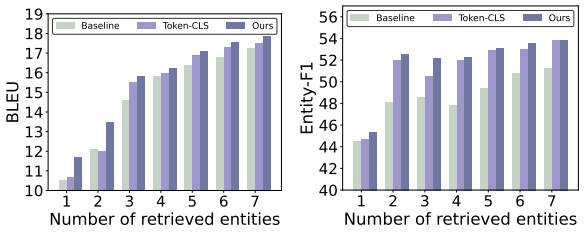
5.1 Ablation Study

We conducted a disambiguation study of our system on MWOZ using both the dialog-level knowledge base and the total knowledge base because of its relatively uniform distribution of dialogue-level databases. The results are shown in the first and second parts of Table 3, respectively.

In order to verify that filtering irrelevant attributes by extracting matching information can promote entity retrieval and thus improve the quality of generation, we design the following ablation experiments: firstly, in order to verify that irrelevant attributes affect the selection of entities, we only cancel the process of entity re-ranking after

Model	BLEU	Entity F1
$\text{Ours}_{\text{dialog-level}}$	18.84	54.80
w/o re-rank	17.86 (↓ **0.98**)	53.79 (↓ **1.01**)
w/o L_u	18.19 (↓ **0.65**)	54.41 (↓ **0.39**)
w/o L_e	17.68 (↓ **1.16**)	53.79 (↓ **1.01**)
$\text{Ours}_{\text{total}}$	18.77	53.82
w/o re-rank	17.43 (↓ **1.34**)	53.23 (↓ **0.59**)
w/o L_u	18.34 (↓ **0.43**)	53.48 (↓ **0.34**)
w/o L_e	17.18 (↓ **1.59**)	51.95 (↓ **1.87**)

Table 3: Results of ablation study on MWOZ with T5-base, where "w/o" means without, "re-rank" denotes candidate entities with masks are recoded for similarity calculation and sorted to re-filter them, L_u,L_e can be found in Section 3.



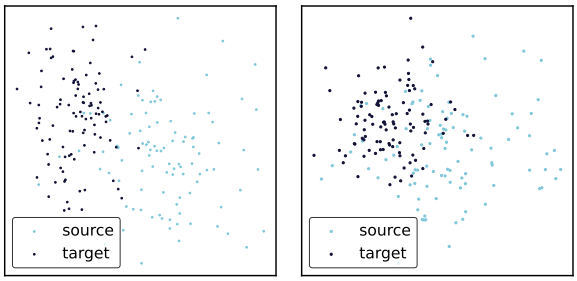
(a) BLEU (b) Entity F1

Figure 3: Performance of different retrieval methods as the number of retrieved entities changes on the full knowledge base in BLEU (a) and Entity F1 (b) scores.

masking irrelevant attributes, and feed it directly into generation, and the results of the experiments show that we found that after removing entity re-ranking (*w/o* re-rank) the system performance decreases significantly by 1.34, 1.01 on BLEU and Entity F1, respectively. This indicates that filtering entities only by relevant attributes change the ordering of the original candidate entities after extracting the matching part between entities and contexts at a fine-grained level, thus proving that filtering entities by attributes is superior to the context-entity interaction at a coarse-grained level; and then, in order to prove the effectiveness of match signal extraction, we design to remove the loss of match signal and filter entities directly by coarse-grain method. Removing L_g and L_e experimental results show that the highest decreases on BLEU and Entity F1 were 0.65 and 0.39, 1.59 and 1.87, respectively. This explains that the extraction of matching information needs to be performed on the premise that the entity information can be represented uniformly by the attributes contained in it, which captures the global information of the entity, while L_e highlights the relevant matching part of the entity, which is indispensable in a retriever.

5.2 Matching Signal Extractor

In computing the matching signals between entities and contexts, we capture the matching representations by homogenizing the global information and extracting the local information and then determining which attributes need to be retained based on a given threshold. We compare different approaches to obtaining matching signals. We designed experiments to compare methods to compute similarity directly from attribute tokens and contextual "[CLS]" tokens as a criterion for filtering attributes. Also, in order to investigate the impact of different numbers of retrieved entities on the system performance as the number of entities increases, we report the entity F1 and BLEU scores of the above



(a) m Representation (b) m Representation(Ours)

Figure 4: T-SNE of the matching representations m for source and target domains.

retrieval methods, and we observe in Fig. 3 that the entity F1 and BLEU scores of all the three methods increase with the increase of the number of entities, while our retriever always achieves the best performance. In Fig. 3, we observe a positive correlation between the entity F1 score and the number of entities over the range of the number of entities we tested, but it does not imply that the higher the number of detected entities the better, and that noisy entities may be introduced as the number of entities increases. We also observe that our system outperforms other methods when the number of entities is small, probably because our system is able to accurately filter out information about irrelevant attributes. The fewer the number of retrieved entities, the more accurate the entity information is required. Therefore, our method achieves good performance when a few entities are retrieved (i.e., when the error tolerance is low).

5.3 Domain-invariant Representation

In order to verify that our method improves the generalization of the model by extracting domain invariant information, we wanted to design experiments to visualize the distribution of the matching feature m. We first used T-SNE to visualize the entity representations in the source and target domains encoded by the baseline (MAKER) and our model, respectively, as shown in Fig. 4. We can find that the representations of the matching features obtained by our method are more domain-indistinguishable concerning the limit, which indicates that the extracted matching features are not affected by domain changes, thus suggesting that our method is more invariant and generalizes better in representing entities from different domains.

6 Conclusion

We propose a knowledge retrieval method that is sensitive to matching attributes to improve general-

isation and interpretability by capturing matching information. Matching information can be used to filter mismatched parts of entities to improve entity selection. We first filter a batch of candidate entities by calculating the rough similarity between the context and the entity, and then calculate the match representation between the context and the entity and compute its match score with each attribute in the entity to obtain the relevance of the attribute to the context. Based on this relevance score, we filter out irrelevant attributes and recalculate the entities used to generate the response. We compare our system with other systems on three benchmark datasets and the results show that our retriever performs knowledge retrieval more effectively than existing methods.

Limitations

There are some potential limitations of the paper worth considering. Firstly, we have not explored the number of rough candidate entities for retrieval, but this may affect the final corresponding results. Second, the pseudo-tagging approach we utilized when extracting relevant information could be explored further. We will explore more efficient architectures for the response generator in future work.

Acknowledgement

We would like to thank all reviewers for their insightful comments. This paper was partially supported by NSFC (No: 62176008). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for its support.

Ethics Statement

All the experimental procedures in this study have been performed exclusively on publicly accessible datasets that do not contain any sensitive or private information. The datasets used in our research are publicly available and do not infringe upon individual privacy rights. Our work strictly adheres to ethical guidelines and does not involve the analysis or consideration of personal identity characteristics, including but not limited to gender and race. We emphasize that our research is focused solely on the scientific aspects of the problem at hand and does not engage in any form of discriminatory practices or bias based on gender, race, or any other protected attributes. The primary objective of our study is to contribute to the advancement of knowledge and technology in a fair, unbiased, and inclusive manner.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Dan Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537.

Zhenhao He, Yuhong He, Qingyao Wu, and Jian Chen. 2020a. Fg2seq: Effectively encoding knowledge for end-to-end task-oriented dialog. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8029–8033. IEEE.

Zhenhao He, Jiachun Wang, and Jian Chen. 2020b. Task-oriented dialog generation with enhanced entity representation. In *INTERSPEECH*, pages 3905–3909.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*.

Guanhuan Huang, Xiaojun Quan, and Qifan Wang. 2022. Autoregressive entity generation for end-to-end task-oriented dialog. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 323–332.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Yi Liu, Yuan Tian, Jianxun Lian, Xinlong Wang, Yanan Cao, Fang Fang, Wen Zhang, Haizhen Huang, Denvy Deng, and Qi Zhang. 2023. Towards better entity linking with multi-view enhanced distillation. *arXiv preprint arXiv:2305.17371*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. 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 133–142, Hong Kong, China. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Dinesh Raghu, Atishya Jain, Sachindra Joshi, et al. 2021. Constraint based knowledge base distillation in end-to-end task oriented dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5051–5061.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure aware task-oriented dialogue generation. *arXiv preprint arXiv:2204.09149*.

Weizhou Shen, Yingqi Gao, Canbin Huang, Fanqi Wan, Xiaojun Quan, and Wei Bi. 2023. Retrieval-generation alignment for end-to-end task-oriented dialogue system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8261–8275.

Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. Dual-feedback knowledge retrieval for task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6566–6580.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. 2019. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. 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 100–110.

Xin Tian, Yingzhan Lin, Mengfei Song, Siqi Bao, Fan Wang, Huang He, Shuqi Sun, and Hua Wu. 2022. Q-tod: A query-driven task-oriented dialogue system. *arXiv preprint arXiv:2210.07564*.

Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. 2023. Multi-grained knowledge retrieval for end-to-end task-oriented dialog. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11196–11210, Toronto, Canada. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašic, Nikola Mrkšic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019a. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019b. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Jie Wu, Ian G Harris, and Hongzhi Zhao. 2022. Graph-memdialog: Optimizing end-to-end task-oriented dialog systems using graph memory networks.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Berm: Training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. *arXiv preprint arXiv:2305.11052*.

Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 247–256.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.