From Retrieval to Generation: A Simple and Unified Generative Model for End-to-End Task-Oriented Dialogue

Zeyuan Ding, Zhihao Yang*, Ling Luo, Yuanyuan Sun, Hongfei Lin

School of Computer Science and Technology, Dalian University of Technology, China
zeyuanding@mail.dlut.edu.cn, {yangzh, lingluo, syuan, hflin}@dlut.edu.cn

Abstract

Retrieving appropriate records from the external knowledge base to generate informative responses is the core capability of end-to-end task-oriented dialogue systems (EToDs). Most of the existing methods additionally train the retrieval model or use the memory network to retrieve the knowledge base, which decouples the knowledge retrieval task from the response generation task, making it difficult to jointly optimize and failing to capture the internal relationship between the two tasks. In this paper, we propose a simple and unified generative model for task-oriented dialogue systems, which recasts the EToDs task as a single sequence generation task and uses maximum likelihood training to train the two tasks in a unified manner. To prevent the generation of non-existent records, we design the prefix trie to constrain the model generation, which ensures consistency between the generated records and the existing records in the knowledge base. Experimental results on three public benchmark datasets demonstrate that our method achieves robust performance on generating system responses and outperforms the baseline systems. To facilitate future research in this area, the code is available at https://github.com/dzy1011/Uni-ToD.

Introduction

Task-oriented Dialogue systems (ToDs) aim to assist users in accomplishing various tasks, such as hotel and restaurant reservations (Luong, Pham, and Manning 2015; Wang et al. 2020). Since the system response is guided not only by the dialogue history but also by the query knowledge base results, the ability to query the external knowledge base is essential in the EToDs (He et al. 2020a; Qin et al. 2021; He, Wang, and Chen 2020). Figure 1 illustrates such an example where the user asks for information about the *chinese restaurant*. By querying the knowledge base, the system provides the correct entities from the knowledge base to answer the user in natural language form.

Recent researches focus on various knowledge retrieval methods for task-oriented dialogue systems, which retrieve relevant records from the knowledge base for response generation. As shown in Figure 2(a), some studies (Wu, Socher, and Xiong 2019; Qin et al. 2020; Wang et al. 2020; Wu, Harris, and Zhao 2022; Qin et al. 2023) employ the memory network to encode the knowledge base and then generate the sketch response[1] based on the dialogue history. By retrieving the memory network, the sketch tags are replaced with the knowledge base entities. Since these methods need to retrieve the memory network frequently, it is difficult for joint optimization when dealing with a large-scale knowledge base. Unlike the previous studies, another approach (Rony, Usbeck, and Lehmann 2022; Xie et al. 2022) concatenates knowledge base and dialogue history as input to the language model for generating system responses, which is shown in Figure 2(b). These methods can avoid querying the knowledge base and capture the relation between the knowledge base and the dialogue. However, even large language models have input length limitations, in practice, there are thousands of records in the knowledge base, resulting in long input sequences that cannot be completely fed into the language model.

To mitigate this issue, as shown in Figure 2(c), Tian et al. (2022) uses the language model to generate query sentences and then employs an off-the-shelf retrieval model to retrieve the knowledge records based on the query sentences. Finally,

name	food	area	price	address	postcode
thanh binh	italian	west	cheap	13 histon roal	cb234p
chiquito	mexican	west	expensive	21 park road	cb17dy
golden wok	**chinese**	**south**	**moderate**	**19 histon road**	**b126le**
chiquito	italian	south	cheap	14 cherry road	c24lop
la raza	spanish	center	cheap	77 alger road	cb23ll

User : i am looking for a restaurant that serves chinese food in the south of town.
Systems: **golden wok** is a **moderate chinese** restaurant in the **south** of town.
User : what is the address and postcode ?
Systems: the address to **golden wok** is **19 histon road** and its postcode is **b126le**

Figure 1: A restaurant reservation example based on the CamRest dataset along with the knowledge base information. The blue word in the dialogue is the entity in the knowledge base.

*Corresponding author

[1] sketch response refers to employing entity types in sentences to replace specific entities. For example, the response: 'golden wok is a Chinese restaurant'. The sketch-response: '@name is a @food restaurant,' where '@name' and '@food' are sketch tags.

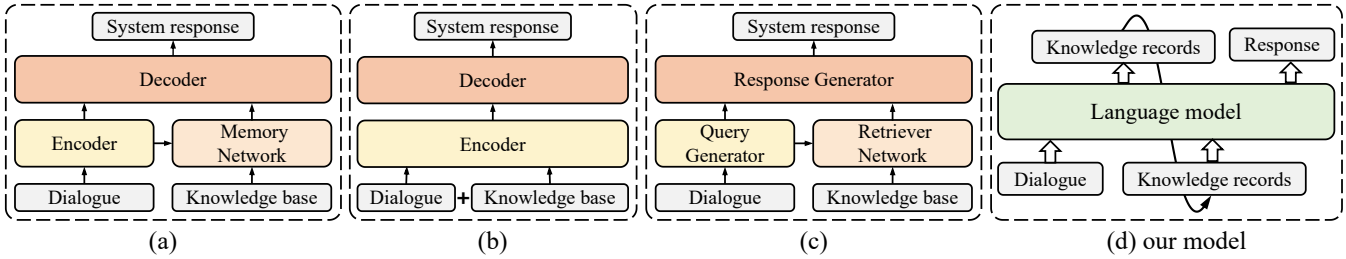

Figure 2: Methods for task-oriented dialogue. (a) This kind of method employs the memory network to encode the knowledge base. (b) This kind of method concatenates knowledge base and dialogue history as input. (c) This kind of method uses the language model to generate query sentences and then uses the retrieval model to retrieve relevant knowledge. (d) This kind of method is our proposed generative model, which only uses the dialogue history as input, then generates relevant knowledge records, and finally generates responses.

the responses are generated by the language model according to the retrieved records and dialogue history. Although this method proposes an effective strategy to retrieve a large-scale knowledge base, it still has two issues: (1) Since this method introduces additional query sentences, it is necessary to manually annotate the query sentences based on the dialogue history. (2) This method uses the pipeline where the three modules are separated from each other and cannot be jointly optimized, which makes it difficult to avoid error propagation and capture the relation between the two modules. A natural question arises can we use a unified model to handle retrieval and generation tasks in task-oriented dialogue?

In this paper, we propose a simple and unified model for task-oriented dialogue systems, which is shown in Figure 2(d), where we design task-oriented dialogue as a simple causal language modeling task. To transform the retrieval task into the generation task, we design the prefix trie to constrain the model generation, which ensures consistency between the generated records and the existing records in the knowledge base. Our model enables modeling of the inherent dependencies between the retrieval and the generation tasks of task-oriented dialogue, by optimizing for two tasks in a unified manner. To the best of our knowledge, we are the first to explore a unified approach in ToDs. Experimental results on three publicly available datasets demonstrate that our model achieves state-of-the-art performance.

Our contributions are summarized as follows:

- We propose a simple generative model for task-oriented dialogue systems, which recasts the EToDs task as a language modeling task using maximum likelihood training to solve the two tasks in a unified way.
- To prevent the generation of non-existent records from the knowledge base, we design the prefix trie to constrain the model generation, which can transform retrieval tasks into generation tasks.
- Experimental results on three public benchmark datasets show that our model outperforms the baseline models. Moreover, we provide extensive experiments to show the advantages of our model.

Related Work

Existing EToDs can be classified into three categories. The first category of research uses the memory network to encode the knowledge base and employs the sketch decoder to generate the sketch response. This method fills in the sketch tags with knowledge graph entities. For instance, Madotto, Wu, and Fung (2018) integrated end-to-end memory networks into response generation, while Wu, Socher, and Xiong (2019) proposed a global-to-locally pointer network for querying the knowledge database. Qin et al. (2020) introduced the dynamic fusion model that explicitly models domain knowledge for multi-domain dialogues, and He et al. (2020b) proposed an effective model to encode knowledge, which can model the structural information of the knowledge graph and the semantic from the dialogue history. Additionally, Raghu et al. (2021) utilized a pairwise similarity-based score function to improve the distillation of relevant knowledge-base records. Qin et al. (2023) fine-tune the pre-trained generation module and knowledge-retriever module to generate the response of the system.

The second category of research directly generates the final system response based on the dialogue history and the corresponding knowledge base. Madotto et al. (2020) introduced the method to encode the knowledge base directly into the model parameters, eliminating the need for a dialogue state tracker or template responses and enabling direct response generation. Rony, Usbeck, and Lehmann (2022) presented the novel EToDs that integrate knowledge entities into the language model effectively, with the model selectively incorporating relevant information during the dialogue generation process. Xie et al. (2022) proposed a UnifiedSKG method that employs a text-to-text framework that fuses the dialogue history and the knowledge base to generate system responses. The third category of studies uses the language model to generate query sentences and then uses the retrieval model to retrieve relevant knowledge based on the query sentences. Tian et al. (2022) proposed a query-driven task-oriented dialogue system, where they rewrite a natural language query for dialogue context, and use the query to retrieve the knowledge base.

Compared with previous work, the main differences of our model are as follows: (1) We propose a unified model

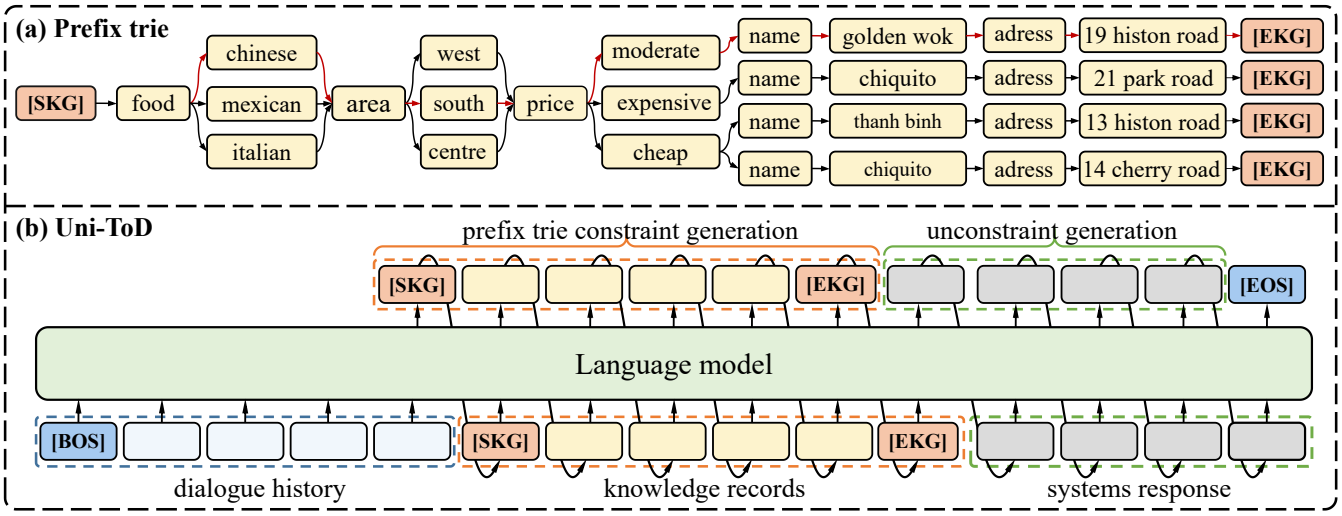

Figure 3: The end-to-end model framework. Figure (a) is a part of the prefix trie, which is constructed from the knowledge base. Where *'[SKG]'* represents the beginning of records in the knowledge base, and *'[EKG]'* represents the end of records. Figure (b) is the overall flow of the model. When the language model generates the special token *'[SKG]'*, the language model generates related records under the constraints of the prefix trie. After the language model generates the special token *'[EKG]'*, the language model starts to generate the system response without constraint.

that can generate records and the system response in a unified way. (2) We design a novel prefix trie, constructed from the knowledge base, that can transform the knowledge retrieval task into a knowledge generation task. (3) We use a unified loss to complete the optimization of the model, which can better capture the internal relationship between the knowledge base and the dialogue.

Method

In this section, we describe our model (Uni-ToD) for EToDs. The architecture of Uni-ToD is shown in Figure 3, which first builds the knowledge base into a prefix trie, and then uses a simple language model and a maximum likelihood loss to build a task-oriented dialogue system. Specifically, our model first generates related knowledge records under the constraints of the prefix trie, and then generates the system response without constraint.

Problem Definition

Unlike the previous methods, Uni-ToD aims to generate relevant knowledge records and informative responses based on the dialogue history. Therefore, we have to redefine the task. Given the dialogue between the user (U) and the system (S), n-turn dialogue utterances are represented as $(U_1, S_1), (U_2, S_2), ..., (U_n, S_n)$. The dialogues are associated with the knowledge database kb. At the i-th turn of the dialogue, our model takes the dialogue history $H_i = (U_1, S_1, ..., U_{i-1}, S_{i-1}, U_i)$ as input to generate the relevant knowledge base records kb_i and the system's response S_i. Specifically, the probability distribution of generating the system response using the language model is formally defined as follows,

$$p(S_i, kb_i|H_i) = \prod_{i=0}^{n} p(S_i|H_i, kb_i)p(kb_i|H_i) \quad (1)$$

where S_i is the system response in turn i and kb_i is the relevant knowledge records.

Language Model

A transformer based language model (Radford et al. 2019) is utilized to generate the response. The transformer layer in the language model consists of two blocks. The first block uses multi-head attention with h heads. Taking the (i+1)-th layer as an example, X_i is the input of the (i+1)-th layer. The attention mechanism is defined as follows,

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + M)V \quad (2)$$

$$X_i = layernorm(X_i) \quad (3)$$

$$h_k(X_i) = Attn(X_i W_k^Q, X_i W_k^K, X_i W_k^V) \quad (4)$$

where $Attn$ calculates the masked attention, h_k represents the k-th head. The W_k^Q, W_k^K and W_k^V are trainable parameters. The output of the first block is $T_i = [h_1, ..., h_k] + X_i$.

The second block uses a feedforward network with ReLU activation,

$$FF(T_i) = max(0, layernorm(T_i)U)V \quad (5)$$

where the output of the second block in the (i+1)-th layer is $X_{i+1} = FF(T_i) + T_i$.

Scores are computed from the output of the last layer X_l.

$$Scores = layernorm(X_l)W_{vocab} \quad (6)$$

During training, these scores are the inputs of a cross-entropy loss function.

Decoding Based on Prefix Trie

Inspired by De Cao et al. (2021), we design a prefix trie that can constrain the process of model generation. For task-oriented dialogue tasks, our model takes the dialogue history as input to generate the relevant records under the constraint of the prefix trie. As shown in Figure 3(a), we first build the knowledge base into a prefix trie. Specifically, we aggregate the same tokens in the same position from left to right. Each complete path on the prefix trie represents a record in the knowledge base. The record starts with *'[SKG]'* and ends with *'[EKG]'*. Since entities such as *'phone number'* and *'address'* are not directly related to dialogue history, it is difficult for the model to generate these entities only based on dialogue history. Therefore, in the process of constructing the prefix trie, we use attributes such as *'name'*, *'address'*, and *'phone number'* as a single path, forcing the model to generate completely correct attributes.

During constrained generation, as shown in Figure 4, if the first token is *'food'*, we set the model's probability predictions for all other tokens to zero, ensuring the model only generates *'food'*. When generating the second token, we set the probabilities of all tokens except *'m'*, *'ch'*, and *'ital'* to zero, forcing the model to choose from the three tokens. At this time, the generated result must be one of *'food m'*, *'food ch'*, and *'food ital'*. If the first and second tokens generated are *'food'* and *'ch'*, the third token can only be *'inese'*, because the next node of *'food ch'* only has one node *'inese'*, and the probabilities of tokens except *'inese'* are set to zero. By analogy, by zeroing out the candidate tokens that are not on the prefix trie, it is ensured that the decoding process only takes the branches of the prefix trie, and must be decoded to the *'[EKG]'* token.

Optimization

Our model first generates the *'[SKG]'* special token, and then uses the prefix trie to constrain the generation of records in the knowledge base. After generating the *'[EKG]'* special token, our model starts to generate the system response without constraints. The training objective of our model is the simple language modeling objective, which minimizes the likelihood of the next token from given tokens. The training loss is as follows,

$$\mathcal{L} = -logp(S_i, kb_i|H_i) \tag{7}$$

Where $\mathcal{L}$ is the loss function of our model, S_i is the system response in turn i, kb_i is the relevant knowledge records. During training, all parameters of our model are updated in back-propagation.

Experiments

Datasets

To assess the effectiveness of our model, we conduct experiments on three public benchmark datasets: **(1) CamRest dataset** (Wen et al. 2017) focuses on dialogues in the restaurant domain, consisting of 676 multi-turn dialogues with an average of 5 turns per dialogue. Each dialogue contains an average of 22.5 KB triples. Following previous work, we split this dataset into training/validation/test sets with 406/135/135 dialogues, respectively. **(2) In-Car Assistant dataset** (Eric et al. 2017) comprises 3,031 multi-turn dialogues spanning three distinct domains: weather, navigation, and schedule. On average, each dialogue consists of 2.6 turns and an average of 62.3 triples per dialogue. Following previous work, we split this dataset into training/-validation/test sets with 2425/302/304 dialogues, respectively. **(3) Multi-WOZ 2.1 dataset** contains three distinct domains: hotel, attraction and restaurant, with an average of 5.6 turns and 54.4 KB triples per dialogue. Following Qin et al. (2020), we split the dataset into training/validation/test sets with 1,839/117/141 dialogues, respectively.

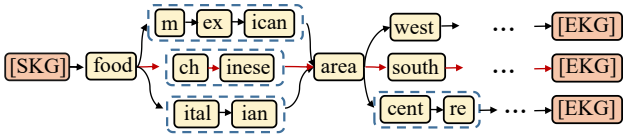

Figure 4: The decoding process for task-oriented dialogue. The dashed rectangles contain sub-words obtained through tokenization, and the red arrows represent the decoding path taken by the model.

Implementation Details

During training, we query the knowledge base for records by utilizing entities within the response and then employ the queried records and response as standard labels to train a model for generating relevant records. The batch size in our method is selected from a range of $[8, 16]$. Our method utilizes the GPT-2. We employ the AdamW optimizer with a learning rate of 6.25e-5 and a weight decay of 1e-8. The GELU activation is applied to our approach. All of our experiments are carried out using NVIDIA RTX 4090 GPUs with 24 GB of memory.

Automatic Evaluation

For evaluation metrics, we follow the baseline models and employ Entity F1 scores (Eric et al. 2017) and BLEU scores (Papineni et al. 2002) as automatic evaluation measures. BLEU computes the n-gram overlap between the generated responses and the gold responses. The Entity F1 scores assess the model's ability to generate responses grounded in knowledge by computing the F1 scores between the entities present in the ground truth and the system response.

Baselines

We compare our model with several representative works: (1) Mem2Seq (Madotto, Wu, and Fung 2018); (2) GLMP (Wu, Socher, and Xiong 2019); (3) DDMN (Wang et al. 2020); (4) GPT2+KE (Madotto et al. 2020); (5) DF-Net (Qin et al. 2020); (6) TTOS (He et al. 2020a); (7) FG2Seq (He et al. 2020b);(8) EER (He, Wang, and Chen 2020); (9) MCL (Qin et al. 2021); (10) CDNet (Raghu et al. 2021);(11)ECO (Huang, Quan, and Wang 2022);(12)Graph-MemDialog (Wu, Harris, and Zhao 2022) ; (13) DialoKG

Model	CamRest		In-Car Assistant		Multi-WOZ 2.1	
	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1
Mem2Seq (Madotto, Wu, and Fung 2018)	13.5	33.6	12.6	33.4	6.6	21.6
GLMP (Wu, Socher, and Xiong 2019)	13.9	59.6	13.9	59.6	6.9	32.4
DDMN (Wang et al. 2020)	19.3	58.9	17.7	55.6	12.4	31.4
GPT2+KE (Madotto et al. 2020)	18.0	54.9	17.4	59.8	**15.0**	39.6
TTOS (He et al. 2020a)	20.5	61.5	17.4	55.4	-	-
DF-Net (Qin et al. 2020)	-	-	14.4	62.7	9.4	35.1
EER (He, Wang, and Chen 2020)	19.2	65.7	17.2	59.0	13.6	35.6
FG2Seq (He et al. 2020b)	20.2	66.4	16.8	61.1	14.6	36.5
MCL (Qin et al. 2021)	20.1	59.2	17.2	60.9	13.6	32.6
CDNet (Raghu et al. 2021)	21.8	68.6	17.8	62.9	11.9	38.7
ECO (Huang, Quan, and Wang 2022)	18.4	71.6	-	-	12.6	40.9
GraphMemDialog (Wu, Harris, and Zhao 2022)	22.3	64.4	18.8	64.5	<u>14.9</u>	40.0
DialoKG (Rony, Usbeck, and Lehmann 2022)	<u>23.4</u>	<u>75.6</u>	<u>20.0</u>	<u>65.9</u>	12.6	<u>43.5</u>
MPEToDs (Qin et al. 2023)	19.3	58.9	17.7	55.6	13.6	36.3
Our model (Uni-ToD)	**24.7***	**77.8***	**22.0***	**66.6***	12.3*	**44.3***

Table 1: Comparison of our model with baselines on CamRest, In-Car and Multi-WOZ 2.1. The best result is highlighted in bold, and $*$ denotes a significant difference between our model and the DialoKG result according to a t-test at $p < 0.05$.

(Rony, Usbeck, and Lehmann 2022); (14) MPEToDs (Qin et al. 2023).

Main Results

Table 1 shows the experimental results of our model on three public datasets. The results clearly indicate that our model significantly surpasses all baseline models in terms of both BLEU score and entity F1, highlighting its effective contribution to dialogue response generation. Specifically, on the CamRest dataset, our method outperforms the previous DialoKG, by 1.3 points in BLEU score and nearly 2.2% in entity F1. The improvement in the BLEU score suggests a substantial reduction in generation errors by our model, while the gain in entity F1 demonstrates our model's superior accuracy in incorporating entities from external knowledge sources compared to the baselines. This proves that our model can effectively model the interaction between dialogue history and KB entities through unified modeling.

Furthermore, Our model achieves the highest entity F1 on Multi-WOZ 2.1 dataset and the highest BLEU and entity F1 score on In-Car dataset, indicating its superior generalization capability. Notably, our model surpasses DialoKG by 2 points in BLEU score and 0.7% in entity F1 on the In-Car dataset. On the Multi-WOZ 2.1 dataset, our method outperforms the previous DialoKG, by 5.4 points 0.9% in entity F1. The observed improvement in entity F1 highlights our model's remarkable reasoning ability in diverse dialogue history contexts, particularly considering the more complex KB information in the In-Car Assistant and Multi-WOZ 2.1 datasets. Despite the notable advancements made by DialoKG and MPEToDs, our model demonstrates a significant performance advantage over them.

Human Evaluation

We provide a human evaluation of our model and other baseline models. We randomly select 100 dialogues from the test set of three datasets for human evaluation. Following Qin et al. (2020), we invite three annotators to independently assign the score scale from 0 to 5 for each generated response. We report the average rating scores from all annotators. The agreement ratio computed with Fless' kappa (Landis and Koch 1977) is 0.73, showing moderate agreement. As shown in Table 2, we can see that our model outperforms baselines on all metrics, which is consistent with the automatic evaluation.

Model	Correct	Fluent	Humanlike
MPEEToDs	3.5	3.9	3.8
DialoKG	3.9	4.2	4.0
Our model	4.2	4.4	4.2

Table 2: Human evaluation results. The agreement ratio computed with Fless' kappa is 0.73.

Analysis

Ablation Study

In this section, we introduce multiple ablation experiments on CamRest and In-Car dataset using our model and present the results in Table 3. These results highlight the efficacy of various components within our model in achieving the ultimate performance.

W/o prefix trie. In this setting, we remove the prefix trie and only use GPT2 to generate the knowledge records and systems response directly. From the results, we can see that on the camrest and In-Car dataset, the BLUE score and the Ent. F1 drops. After removing the prefix trie, our model can easily generate records that do not exist in the knowledge base. This led to a decrease in model performance. In addition, when the model generates the attributes of food, the model can be generated correctly, but when it generates the attributes such as address and phone, the generation of the model is meaningless. This demonstrates that the prefix trie constraint model generates records that are consistent with

Model	CamRest		In-Car Assistant	
	BLEU	Ent.F1	BLEU	Ent.F1
Default	24.7	77.8	22.0	66.6
-w/o prefix trie	17.2	65.5	16.0	49.0
concatenate KB	12.7	51.6	16.5	59.2

Table 3: Ablation results with different settings on CamRest and In-Car test sets.

Model	CamRest		In-Car Assistant	
	BLEU	Ent.F1	BLEU	Ent.F1
DialoGPT	12.3	44.4	14.6	47.0
GPT+KB	12.7	51.6	16.5	59.2
GPT+KE	18.0	54.9	17.4	59.8
GPT+pre-train	22.0	63.9	18.8	63.8
Our model	24.7	77.8	22.0	66.6

Table 4: The impact of using different knowledge base methods on performance

those in the knowledge base. Directly transforming the retrieval task into a generation task, which introduces serious factual errors, will make the dialogue system unreliable. This phenomenon also shows that our model can accurately integrate the knowledge base, and proves that the model can convert the retrieval task into a generation task.

Concatenate KB. In this setting, we remove the prefix trie, directly concatenate the knowledge base into the dialogue history, and then send it to the input of the language model. From the results, we can see that on the In-Car dataset, the BLUE score drops by 5.5 points, and the Ent. F1 drops by 7 points. The method of direct concatenation suffers from performance degradation due to the large size of the knowledge base, making it difficult to fully input the dialogue history and knowledge base into the model. The method we proposed can directly generate the knowledge base that can directly generate related databases. This also demonstrates the effectiveness of our model.

The Impact of Using Different Knowledge Base Methods on Performance

In this section, for a fair comparison, we compare the models using GPT2 as the base model. DialoGPT (Zhang et al. 2020) directly generates system response based on the dialogue history; GPT+KB first linearizes a knowledge base into a sequence. Then it concatenates the linearized KB and the dialogue history as input and directly models the task-oriented dialogue task as the language modeling task with DialoGPT. GPT+KE (Madotto et al. 2020) directly embeds the knowledge base into the model's parameters. GPT+pre-train pre-trained generation module and knowledge-retriever module and fine-tune the pre-trained generation module and knowledge-retriever module to generate the system's response. The results show that our model performs best when the base language model is GPT2. Compared with other methods, our model can generate accurate knowledge records, does not need to query the knowledge base frequently, and can maintain knowledge record consistency.

Model	BLEU	Ent.F1
EER (He, Wang, and Chen 2020)	20.6	57.6
FG2Seq (He et al. 2020b)	19.2	59.4
CDNet (Raghu et al. 2021)	16.5	63.6
Q-ToD (Tian et al. 2022)	21.4	63.9
Our model (Uni-ToD)	21.2	72.3

Table 5: Results on the large-scale knowledge base

This result demonstrates the strength of our model.

The Impact of Large Scale Knowledge Base

Previous methods are trained and evaluated on the normal-sized knowledge base, but in real-world scenarios, we need to retrieve knowledge from the large-scale knowledge base. Therefore, we aggregate all knowledge bases in the CamRest to simulate real-world scenarios and use this dataset to evaluate the performance of the EToDs. We observe that our model outperforms baselines when using the large-scale knowledge base. Comparing the results in Table 1 and Table 5, we observe that the performance of existing systems deteriorates severely when using the large-scale knowledge base. The reason is that our model transforms the retrieval task into the generation task, which fully leverages the language model and enables the unified optimization of both tasks within a single model. Furthermore, our model constructs the kb into the prefix trie, which makes the model generate relevant knowledge records rather than taking all knowledge records into the model for records retrieval. This method can prevent the noise from the kb in the generation process. This also verifies the superiority of our model in dealing with large-scale knowledge base scenarios and the feasibility of applying it to practical scenarios.

The Performance of Precise Knowledge

In this section, we aim to explore the impact of accuracy in generating records on model performance. In Table 7, *"oracle record"* means that we directly use related records, and then generate system responses, and *"generate record"* means that our model generates records in the knowledge base, and then generates system responses. *"generation (F1)"* represents the F1 value of the record generated by our model. In order to explore the performance limitation of knowledge generation on the model, we directly linearize the records into sequence and concatenate them with the dialogue history, and then input the language model. As shown in Table 7, compared with the generated records, oracle records improve the BLUE score by 0.6 points and the Ent. F1 score by 3.8 points. We can see that the performance of the model has improved significantly. Generating inaccurate records can limit the performance of our model, which means that if we can generate more accurate knowledge base records, this can further improve our model performance.

The Impact of Prefix Trie Ordering

To explore the impact of different prefix trie orderings on model performance, we maintain the same experimental settings, excluding the prefix trie ordering as a vari-

	Point of interest (Poi)	Distance	Poi type	Traffic info	Address
Knowledge base	ravenswood shopping center	4 miles	shopping center	heavy traffic	434 arastradero rd
	civic center garage	**4 miles**	**parking garage**	**no traffic**	**270 altaire walk**
	jills house	4 miles	friends house	heavy traffic	347 alta mesa ave
	sigona farmers market	4 miles	grocery store	no traffic	638 amherst st
	trader joes	5 miles	grocery store	no traffic	408 university ave
	mandarin roots	4 miles	chinese restaurant	moderate traffic	271 springer street
	chevron	3 miles	gas station	moderate traffic	783 arcadia pl
User Query	what are the directions to the closest **parking garage**				
Ground truth	response: the closest **parking garage** is **civic center garage**, located **4 miles** away at **270 altaire walk**				
Baseline	response: the closest **parking garage** is **4 miles** away and it s name is stanford oval parking				
Our model	generates record: civic center garage, 4 miles parking garage, no traffic, 270 altaire walk				
	response: the closest **parking garage** is **civic center garage** at **270 altaire walk 4 miles** away would you like directions there?				

Table 6: Responses generated by our model and baseline model on In-Car Assistant dataset. The gold entities in each response are highlighted in bold. The incorrect entities in each response are highlighted in the underline.

oracle/generate	CamRest		In-Car Assistant	
	BLEU	Ent.F1	BLEU	Ent.F1
oracle record	25.0	84.0	22.1	84.3
generate record	24.7	77.8	22.0	66.6
generation (F1)	83.1		77.8	

Table 7: The performance of precise knowledge

prefix trie order	BLEU	Ent.F1
prefix trie order 1	24.7	77.8
prefix trie order 2	23.5	78.1
prefix trie order 3	24.4	77.6

Table 8: The impact of prefix trie ordering on our model

able. We establish three distinct orderings, labeled as *'food, area, pricerange, name, address, phone, postcode'*, *'area, food, pricerange, name, address, phone, postcode'* and *'pricerange, area, food, name, address, phone, postcode'*. As shown in Table 8, we observe that the performance of our model remains relatively stable across these different trie orderings, indicating the strong robustness of our approach. The performance of this model seems to be independent of the order of the prefix trie.

Case Study

In this section, we provide an example of responses generated by our model and the baseline model, which is shown in Table 6. Through the dialogue example, we can observe that our model successfully generates the correct entity(e.g., *civic center garage*, *270 altaire walk*, and *4 miles*). The type entity and distance entity generated by the baseline is generated correctly, but the poi entity (*civic center garage*) is generated incorrectly. Our model first generates records and then generates system responses. Since the existence of the prefix trie, the generated record must be complete, and the entity attributes must match. This further demonstrates the effectiveness and reliability of our model.

Discussions

The previous EToDs mainly focus on the retrieval of the knowledge base. They additionally train retrieval models or use memory networks and pointers to retrieve databases, which makes the generation and retrieval modules separate from each other and cannot be jointly optimized. Our model converts the retrieval task into the generation task through the prefix trie, and completes the unification of tasks. It can build EToDs using only one language model.

The accuracy score of the knowledge base generated by the current model can be further improved (In the Incar dataset, the accuracy score of the generated record is only 77.8%), and we can also design better generation algorithms to make the generated knowledge records more accurate, and there is still room for further improvement in this accuracy score. Furthermore, the method can also provide a new idea for the large language model fusion knowledge base, which transforms the retrieval task into the generation task to fuse the knowledge base. In future work, we will explore how to fuse knowledge bases into a larger language model.

Conclusion

In this paper, we propose a simple and unified model for EToDs, which recasts the ToD task as a simple language modeling task and uses maximum likelihood training to train retrieval and generation modules in a unified manner. To ensure consistency between the generated and the existing records in the knowledge base, we design the prefix trie to constrain the model generation. To the best of our knowledge, we are the first to explore a unified approach in ToDs. Experimental results on three public datasets demonstrate that our model achieves state-of-the-art performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62276043) and the Fundamental Research Funds for the Central Universities (No.DUT22ZD205).

References

De Cao, N.; Izacard, G.; Riedel, S.; and Petroni, F. 2021. Autoregressive Entity Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 37–49. Saarbrücken, Germany: Association for Computational Linguistics.

He, W.; Yang, M.; Yan, R.; Li, C.; Shen, Y.; and Xu, R. 2020a. Amalgamating Knowledge from Two Teachers for Task-oriented Dialogue System with Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3498–3507. Online: Association for Computational Linguistics.

He, Z.; He, Y.; Wu, Q.; and Chen, J. 2020b. 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)*, 8029–8033.

He, Z.; Wang, J.; and Chen, J. 2020. Task-Oriented Dialog Generation with Enhanced Entity Representation. In *Proc. Interspeech 2020*, 3905–3909.

Huang, G.; Quan, X.; and Wang, Q. 2022. Autoregressive Entity Generation for End-to-End Task-Oriented Dialog. In *Proceedings of the 29th International Conference on Computational Linguistics*, 323–332. International Committee on Computational Linguistics.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics.

Madotto, A.; Cahyawijaya, S.; Winata, G. I.; Xu, Y.; Liu, Z.; Lin, Z.; and Fung, P. 2020. Learning Knowledge Bases with Parameters for Task-Oriented Dialogue Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2372–2394.

Madotto, A.; Wu, C.-S.; and Fung, P. 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)*, 1468–1478. Melbourne, Australia: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Association for Computational Linguistics.

Qin, B.; Yang, M.; Bing, L.; Jiang, Q.; Li, C.; and Xu, R. 2021. Exploring Auxiliary Reasoning Tasks for Task-oriented Dialog Systems with Meta Cooperative Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15): 13701–13708.

Qin, L.; Xu, X.; Che, W.; Zhang, Y.; and Liu, T. 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*, 6344–6354. Online: Association for Computational Linguistics.

Qin, L.; Xu, X.; Wang, L.; Zhang, Y.; and Che, W. 2023. Modularized Pre-Training for End-to-End Task-Oriented Dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1601–1610.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raghu, D.; Jain, A.; Mausam; and Joshi, S. 2021. Constraint based Knowledge Base Distillation in End-to-End Task Oriented Dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5051–5061.

Rony, M. R. A. H.; Usbeck, R.; and Lehmann, J. 2022. DialoKG: Knowledge-Structure Aware Task-Oriented Dialogue Generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2557–2571. Seattle, United States: Association for Computational Linguistics.

Tian, X.; Lin, Y.; Song, M.; Bao, S.; Wang, F.; He, H.; Sun, S.; and Wu, H. 2022. Q-TOD: A Query-driven Task-oriented Dialogue System. *arXiv preprint arXiv:2210.07564*.

Wang, J.; Liu, J.; Bi, W.; Liu, X.; He, K.; Xu, R.; and Yang, M. 2020. Dual Dynamic Memory Network for End-to-End Multi-turn Task-oriented Dialog Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4100–4110. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gašić, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 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*, 438–449.

Wu, C.-S.; Socher, R.; and Xiong, C. 2019. Global-to-local Memory Pointer Networks for Task-Oriented Dialogue. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wu, J.; Harris, I. G.; and Zhao, H. 2022. GraphMemDialog: Optimizing End-to-End Task-Oriented Dialog Systems Using Graph Memory Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 11504–11512.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C. S.; Zhong, M.; Yin, P.; and Wang, S. I. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. *arXiv e-prints*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 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*, 270–278.