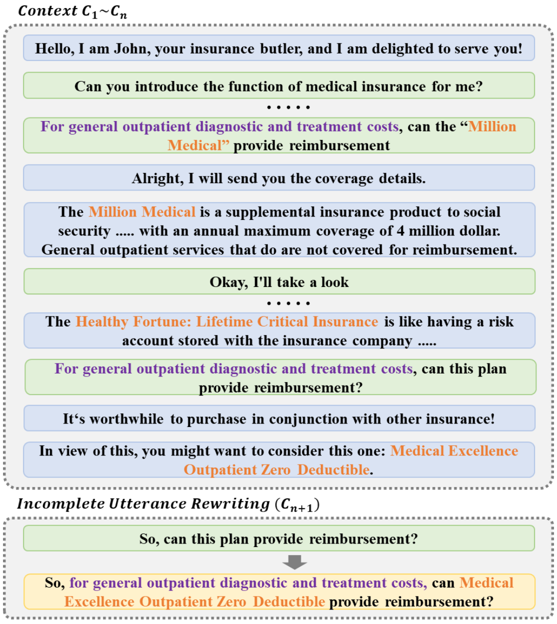
Context-Aware Tracking and Dynamic Introduction for Incomplete Utterance Rewriting in Extended Multi-Turn Dialogues

Xinnan Guo[1], **Qian Zhu**[1], **Qiuhui Shi**[1], **Xuan Lin**[1], **Liubin Wang**[1], **DaqianLi**[1], **Yongrui Chen**

[1]Ant Group, China
guoxinnan0727@163.com, zq371417, qiuhui.sqh, daxuan.lx, daqian.ldq@antgroup.com,
ubiwang@gmail.com, yrchen@seu.edu.cn

Abstract

Incomplete utterance rewriting (IUR) aims to reconstruct the utterance with omitted information and pronouns to be standalone and complete based on the context. The existing works predominantly focus on simple ellipsis and coreference problems in brief multi-turn dialogues. But in actual scenarios: 1) the context of the dialogues frequently comprises multiple similar candidates for ellipsis and coreference resolution, pouring to confuse. 2) the number of turns tends to be more extensive, while the content with various topics also grows more complex. This paper proposes a novel method called CaT to address these issues. In particular, we first devise a tacker model, distilled from GPT4-turbo, to adopt Context Tracking that dynamically updates a list of key phrases turn by turn, as accurate candidates for ellipsis and coreference resolution. Second, we further present the Dynamic Context Introduction mechanism to filter irrelevant preceding contexts that are not relied on by any element within the key phrase list to condense extended dialogues. Comprehensive experiments indicate that our solution provides a significant improvement over the existing baselines, and achieves state-of-the-art on three benchmarks[1].



Hello, I am John, your insurance butler, and I am delighted to serve you!

Can you introduce the function of medical insurance for me?

• • • • •

For general outpatient diagnostic and treatment costs, can the “Million Medical” provide reimbursement

Alright, I will send you the coverage details.

The Million Medical is a supplemental insurance product to social security ..... with an annual maximum coverage of 4 million dollar. General outpatient services that do are not covered for reimbursement.

Okay, I'll take a look

• • • • •

The Healthy Fortune: Lifetime Critical Insurance is like having a risk account stored with the insurance company .....

For general outpatient diagnostic and treatment costs, can this plan provide reimbursement?

It‘s worthwhile to purchase in conjunction with other insurance!

In view of this, you might want to consider this one: Medical Excellence Outpatient Zero Deductible.

So, can this plan provide reimbursement?

So, for general outpatient diagnostic and treatment costs, can Medical Excellence Outpatient Zero Deductible provide reimbursement?

Figure 1: An example of IUR in a product consulting and sales scenario. Within, green denotes the user’s input, while blue signifies the salesperson’s. The full dialogue is upon 35 turns.

1 Introduction

Incomplete Utterance Rewriting (IUR) serves as a vital component in multi-turn dialogue systems, which reconstruct the current utterance by integrating the omitted information and resolving coreference based on the context to ensure its completeness. Currently, IUR is widely employed in prevalent tasks such as information retrieving (Li et al., 2022b; Mo et al., 2023), question answering (Vakulenko et al., 2021), and web searching (Li et al., 2022a; Mohankumar et al., 2023). With the advent of large language models (LLMs) (Ouyang et al., 2022; OpenAI, 2023) and breakthroughs in multi-turn dialogue-oriented tasks, IUR has received increasing attention in recent years.

In contrast to the brief (in less than 10 turns on average) and simple (usually centering around a single topic) dialogue observed between users and systems, which is primarily focused on by current IUR researches as the benchmark (Elgohary et al., 2019; Su et al., 2019; Pan et al., 2019; Regan et al., 2019; Martin et al., 2020), customer service and sales scenarios often entail more complex dialogues between users and salespersons, as shown in Figure 1. These scenarios present two principal challenges: a) complex ellipsis and coreference problems delivered by the diverse dialogue content and topics. For instance, according to the demand for reimbursement, several insurance products are mentioned throughout the dialogues, which may lead to confusion in figuring out which product

[1]https://github.com/ygxw0909/CaT

Findings of the Association for Computational Linguistics: ACL 2024, pages 2138–2148
August 11-16, 2024

is the user inquiring about. b) extended turns, at times, reaching upwards of hundreds, encompassing multiple distinct topics. Within the dialogue in Figure 1, the topic transitions from consulting the function of medical insurance to recommending various products for purchase.

Some of the existing methods (Liu et al., 2020; Du et al., 2023; Li et al., 2023a) apply cross attention between the current utterance and its context to gain the rewriting matrix, using it for token-level actions (such as insertion, deletion, etc.). However, these methods are limited to copying tokens from the context, lack the flexibility for modifications, and also struggle with lengthy contexts. Meanwhile, many other methods (Hao et al., 2021; Xu et al., 2020; Jin et al., 2022) decompose the rewriting task into action prediction and span prediction, performing sequence tagging for action selection in each token of the current utterance and predicting candidate spans in the context. These methods do not adequately address the issue of ellipsis and coreference resolution among various similar candidates. (Inoue et al., 2022) jointly optimizes important tokens picking and generative utterances rewriting, but the design of the picker is too straightforward and still ignores the challenges.

To handle the challenges, this paper proposes a novel method called CAT (**C**ontext-**A**ware **T**racking and Dynamic Introduction), adopting Context Tracking (CT) and Dynamic Context Introduction (DCI) for resolve ellipsis and coreference, and requisite irrelevant context filtering: **1) Context Tracking.** To facilitate effective tracking and extracting the key content of complex dialogues that the current utterance depends on, we leverage a trainable tracker and propose Context Tracking to perpetually maintain a key-phrase list (kp list) for ellipsis and coreference resolution. Specifically, an empty kp list is initialized at first, as the dialogue advances, the tracker dynamically updates the kp list turn by turn. The updates include: a) add when new key phrases occur. b) delete when the topic shifts and key phrases are not dependent anymore. c) reserve when the previous key phrases may still be mentioned. **2) Dynamic Context Introduction.** In order to further filter out the turns which irrelevant to the current utterance, while ensuring the contextual integrity required for rewriting, the Dynamic Context Introduction mechanism is applied. Upon the update of the key phrases list, each key phrase is tagged with the index of the most recent turn it is mentioned. Based on that, the minimal dependency context is obtained by filtering out the part preceding the furthest turn tagged in the key phrases list. Ultimately, the rewriter takes the current utterance, kp list, as well as the minimal dependency context as input to generate the final complete utterance. To fully evaluate our proposed CAT, we collect the real dialogues between users and salespersons from our platform with manual annotation and construct a new challenging benchmark INSQR. Experimental results across three datasets prove the effectiveness of our method.

Our contributions are summarized as follows:

- We propose a novel Incomplete Utterance Rewriting method, which equips a tracker to maintain a key-phrases list for ellipsis and coreference resolution, and a rewriter to generate the complete utterance.
- We design the Dynamic Context Introduction mechanism to filter out the irrelevant context in order to handle the extended multi-turn dialogue, which further enhances the performance of the Rewriter.
- Our method achieves state-of-the-art results on all three IUR benchmarks, obtaining significant improvements compared to the existing methods.

2 Preliminaries

Given a dialogue context $\mathcal{C} = \{C_1, C_2, ..., C_{|\mathcal{C}|}\}$ and a following incomplete utterance $U = \{u_1, u_2, ..., u_{|U|}\}$, where $C_i = \{c_1, c_2, ..., c_{|C_i|}\}$ is the utterance of the i-th turn, u_j and c_k are the tokens of the utterance. The goal of IUR is to learn a mapping function,

$$U^* = f(\mathcal{M}, U, \mathcal{C}), \tag{1}$$

where U^* denotes a complete and standalone utterance resolving ellipsis and coreference for U, $\mathcal{M}$ denotes the trainable model parameters. Note that the textual content newly introduced in U^* is not required to maintain strict congruence with the text appearing in $\mathcal{C}$, and modification is allowed to ensure semantic coherence.

3 Methodology

3.1 Overview

Figure 2 illustrates the overview of our proposed CAT, which wraps a generative rewriter $\mathcal{R}$ with a context tracker $\mathcal{T}$. The tracker $\mathcal{T}$ adopts CT by maintaining a key phrase list (kp list) $\mathcal{K}$ turn by turn.

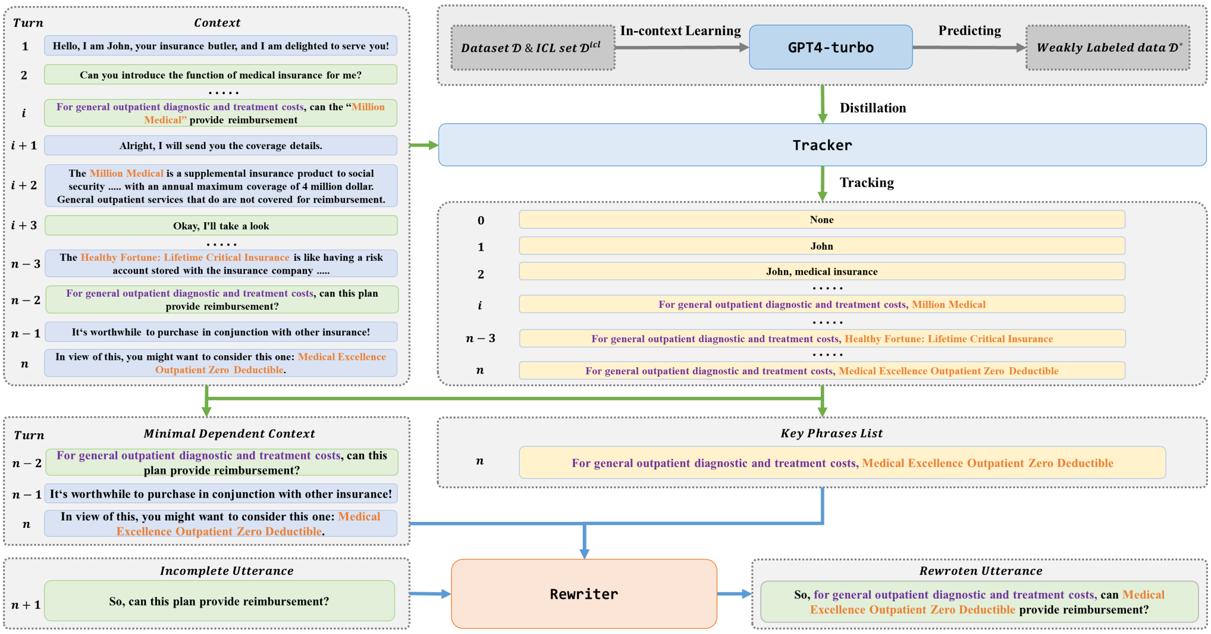

Turn
Context
1
Hello, I am John, your insurance butler, and I am delighted to serve you!
2
Can you introduce the function of medical insurance for me?
.....
i
For general outpatient diagnostic and treatment costs, can the "Million Medical" provide reimbursement
i + 1
Alright, I will send you the coverage details.
i + 2
The Million Medical is a supplemental insurance product to social security ..... with an annual maximum coverage of 4 million dollar. General outpatient services that do are not covered for reimbursement.
i + 3
Okay, I'll take a look
.....
n − 3
The Healthy Fortune: Lifetime Critical Insurance is like having a risk account stored with the insurance company .....
n − 2
For general outpatient diagnostic and treatment costs, can this plan provide reimbursement?
n − 1
It's worthwhile to purchase in conjunction with other insurance!
n
In view of this, you might want to consider this one: Medical Excellence Outpatient Zero Deductible.
Dataset 𝒟 & ICL set 𝒟^icl
In-context Learning
GPT4-turbo
Predicting
Weakly Labeled data 𝒟'
Distillation
Tracker
Tracking
0
None
1
John
2
John, medical insurance
.....
i
For general outpatient diagnostic and treatment costs, Million Medical
.....
n − 3
For general outpatient diagnostic and treatment costs, Healthy Fortune: Lifetime Critical Insurance
.....
n
For general outpatient diagnostic and treatment costs, Medical Excellence Outpatient Zero Deductible
Turn
Minimal Dependent Context
n − 2
For general outpatient diagnostic and treatment costs, can this plan provide reimbursement?
n − 1
It's worthwhile to purchase in conjunction with other insurance!
n
In view of this, you might want to consider this one: Medical Excellence Outpatient Zero Deductible.
Key Phrases List
n
For general outpatient diagnostic and treatment costs, Medical Excellence Outpatient Zero Deductible
Incomplete Utterance
n + 1
So, can this plan provide reimbursement?
Rewriter
Rewroten Utterance
So, for general outpatient diagnostic and treatment costs, can Medical Excellence Outpatient Zero Deductible provide reimbursement?

Figure 2: An overview of CAT. Initially, a tracker distilled from GPT4-turbo conducts CT to generate a key phrase list for each turn in the input context. Then, the DCI mechanism is utilized to obtain the minimal dependency context. Consequently, the rewriter takes over them combined with the incomplete utterance, and performs rewriting to output the complete utterance.

At the i-th turn, $\mathcal{K}_i = \{k_1^i, k_2^i, ..., k_{|\mathcal{K}_i|}\}$, where k_j^i is a key phrase mentioned in the previous context or the current utterance which can be referred to by the ongoing dialogue topic and content. Meanwhile, contrary to the existing methods that input the whole context $\mathcal{C}$ into the rewriter while ignoring the scenarios where $\mathcal{C}$ is particularly verbose, we proposed a DCI mechanism to filter the irrelevant context base on the latest kp list. At the i-th turn, the input minimal dependency context is denoted as $\mathcal{C}_i^{min} = \{C_m, C_{m+1}, ..., C_{i-1}\}, m \leq i-1$, where m is the furthest turn that the kp list has dependent. Eventually, the rewriter performs IUR based on the kp list of last turn $\mathcal{K}_{i-1}$, the minimal dependency context $\mathcal{C}_i^{min}$, combined with the current utterance U_i to obtain the rewritten complete utterance U_i^*. In addition, We leverage LLM and in-context learning to generate weakly supervised labels, thereby enabling the training of the tracker and gaining the capabilities of CT by LLM distillation.

3.2 Tracker for CT

The most common way (Liu et al., 2020; Hao et al., 2021; Jin et al., 2022; Inoue et al., 2022; Li et al., 2023a) to address ellipsis and coreference resolution is to directly predict the corresponding phrases from the context $\mathcal{C}$ for IUR, which is too straightforward to solve the challenge proposed in Section 1 Thus, we design CT that leverages a tracker $\mathcal{T}$ to dynamically maintain a kp list $\mathcal{K}$ turn by turn.

In particular, the kp list is first initialized as an empty list, denoted as $\mathcal{K}_0$. In each turn of the dialogues, $\mathcal{K}_i$ is predicted base on the last kp list $\mathcal{K}_{i-1}$, the current utterance U_i, and the limited previous context $\mathcal{C}_i^{\mathcal{T}}$, denoted as

$$\mathcal{K}_i = \mathcal{T}(\mathcal{K}_{i-1}, U_i, \mathcal{C}_i^{\mathcal{T}}), \tag{2}$$

where $\mathcal{C}_i^{\mathcal{T}} = \{C_j, C_{j+1}, ..., C_{i-1}\}$, $j = max(1, i - \gamma)$, γ is a hyper-parameter. The newly $\mathcal{K}_i$ is edited from $\mathcal{K}_{i-1}$ by these actions:

- **add**: insert a newly occurring key-phrase
- **reserve**: retain $k_j^{i-1} \in \mathcal{K}_{i-1}$, while k_j^{i-1} may still be referred to.
- **delete**: remove $k_j^{i-1} \in \mathcal{K}_{i-1}$, while k_j^{i-1} is not related to the current topic or content.

The update process persists across each turn. For instance, in Figure 2, "medical insurance" is added to $\mathcal{K}_2$ mentioned in C_2, due to its relevance to the dialogue topic. Beginning from the i-th, the customer consistently consults about reimbursed "general outpatient diagnostic and treatment costs", so it is reserved persistent in the kp list. Meanwhile, as the salesperson consistently recommends different insurance products, the user's focus shifts from one insurance product to another, such as from "Healthy Fortune: Lifetime Critical Insurance" to "Medical Excellence Outpatient Zero Deductible", and the old one that probably is not be mentioned in the after dialogue is removed while the new one is added accordingly. Ultimately, the updated kp list at the n-th turn serves as the input for rewriting the incomplete utterance in the $(n+1)$-th turn.

In light of the diversity of expressions for the same subject in the context, we utilize a generative way instead of span extraction. We employ T5 (Raffel et al., 2020) as the $\mathcal{T}$, to take over the task instruction combined with $(\mathcal{K}_{i-1}, U_i, \mathcal{C}_i^{\mathcal{T}})$ to create input prompts, and require the output of $\mathcal{K}_i$ in a predefined structured format

$$\begin{aligned}&\texttt{<bos>}x^{kp}_{1,1}, x^{kp}_{1,2}, ..., x^{kp}_{1,|k_1|}, \texttt{<sep>}, x^{kp}_{2,1},\\ &x^{kp}_{2,2}, ..., x^{kp}_{2,|k_2|}, \texttt{<sep>}, ..., x^{kp}_{|\mathcal{K}|,|k_{|\mathcal{K}|}|}, \texttt{<eos>},\end{aligned} \quad (3)$$

where <bos> and <eos> denotes the beginning and the end of the decoding, <sep> is the separate tokens and $x^{kp}_{i,j}$ is the j-th token of k_i.

3.3 Rewriter Equipped with DCI

Existing work (Inoue et al., 2022; Li et al., 2023b) has proved the efficiency of generative methods for IUR, hence, we also employ T5 as the backbone for the rewriter $\mathcal{R}$. To handle the extended dialogues, we design a DCI mechanism to dynamically truncate the irrelevant preceding turns, instead of feeding the entire context into the model.

In brief, the DCI mechanism iteratively records index $\mathcal{I} = \{I_1, I_2, ..., I_{|\mathcal{K}|}\}$ for kp list $\mathcal{K}$, which denotes the closest turn that each key phrase is mentioned. In i-th turn, the index $I^i_j \in \mathcal{I}_i$ of $k^i_j \in \mathcal{K}_i$ is obtained by the following way: 1) If k^i_j is mentioned in C_i and does not exist in $\mathcal{K}_{i-1}$, it means that k^i_j is a newly added key phrase, thus $I^i_j = i$. 2) If k^i_j is not mentioned in C_i but exists in $\mathcal{K}_{i-1}$, it signifies that k^i_j is a reserved key phrase, hence $I^i_j = I^{i-1}_n$, where $k^{i-1}_n \in \mathcal{K}_{i-1}$and $k^{i-1}_n == k^i_j$. 3) If k^i_j is mentioned in C_i and also exists in $\mathcal{K}_{i-1}$, this denotes that k^i_j need to update the recorded index, therefore $I^i_j = i$. Here, to judge whether k^i_j is mentioned in C_i, we perform sliding window matching on k^i_j and C_i with a window size $|k^i_j|$, using *Levenshtein distance* to compute similarity, and selecting the maximum score as the similarity score s^i_j. If $s^i_j \geq \delta$, then we considered that k^i_j is mentioned C_i, where δ is a hyper-parameter. After the iteratively recording, in $(i+1)$-th turn, the minimal dependency context is gained as $C_{min(\mathcal{I}_i):i}$ In Figure 2, $\mathcal{K}_n$ includes two key phrases, while the former is mentioned in the $(n-2)$-th turn and the latter appears in the n-th turn. So the minimal dependency context for rewriting is $C_{n-2:n}$.

The reason we employ fuzzy matching and index recording instead of directly requiring the tracker $\mathcal{T}$ to predict the action of $\mathcal{K}$'s updating is that integrating the prediction of the kp list and action would increase the decoding complexity for the rewriter. Furthermore, it would significantly diminish the accuracy of GPT4-turbo during zero-shot predictions, which is introduced in the next subsection.

During rewriting, we prompt the task instruction combined with the current incomplete utterance U_i, the minimal dependency context $\mathcal{C}^{min}_i$, and the kp list $\mathcal{K}_{i-1}$ predicted in last turn as input. While the rewriter $\mathcal{R}$ directly generate the rewritten complete utterance, as follow

$$U^*_i = \mathcal{R}(U_i, \mathcal{C}^{min}_i, \mathcal{K}_{i-1}), \quad (4)$$

where $\mathcal{K}_{i-1}$ provides precise candidates for ellipsis and coreference resolution, and $\mathcal{C}^{min}_i$ offers a minimal context for comprehending the topic and main content of the dialogue. Consequently, a more semantically accurate and complete U^*_i is attained.

3.4 Weakly Supervised Learning via LLM Distillation

LLMs recently distinguish themselves due to their potent generalization capabilities and robustness, showing great zero-shot performance on different tasks. Meanwhile, in-context learning (ICL), as a method within the prompt construction, is being widely adopted. It facilitates the model's task understanding by providing representative input-output examples indicative of the current task, included in the prompt. To solve the problem of lacking labeled data for training $\mathcal{T}$ since only golden rewriting utterance can be used to ensure fairness, we leverage GPT4-turbo combined with ICL to obtain weakly supervised labels.

Specifically, for each dataset $\mathcal{D}$, we select τ representative examples of dialogue context as the ICL set, denoted as $\mathcal{D}^{icl} = \{\mathcal{C}^{icl}_1, \mathcal{C}^{icl}_2, ..., \mathcal{C}^{icl}_\tau\}$. In order to make sure that the sampled contexts are emblematic and have substantial distinctions amongst each other, we choose them relying on many factors, such as the number of dialogue turns, the topics, the complexity of the content, and so on. We manually annotate these limited examples as $\hat{\mathcal{D}}^{icl} = \{(\mathcal{C}^{icl}_1, \hat{\mathrm{K}}_1), (\mathcal{C}^{icl}_2, \hat{\mathrm{K}}_2), ..., (\mathcal{C}^{icl}_\tau, \hat{\mathrm{K}}_\tau)\}$, where $\hat{\mathrm{K}}_i = \{\hat{\mathcal{K}}^i_1, \hat{\mathcal{K}}^i_2, ..., \hat{\mathcal{K}}^i_{|\mathcal{C}_i|}\}$ is a set of annotated kp list corresponding to each turn of the context $\mathcal{C}_i$, and $\hat{\mathcal{K}}^i_j$ denotes the labeled kp list in j-th turn of the context $\mathcal{C}_i$. We prompt the $\hat{\mathcal{D}}^{icl}$ incorporate with explicit task instruction and the $\mathcal{C}_i$ as input. The inference is formulated as

$$\mathrm{K}_i^* = GPT(\hat{\mathcal{D}}^{icl}, \mathcal{C}_i), \quad (5)$$

where K_i^* is the results parsed from the output. Subsequently, we filter some error returns during requesting and obviously wrong result after parsing, and gain the weakly labeled dataset $\mathcal{D}^* = \{(\mathcal{C}_1, \mathrm{K}_1^*), (\mathcal{C}_2, \mathrm{K}_2^*), ..., (\mathcal{C}_{|\mathcal{D}^*|}, \mathrm{K}^*_{|\mathcal{D}^*|}\}$ with a high quality. After that, we divide each pair of $(\mathcal{C}_i, \mathrm{K}_i^*) \in \mathcal{D}^*$ into training samples $(\mathcal{K}^i_{j-1}, U^i_j, \mathcal{C}^{\mathcal{T}}_{i,j}, \mathcal{K}^*_{i,j})$ for the tracker $\mathcal{T}$, as shown in formulation 2, where $\mathcal{K}^*_{i,j} \in \mathrm{K}_i^*$.

During the training process, the tracker $\mathcal{T}$ firstly performs weakly supervised learning based on the constructed dataset, to distill the capability of GPT4-turbo on the CT task. After that, the rewriter $\mathcal{R}$ is trained with the $\mathcal{K}$ of each IUR sample predicted by $\mathcal{T}$. Here, we do not employ joint learning, since the granularity of the training samples for the tracker $\mathcal{T}$ and rewriter $\mathcal{R}$ is not uniform, which is also due to the consideration of efficiency.

4 Experiments

Datasets

To comprehensively evaluate our method, we used the following three datasets covering both English and Chinese: 1) **INSQR** is a challenging Chinese IUR dataset constructed by the real dialogues between customers and sales collected from Alipay[2] platform, as well as the manual annotation to provide the gold utterance. Within, the longest dialogues can reach 100 turns, and the average turns is about 30, which is much longer than the existing datasets. The detailed comparison is shown in Table 1, and Table 2 shows the distribution of the number of turns. The dataset is divided into 4733/591/593 for Train/Val/Test. 2) **CANARD** (Elgohary et al., 2019) is a representative English IUR dataset derived by QuAC (Choi et al., 2018), an open-domain conversational question answering dataset about specific Wikipedia sections, while CANARD rewrites the originally incomplete questions in QuAC. It contains 31526/3430/5571 in the Train/Val/Test set for evaluation. 3) **CQR** (Regan et al., 2019) is an English IUR dataset extended from task-oriented dialogue (Eric et al., 2017) between drivers and an in-car assistant, which is divided into 2131/271/276 for Train/Val/Test. It provides multiple rewriting utterances while we only use the gold utterance for metric calculation.

Dataset	Train	Dev	Test	avg_c	avg_t
ReWriter (Su et al., 2019)	18000	2000	-	3.0	24.2
Restoration (Pan et al., 2019)	116360	3024	2999	4.9	34.4
MuDoCo (Martin et al., 2020)	5901	691	749	5.4	41.1
CQR (Regan et al., 2019)	2131	271	276	5.7	46.2
CANARD (Elgohary et al., 2019)	31526	3430	5571	9.8	92.9
INSQR	4733	591	593	**30.1**	**600.5**

Table 1: The comparison of INSQR with the existing benchmarks. Here, "avg_c" and "avg_t" denote the average of turns and length of context, separately. We use the three with the longest context for experiments.

Subset	**0-10**	**11-20**	**21-30**	**31-40**	**41-50**	**51+**
Train	723	774	1429	716	394	697
Dev	94	91	191	70	54	91
Test	90	95	189	78	50	91

Table 2: The distribution of the number of dialogue turns in INSQR.

Evaluation Metrics

Following the prior works, we use three different metrics for evaluation, including BLUE-scores (Papineni et al., 2002), ROUGE-scores (Lin, 2004), and Restoration-scores (Pan et al., 2019).

Methods for Comparison

We compared the proposed method with notable IUR methods, including Ptr-Gen (See et al., 2017), RUN (Liu et al., 2020), SRL (Xu et al., 2020), T5 (Raffel et al., 2020), RaST (Hao et al., 2021), HCT (Jin et al., 2022), QUEEN (Si et al., 2022), JET (Inoue et al., 2022), Flan-T5 (Chung et al., 2022), DuReSE (Jiang et al., 2023), MGII (Du et al., 2023), MIUR (Li et al., 2023a).

Implementation Details

Our method ran on Tesla A100 GPUs. Both English and Chinese versions of T5-base are used as the backbone of CAT for evaluation. The hyperparameters in the experiments were set as follows: (1) The batch size and the learning rate were set to 16 and 2e-5 (2) The beam search size of decoding was set to 8 (3) The number of the limited turns of the context γ for the tracker is set to 10 (4) The threshold δ for similarity score was set to 0.8 (5) The number of the examples τ for ICL was set to 5 (6) The maximum length of input and output were set to 512. All the experimental results were repeated 3 times and averaged.[3]

[2] https://www.alipay.com/

[3] The prompt of distillation is shown in the Github page.

	INSQR							CQR						
methods	*B1*	*B2*	*B4*	*R1*	*R2*	*F1*	*F2*	*B1*	*B2*	*B4*	*R1*	*R2*	*F1*	*F2*
RUN	72.6	68.8	61.1	81.4	71.2	45.4	34.3	77.4	68.9	53.5	88.0	74.2	-	-
HCT	72.8	68.4	61.0	80.9	70.2	-	-	79.4	74.8	59.3	88.2	73.2	-	-
MIUR	73.3	68.9	61.1	80.9	69.2	44.8	30.6	82.8	76.0	64.0	89.6	74.1	74.8	52.9
DuReSE	-	-	-	-	-	-	-	83.3	79.1	70.4	92.1	82.9	-	-
T5	73.4	69.8	61.1	82.2	70.3	46.0	36.9	80.8	76.1	67.4	88.1	79.1	73.8	65.0
Flan-T5	75.1	71.4	63.0	86.2	74.7	47.3	38.3	81.6	77.0	68.8	89.2	80.1	75.1	66.9
CAT	**78.8**	**74.7**	**66.5**	**89.8**	**79.9**	**53.0**	**43.6**	**85.4**	**83.9**	**80.5**	**93.7**	**87.7**	**77.9**	**69.5**
CAT+GPT4	77.2	72.1	65.0	87.9	76.2	51.1	40.8	84.3	82.4	80.3	91.1	84.2	77.2	68.7
CAT+Gold	89.6	85.7	77.9	97.1	88.6	86.8	73.1	93.4	91.6	87.5	94.8	89.9	84.0	76.9

Table 3: Overall results compared with existing baselines on INSQR and CQR. Here, "B1/2/4" denote Bleu-1/2/4, "R1/2" indicate Rouge-1/2, and "F1/2" represent the F1 of the Restoration score with 1/2-gram.

Method	*B1*	*B2*	*B4*	*R1*	*R2*	*RL*
PtrG	67.2	60.3	50.2	78.9	62.9	74.9
RUN	70.5	61.2	49.1	79.1	61.2	74.7
RaST	53.5	47.6	38.1	62.7	50.5	61.9
HCT	68.7	62.3	52.1	80.0	66.5	79.4
QUEEN	72.4	65.2	54.4	82.5	68.1	**81.8**
DuReSE	73.3	65.8	54.7	82.3	66.9	80.6
MGII	72.9	63.2	-	79.2	-	77.0
MIUR	71.3	63.4	51.7	81.6	64.5	77.4
JET	78.8	72.0	-	84.3	71.1	-
T5	77.1	70.7	60.1	81.9	69.8	79.6
Flan-T5	77.9	71.4	60.9	82.5	70.5	80.6
CAT	**79.7**	**72.7**	**62.6**	**84.9**	**71.7**	81.6
CAT+GPT4	79.4	**72.7**	62.2	84.5	71.5	81.6

Table 4: Overall results compared with existing baselines on CANARD. Here, "RL" denotes Rouge-L.

4.1 Overall Results

Initially, we compare our proposed CAT with the existing methods. To ensure fairness, the pre-trained model (BERT and T5) equipped by all the methods is set to the base version. Here, **CAT** denotes the method proposed in this paper including the tracker $\mathcal{T}$ and the rewriter $\mathcal{R}$, **CAT+GPT4** signifies the direct use of results $\mathcal{K}$ returned by GPT4-turbo, and **CAT+Gold** indicates the employment of golden phrases provided by the dataset, required in ellipsis and coreference resolution during rewriting. Table 3 and Table 4 show the experimental results, wherein our CAT outperforms the other methods and achieves state-of-the-art on all three datasets.

Results on INSQR and CQR

In INSQR and CQR, the main challenges of the rewriting are ellipsis and coreference resolution, while minimally revising the rest of the original sentence. Therefore, the metrics such as BLEU and ROUGE remain at a high level. However, the Restoration score reveals the difference in the accuracy of the actual modified part, where INSQR is obviously much lower than CQR, indicating that INSQR presents more complex scenarios. Experimental results demonstrate that, in comparison to the backbone T5 model, CAT achieves a 7% and 4.1% promotion in Restoration score, respectively, which prove the accurate $\mathcal{K}$ providing and irrelevant context filtering, can effectively improve the performance of IUR. Meanwhile, the comparison between **CAT** and **CAT+GPT4** yields an interesting result that, the tracker $\mathcal{T}$ surprisingly achieves a greater enhancement than GPT4-turbo in assisting the rewriter $\mathcal{R}$. we consider the reason is that the output from GPT4-turbo including its inherent randomness and error rate on both request and return, may introduce noise and bias when it is directly used as $\mathcal{K}$ which also serves as the foundation for DCI. In contrast, after training, $\mathcal{T}$ predicts $\mathcal{K}$ following its stable task understanding, hence achieving higher quality. **CAT+Gold** delineates the upper bound of dataset metrics. It is observable that, although the current CAT attains a notable improvement in INSQR and CQR, there remains a substantial gap to the upper bound, especially the Restoration score. It again highlights the complexity inherent in INSQR since the gap is huge.

Results on CANARD

CANARD presents a distinct scenario compared to the above two datasets. Beyond ellipsis and coreference resolution, it also entails modifications to the key phrases themselves, and sometimes also to the rest part of the original sentences. These modifications lack a definitive uniqueness, making

	Method	B1	B2	R1	R2
INSQR	**CAT**	**78.8**	**74.7**	**89.8**	**79.9**
	w/o CT	75.2	71.4	84.8	73.5
	w/o DCI	76.8	72.7	86.4	76.0
	T5	73.4	69.8	82.2	70.3
CQR	**CAT**	**85.4**	**83.9**	**93.7**	87.7
	w/o CT	81.6	77.4	89.7	81.2
	w/o DCI	**85.4**	83.5	92.7	**87.9**
	T5	80.8	76.1	88.1	79.1
CANARD	**CAT**	**79.7**	**72.7**	**84.9**	**71.7**
	w/o CT	77.9	71.5	82.9	70.6
	w/o DCI	78.9	72.2	84.2	71.2
	T5	77.1	70.7	81.9	69.8

Table 5: Overall ablation tests on three datasets.

it challenging for the Restoration score to evaluate the performance. The results shown in Table 4 present that the generative methods outperform the editing-based methods due to their flexibility in conforming to the various modifications present in the labels. CAT further delivers an improvement over existing methods. Within, JET also employs a picker for directly acquiring referential candidate phrases from the entire preceding context. In contrast, our proposed CT decomposes the complexity of prediction into each turn, maintaining a more precise $\mathcal{K}$, combined with a brief context $\mathcal{C}^{min}$ provided by DCI, thereby achieving superior performance on IUR. During the distillation, GPT4-turbo exhibits a more stable predictive performance on CANARD because all the dialogues are about Wikipedia, so the results of **CAT** and **CAT+GPT4** are close. Here, CANARD does not provide annotations of key phrases, hence the absence of the **CAT+Gold** setting.

4.2 Ablation Tests

To evaluate the contributions of each component of CAT, the ablation test is constructed. Here, we consider the following settings

- *w/o CT*: Removing $\mathcal{T}$ and rewriting without $\mathcal{K}$, while $\mathcal{C}^{min}$ is obtained by directly adopt DCI on the prediction of GPT4-turbo.
- *w/o DCI*: Removing the DCI while inputting the entire $\mathcal{C}$ to R.

The results are shown in Table 5. On our challenge dataset INSQR, both CT and DCI demonstrate significant improvement, which shows the effectiveness of CAT to handle the scenario with confusing ellipsis and coreference resolution and extended context. In the case of CQR, which primarily focuses on ellipsis and coreference resolution with short context, CT plays a pivotal role, while the impact of DCI is less pronounced. Meanwhile, both CT and DCI also exhibit a certain improvement in CANARD while the gold rewriting labels with non-unique expressions constrains the performance of the components.

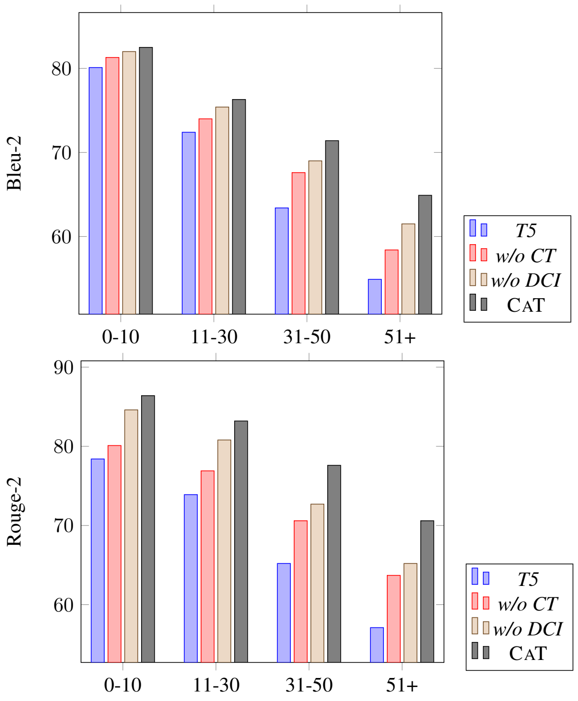


Figure 3: Detailed ablation tests of four subsets with different numbers of turns.

4.3 The Impact of Context Turns

For a detailed analysis of the efficacy of our proposed methods when dealing with the extended dialogues of varying turn and complexity, we partitioned the INSQR test set into four subsets based on the number of turns: 0-10, 11-30, 31-50, and 51+ (with the maximum being 100 turns). We then evaluate the score of Bleu-2 and Rouge-2 on these subsets, employing the same settings used in the ablation tests. The results are shown in Figure 3. With the increase in the number of turns, the improvement of CAT over the backbone T5 is progressively amplified, which proves the effectiveness of CAT in handling complex and extended dialogues. CT consistently delivers high contributions across all the subsets with various lengths, demonstrating that accurate $\mathcal{K}$ can steadily improve the performance of ellipsis and coreference resolution during rewriting. Meanwhile, when the context is brief with fewer turns, particularly within 10 turns, the

	Method	P	R	F1
INSQR	**GPT4-turbo**	50.7	**68.4**	58.2
	Tracker	**57.2**	64.4	**60.6**
CQR	**GPT4-turbo**	55.2	**76.1**	64.0
	Tracker	**61.6**	70.1	**65.6**

Table 6: The results of CT. Here, P, R, and F1 denote precision score, recall score, and F1 score.

enhancement brought by DCI is slight. However, after the increase in dialogues (31-50 and 50+), the demand for filtration of irrelevant context is amplified, thereby the contribution of DCI is highlighted.

.

4.4 Performance on CT

In addition to rewriting, we also conduct experiments to evaluate the results of CT. Herein, we compare the prediction results of $\mathcal{T}$ by the trained $\mathcal{T}$ with the zero-shot GPT-4-turbo on INSQR and CQR which both provide golden annotations for the key phrases leveraged by rewriting utterance. Note that we only evaluate the precision, recall, and F1 of the $\mathcal{T}$ on the rewriting turns due to the absence of annotations for the other turns. The results, shown in Table 6 indicate that GPT4-turbo retains the great capability to identify key phrases assisted by in-context learning, thus performing well in recall. However, it also tends to predict phrases that are extraneous to the core topic, resulting in a lower precision. In contrast to the zero-shot prediction, the trained tracker $\mathcal{T}$ demonstrates a deeper comprehension of the overall data distribution and the key phrases required to be extracted, thereby achieving a marked improvement in precision. While this comes with a little decrease in recall. Nonetheless, overall performance of $\mathcal{T}$ remains marginally superior to that of GPT-4-turbo. It also explains that, as shown in Table 3, employing the trained $\mathcal{T}$ achieves better performance than directly using the outputs of GPT-4-turbo.

5 Related Work

Early works on IUR primarily combine sequence-to-sequence based model with a copy mechanism to fetch the relevant information in the context (Su et al., 2019; Elgohary et al., 2019; Kumar and Joshi, 2017). Due to the strong representation capabilities of pre-trained models (Devlin et al., 2019), RUN (Liu et al., 2020) introduces a rewriting matrix derived from the embedding of context and utterance to adjudicate on a token-level whether insertion or replacement operations should be carried out. SRL (Xu et al., 2020) pre-identifies the candidate lexicon for the subject, predicate, and object within the context, and incorporates these as extra features into the encoder to achieve better performance. RaST (Hao et al., 2021) formulates the IUR as sequence labeling, effectuating the rewrite by predicting actions (insertion, deletion, and None) and the spans of the context separately. HCT (Jin et al., 2022) enhances the action predictor of the previous work to a rule predictor, optimizing for the generation of modified words not exactly present in the context. QUEEN (Si et al., 2022) proposed a query template that explicitly brings guided semantic structural knowledge between the incomplete utterance and the rewritten utterance. JET(Inoue et al., 2022) jointly optimizes important token picking and rewritten utterances generation (Raffel et al., 2020), but the design of the picker is too straightforward to perform accurate recall in extended dialogues. (Li et al., 2023b) employs continued pre-training on the T5 (Raffel et al., 2020) model to enhance its performance on IUR, while (Du et al., 2023) and (Li et al., 2023a) further devise fine-grained subtasks and refined model architectures based on the edit matrix. However, the existing benchmarks targeted by all the work referenced above are limited to simple ellipsis and coreference problems, while none of them pay attention to the challenges mentioned in section 1.

6 Conclusion

In this paper, we presented a novel method for incomplete utterance rewriting among extended dialogues with complex content, called CAT which is a component of a context tracker and a generative rewriter. Context Tracking is adopted by the tracker to maintain a list turn by turn, which includes the key phrase mentioned in the previous context that may still be referred to in the current utterance, as candidates to promote the performance of ellipsis and coreference resolution. Furthermore, the Dynamic Context Introduction is designed to filter the irrelevant previous turns and retain the minimal dependency context for the rewriter. Extensive experiments indicate that CAT achieves state-of-the-art results on three benchmarks compared with the existing baselines. In future work, we will expand our work on a wider range of scenarios, and also explore the optimization of the tracker.

7 Limitations

This paper proposed a novel Incomplete Utterance Rewriting method, namely CAT, which primarily focuses on the task in dialogue systems with extended multi-turn dialogue and complex content. Nevertheless, we have not combined IUR with some downstream tasks such as recommendation or question answering for further discussion. Meanwhile, though the design of Context Tracking can yield significant improvement, it necessitates real-time computation and prediction as the dialogue moves on, which results in substantial computational resources required in actual scenarios. Simultaneously, the efficacy of the tacker is heavily contingent upon the quality of weakly supervised labels obtained from the LLMs.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Haowei Du, Dinghao Zhang, Chen Li, Yang Li, and Dongyan Zhao. 2023. Multi-granularity information interaction framework for incomplete utterance rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2576–2581. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, François 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, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49. Association for Computational Linguistics.

Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. RAST: domain-robust dialogue rewriting as sequence tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4913–4924. Association for Computational Linguistics.

Shumpei Inoue, Tsungwei Liu, Son Nguyen, and Minh-Tien Nguyen. 2022. Enhance incomplete utterance restoration by joint learning token extraction and text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3149–3158. Association for Computational Linguistics.

Wenhui Jiang, Xiaodong Gu, Yuting Chen, and Beijun Shen. 2023. Durese: Rewriting incomplete utterances via neural sequence editing. *Neural Process. Lett.*, 55(7):8713–8730.

Lisa Jin, Linfeng Song, Lifeng Jin, Dong Yu, and Daniel Gildea. 2022. Hierarchical context tagging for utterance rewriting. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10849–10857. AAAI Press.

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval based sequence to sequence learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 705–714. ACM.

Jiang Li, Xiangdong Su, Xinlan Ma, and Guanglai Gao. 2023a. How well apply simple MLP to incomplete utterance rewriting? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1567–1576. Association for Computational Linguistics.

Sen Li, Fuyu Lv, Taiwei Jin, Guiyang Li, Yukun Zheng, Tao Zhuang, Qingwen Liu, Xiaoyi Zeng, James T. Kwok, and Qianli Ma. 2022a. Query rewriting in taobao search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 3262–3271. ACM.

Xiangsheng Li, Jiaxin Mao, Weizhi Ma, Zhijing Wu, Yiqun Liu, Min Zhang, Shaoping Ma, Zhaowei Wang, and Xiuqiang He. 2022b. A cooperative neural information retrieval pipeline with knowledge enhanced automatic query reformulation. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 553–561. ACM.

Zitong Li, Jiawei Li, Haifeng Tang, Kenny Q. Zhu, and Ruolan Yang. 2023b. Incomplete utterance rewriting by A two-phase locate-and-fill regime. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2731–2745. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2846–2857. Association for Computational Linguistics.

Scott Martin, Shivani Poddar, and Kartikeya Upasani. 2020. Mudoco: Corpus for multidomain coreference resolution and referring expression generation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 104–111. European Language Resources Association.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4998–5012. Association for Computational Linguistics.

Akash Kumar Mohankumar, Bhargav Dodla, Gururaj K, and Amit Singh. 2023. Unified generative & dense retrieval for query rewriting in sponsored search. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 4745–4751. ACM.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, and Xu Jiang. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1824–1833. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Michael Regan, Pushpendre Rastogi, Arpit Gupta, and Lambert Mathias. 2019. A dataset for resolving referring expressions in spoken dialogue via contextual query rewrites (CQR). *arXiv e-prints*, page arXiv:1903.11783.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Shuzheng Si, Shuang Zeng, and Baobao Chang. 2022. Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4839–4847. Association for Computational Linguistics.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 22–31. Association for Computational Linguistics.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 355–363. ACM.

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. 2020. Semantic role labeling guided multi-turn dialogue rewriter. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6632–6639. Association for Computational Linguistics.