Don't be Contradicted with Anything! CI-ToD: Towards Benchmarking Consistency for Task-oriented Dialogue System

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, Wanxiang Che*
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{lbqin,tianbaoxie,sjhuang,qgchen,xxu,car}@ir.hit.edu.cn

Abstract

Consistency Identification has obtained remarkable success on open-domain dialogue, which can be used for preventing inconsistent response generation. However, in contrast to the rapid development in open-domain dialogue, few efforts have been made to the task-oriented dialogue direction. In this paper, we argue that *consistency problem* is more urgent in task-oriented domain. To facilitate the research, we introduce CI-ToD, a novel dataset for **C**onsistency **I**dentification in **T**ask-**o**riented **D**ialog system. In addition, we not only annotate the single label to enable the model to judge whether the system response is contradictory, but also provide more fine-grained labels (i.e., Dialogue History Inconsistency, User Query Inconsistency and Knowledge Base Inconsistency) to encourage model to know what inconsistent sources lead to it. Empirical results show that state-of-the-art methods only achieve 51.3%, which is far behind the human performance of 93.2%, indicating that there is ample room for improving consistency identification ability. Finally, we conduct exhaustive experiments and qualitative analysis to comprehend key challenges and provide guidance for future directions. All datasets and models are publicly available at https://github.com/yizhen20133868/CI-ToD.

1 Introduction

Task-oriented dialogue systems (ToDs) (Young et al., 2013) aim to achieve user goals such as hotel booking and restaurant reservation, has gained more attention recently in both academia and industries. Over the last few years, two promising research directions in ToDs have emerged. The first focuses on a pipeline approach, which consists of modularly connected components (Wu et al., 2019a; Takanobu et al., 2020; Peng et al., 2020; Li et al., 2020). The second direction employs an end-to-end model, which directly takes the sequence-to-sequence (Seq2Seq) model to generate a response from a dialogue history and a corresponding knowledge base (KB) (Eric et al., 2017; Madotto et al., 2018; Wen et al., 2018; Qin et al., 2019b; Wu et al., 2019b; Qin et al., 2020b)

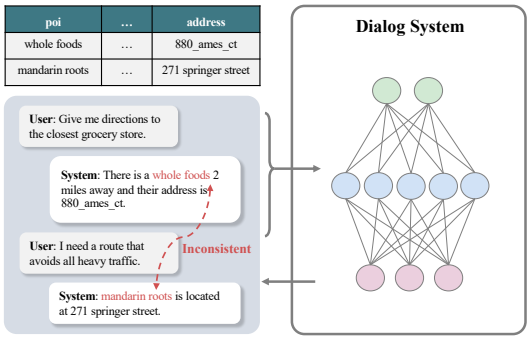

Figure 1: A system response generation example by the state-of-the-art end-to-end task-oriented dialogue model *DF-Net* (Qin et al., 2020b).

In recent years, with the burst of deep neural networks and the evolution of pre-trained language models, the research of ToDs has obtained great success. While the success is indisputable, previous work have shown that it's inevitable to generate inconsistent response with the neural-based model, resulting in a contradiction (Welleck et al., 2019; Song et al., 2020; Nie et al., 2021). Such contradictions caused by these bots are often jarring, immediately disrupt the conversational flow. To address the above issue, some work try to improve consistency in dialogue by posing a consistency identification into dialogue. Welleck et al. (2019) made an early step towards performing consistency identification in dialogue agent. Nie et al. (2021) proposed dialogue contradiction detection task to prevent the system response from being inconsistent with dialogue history. Song et al. (2020) further proposed a profile consistency identification to consider whether response is consistent with the corresponding profile. Though achieving

*Email corresponding.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2357–2367
November 7–11, 2021.

Dataset	Open Domain/ Task-Oriented Dialoge System	External Knowledge	Multi-turn / Single-turn	Single Label / Fine-grained Labels
Dialogue NLI (Welleck et al., 2019)	Open domain	✖	Single-Turn	Single Label
InferConvAI (Dziri et al., 2019)	Open domain	✖	Multi-Turn	Single Label
KvPI (Song et al., 2020)	Open domain	✔	Single-Turn	Single Label
DECODE (Nie et al., 2021)	Open domain	✖	Multi-Turn	Single Label
CI-ToD	Task-Oriented	✔	Multi-Turn	Fine-grained Labels (HI, QI and KBI)

Table 1: Comparison between our dataset and other datasets. HI denotes Dialog History Inconsistency; QI denotes User Query Inconsistency; KBI represents Knowledge Base Inconsistency.

the promising performance, the above work were limited to open-domain dialogue. In this paper, we highlight that *inconsistent generation problems should also be considered in task-oriented dialogue*. For example, as shown in Figure 1, the system expresses about the POI *whole foods* in dialogue history. However, when we run the state-of-the-art model (*DF-Net*) (Qin et al., 2020b), the system generate response *"mandarin roots is located at 271 springer street."*, which incorrectly generates irrelevant POI *mandarin roots*, resulting in contradiction. This is because neural-based models are a black-box and thus make us hard to explicitly control the neural-based dialogue systems to maintain a consistent response generation. From the user's perspective, such inconsistent bots not only fail to meet the requirements of the user but also mislead users to get wrong feedback in the task-oriented domain. Therefore, it's promising to consider *consistency* problem and detect in advance whether the generated response is consistent in task-oriented dialogue direction. Unfortunately, there still has been relatively little research on considering consistency identification in task-oriented dialogue due to the the lacking of public benchmarks.

To fill this research gap, we introduce a novel human-annotated dataset CI-ToD: **C**onsistency **I**dentification in **T**ask-**o**riented **D**ialog system. Dialogue data for CI-ToD is collected from the public dialogue corpora KVRET (Eric et al., 2017). For each final system response in KVRET, we re-write the utterance by crowdsourcing where we deliberately contradict the dialogue history, user query or the corresponding knowledge base (KB). As shown in Table 1, compared to the existing consistency identification for dialogue dataset, CI-ToD has the following characteristic: (1) *Task-oriented Dialogue Domain*. To the best of our knowledge, we are the first to consider dialog consistency in task-oriented dialogue system while the prior work mainly focuses on the open domain dialogue system. We hope CI-ToD can fill the gap of *consistency identification* in the task-oriented dialogue domain; (2) *Fine-grained Annotations.* We provide not only single annotations of whether each sentence is consistent, but also more fine-grained annotations, which can be used for helping the model analyze what source is causing this inconsistency.

To establish baseline performances on CI-ToD, we evaluate the state-of-the-art pre-trained and non pre-trained models for consistency identification. Experimental results demonstrate a significant gap between machine and human performance, indicating there is ample room for improving consistency identification ability. In addition, we show that our best consistency identification detector correlates well with human judgements, demonstrating that it can be suitable for use as an automatic metric for checking task-oriented dialogue consistency. Finally, we perform exhaustive experiments and qualitative analysis to shed light on the challenges that current approaches faced with CI-ToD.

In summary, our contributions are three-fold:

- We make the first attempt to consider consistency identification in task-oriented dialog and introduce a novel human-annotated dataset CI-ToD to facilitate the research.
- We establish various baselines for future work and show well-trained consistency identification model can be served as an automatic metric for checking dialogue consistency.
- We conduct exhaustive experiments and qualitative analysis to comprehend key challenges and provide guidance for future CI-ToD work.

2 Problem Formulation

In our paper, the consistency identification in task-oriented dialogue is formulated as a supervised multi-label classification task, which aims to judge whether the generated system response is inconsistent. To equip the model with the ability to analyze what the inconsistent sources lead to it, we require the model not only provide the final prediction but also the fine-grained sources including dialogue history, knowledge base (KB) and user's

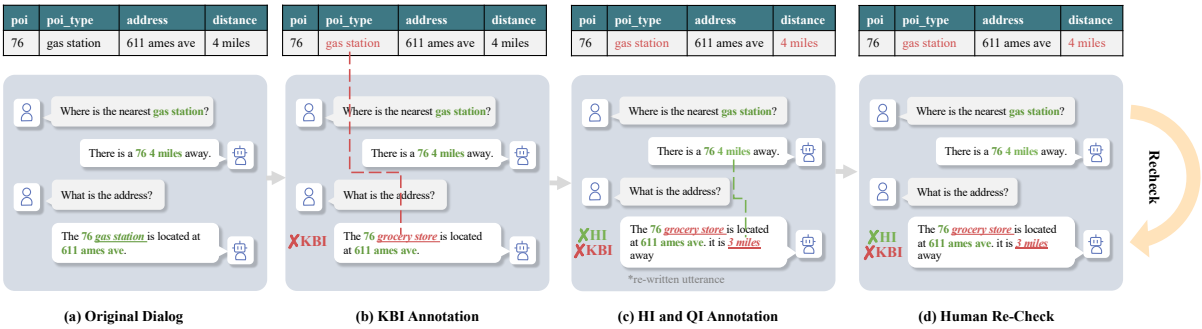

Figure 2: The process of CI-ToD construction.

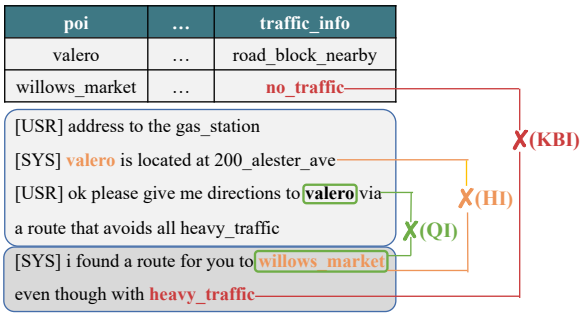

Figure 3: Inconsistent types in CI-ToD. Different colors denote different inconsistent types.

query. More specifically, given a task-oriented dialogue between a user (u) and a system (s), the n-turned dialogue snippet consists of dialogue history $H = \{(u_1, s_1), (u_2, s_2), ..., (u_{n-1}, s_{n-1})\}$, the corresponding knowledge base KB, the user query u_n and system response s_n. More specifically, the task can be defined as:

$$\mathbf{y} = f([H, KB, u_n], s_n), \quad (1)$$

where f denotes the trainable model; $\mathbf{y}$ is an output three-dimension vector, indicating whether the last utterance s_n contradicts any previously mentioned dialogue history H, user query u_n or the corresponding knowledge base KB .

3 Dataset

We construct the CI-ToD dataset based on the KVRET dataset and follow four steps: (a) *Data Pre-Processing*, (b) *KBI Construction*, (c) *QI and HI Construction* and (d) *Human Annotation*, which is illustrated in Figure 2. In the following, we first describe the definition of QI, HI and KBI, then illustrate the four construction steps in detail.

3.1 Inconsistency Types

As show in Figure 3, we give an example to show different inconsistency types, which are illustrated as follows:

User Query Inconsistency (QI) QI denotes that the dialogue system response is inconsistent with the current user query. Take the dialogue in Figure 3 for example, in the last turn of dialogue, user's query is asking about *valero*, while the final system response don't satisfied with user's requirement, showing a route to *willows_market*, which causes the user query inconsistency.

Dialogue History Inconsistency (HI) HI denotes that the dialogue system response is inconsistent with the dialogue history except the current user query. Take the dialogue in Figure 3 for example, the previous system response is talking about *valero* and the user do not change the theme of the dialogue. However, the final system response turn to discussing about *willows_market* , causing the dialogue history inconsistency.

Knowledge Base Inconsistency (KBI) KBI denotes that the dialogue system response is inconsistent with the corresponding KB, which is an unique challenge in task-oriented dialogue domain. Take the dialogue in Figure 3 for example, the final system response express the traffic_info of *willows_market* is *heavy_traffic*, which is conflict with the corresponding KB (*no_traffic for willows_market*).

3.2 Data Collection and Statistics

3.2.1 Step 1 Data Pre-Processing

We build CI-ToD on existing dialogues KVRET rather than collecting new dialogue from scratch More specifically, given a n-turned dialogue $\{(u_1, s_1), (u_2, s_2), ..., (u_n, s_n), KB\}$ for KVRET,

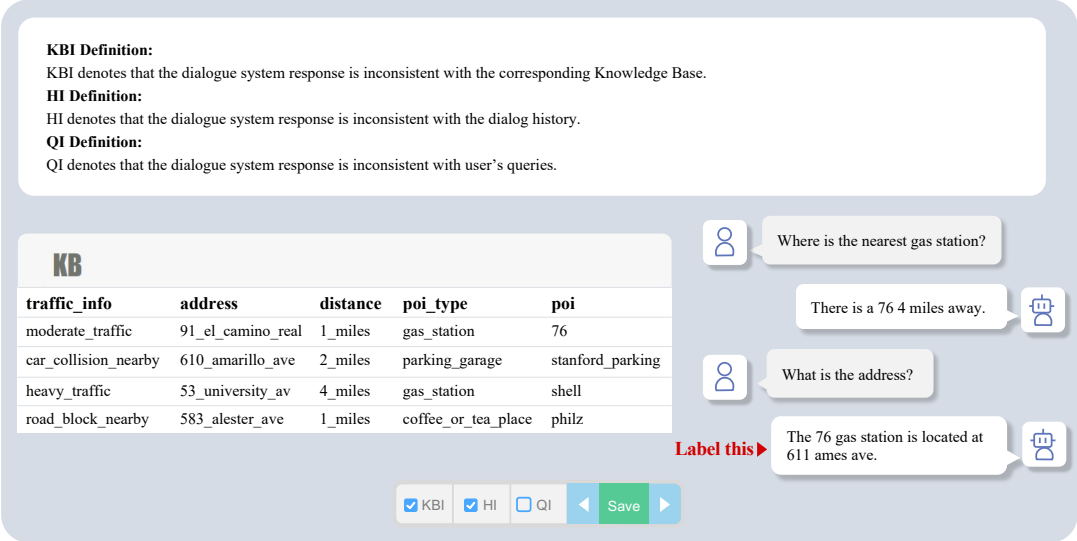

Figure 4: The collection interface.

	CI-ToD
# Domain	3
# Training Dialogues	2,553
# Validation Dialogues	319
# Test Dialogues	318
# Avg. Utterances per Dialogue	3.693
# Inconsistency ratio	0.648
# KBI ratio	0.521
# QI ratio	0.485
# HI ratio	0.214

Table 2: Data statistics of CI-ToD.

we first split it into some sub-dialogues to generate various samples, such as $\{(u_1, s_1), KB\}$, ..., $\{(u_1, s_1), (u_2, s_2), ..., (u_{n-1}, s_{n-1}), KB\}$ and $\{(u_1, s_1), (u_2, s_2), ..., (u_n, s_n), KB\}$. In addition, to ensure the system response is informative, we filter these general response, such as *"Thanks"* and *"You are welcome"*. Finally, we obtain the pre-processed dialogues.

3.2.2 Step 2 KBI Annotation

Given the pre-processed dialogues, we first construct KBI for each dialogue. KBI denotes that the final system response is inconsistent with the corresponding KB. We simply replace the knowledge entity value to construct KBI automatically.

More specifically, for each knowledge value in the system response, we sample specific entities from the whole KB to replace the selected slot and ensure that the sampled KB entity is different with the selected value. By this means, the constructed response is inconsistent with the corresponding KB. For example, as shown in Figure 2(b), we replace the entity *"gas station"* with *"grocery store"*, which resulting in KBI (the corresponding KB is (poi_type for *gas station*)).

3.2.3 Step 3 QI and HI Annotation

In this section, we show how we generate QI and HI. Since this require us to have a deep understanding for the corresponding user's query and dialogue history, constructing a system response with QI or HI is non-trivial, To address this issue, we achieve this by human efforts. We hire a human annotation team[1] to (1) randomly assign a sample with QI or HI and re-write each response to make it inconsistent with user query or dialogue history, and (2) check whether each written response is fluent or not by three extra annotators.

3.2.4 Step 4 Human Re-Check

In the final step, we will re-check the fine-grained inconsistent information with human efforts, including QI, HI and KBI. To ensure quality, each sample is annotated by three people, and the annotation process lasts nearly three months. Figure 4 shows the annotation user interface.

The detailed statistics of CI-ToD is summarized in Table 2. The percentage of inconsistency has exceeded 50%, indicating that CI-ToD is challenging.

3.2.5 Quality Control

To control the quality of the annotated dataset, we introduce different verification methods:

[1]All annotators are undergraduates from university in China, who are familiar with English language. (pass the College English Test (CET-6), one of the hardest English level exams in China.)

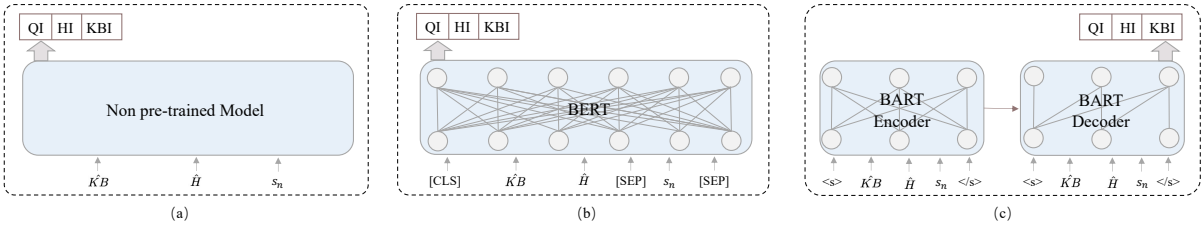

Figure 5: The model structure of non pre-trained model (a) and pre-trained model (b and c).

(1) **Onboarding Test:** Each annotator will have an advance annotation test, where each annotator will first annotate 100 samples and 3 experts check their annotation results. Finally, only those who achieves 80% annotation performance can conduct the following annotation work; (2) **Double Check** We randomly sampled 1,000 samples from the final annotated dataset and ask two new annotators to annotate the inconsistent information. Following (Bowman et al., 2015), we calculated the Fleiss' Kappa among the previous labels and two new labels and obtained a kappa of 0.812, which means almost perfect agreement (Landis and Koch, 1977).

4 Models

In this section, we establish several strong baseline methods using the state-of-the-art non pre-trained models (§4.1) and pre-trained models (§4.2). Since multi-task framework has obtained remarkable success on various NLP tasks (Fan et al., 2021; Qin et al., 2019a, 2020a; Liang et al., 2020; Xu et al., 2021; Qin et al., 2021), we adopt a vanilla multi-task framework to simultaneously perform QI, HI, and KBI, which has the advantage of extracting the shared knowledge across three tasks.

For both pre-trained models and non-pre-trained models, we introduce delimiter tokens `[SOK]`, `[USR]` and `[SYS]` to signal the beginning of KB, user utterance and system response, respectively, aiming to learn to distinguish the role of KB, user and system behavior in multi-turn dialogues. Specifically, the input of KB is denoted as $\hat{KB}$ = "$[SOK]$ KB $[EOK]$" while input of H is defined as $\hat{H}$ = "$[USR]$ u_1 $[SYS]$ s_1 ... $[USR]$ u_n".

4.1 Non Pre-trained Models

In this approach, we simply concatenate all the previous utterances in the dialogue history and the corresponding KB to form a single textual context, which is shown in Figure 5. For KB representation, we format each knowledge entity into "column name, cell value" pairs instead of "subject, relation, object" triples to save length space. KB representation for ToDs is actually an important issue which is mentioned in our challenge section. Then, we apply f_{non} as the non pre-trained models to obtain the final prediction, which is defined as:

$$\mathbf{y} = f_{\text{non}}([\hat{KB}, \hat{H}, u_n], s_n). \quad (2)$$

In our work, we explore some state-of-the-art non pre-trained models including: *ESIM* (Chen et al., 2017), *InferSent* (Conneau et al., 2017) and *RE2* (Yang et al., 2019).

4.2 Pre-trained Models

We investigate several state-of-the-art BERT-based and BART models, which are illustrated in Figure 5. Given a dialogue $\{(u_1, s_1), \ldots, (u_n, s_n), KB\}$, for BERT-based models, following (Chen et al., 2020), the input can be denoted as (`[CLS]`, $\hat{KB}$, $\hat{H}$, `[SEP]`, s_n, `[SEP]`), where `[CLS]` and `[SEP]` are special symbol for classification token and separator token. After pre-training model encoding, the last layer's hidden representation $\mathbf{h}_{\texttt{CLS}}$ from the `[CLS]` token is used for classification, which can be defined as:

$$\mathbf{y} = \text{Softmax}(\mathbf{W}\mathbf{h}_{\texttt{CLS}} + \mathbf{b}), \quad (3)$$

where $\mathbf{W}$ and $\mathbf{b}$ are the trainable parameters.

For BART, we feed the same sequence to both the encoder and the decoder, using the last hidden state for classification. The class that corresponds to the highest probability is chosen as model prediction, which is illustrated in Figure 5(b).

More specifically, we explore `BERT` (Devlin et al., 2019), `RoBERTa` (Liu et al., 2019), `XLNet` (Yang et al., 2020), `Longformer` (Beltagy et al., 2020) and `BART` (Lewis et al., 2020).

4.3 Training Objective

The training objective is the binary cross-entropy loss, which is defined as:

$$\mathcal{L} \triangleq -\sum_{i=1}^{3} \left(\hat{y}_i \log\left(y_i\right) + \left(1-\hat{y}_i\right) \log\left(1-y_i\right)\right) \tag{4}$$

where y_i is the predicted score between 0 and 1 while $\hat{y}_i$ is the gold label for the i inconsistent type.

5 Experiments

5.1 Implementation Details

For pre-trained models the batch size we use in our framework is selected from $\{4, 8\}$ and learning rate is selected from $\{5e^{-6}\}$ to $\{2e^{-5}\}$ with a step of $\{1e^{-6}\}$. We set the max length to 512 tokens for all models except `Longformer`, of which 3,000 tokens are the max length we take. For the non pre-trained models, we adopt the suggested hyper-parameters in their open-sourced code. All experiments are conducted at TITAN Xp and Tesla V100 GPUs. For all experiments, we select model which performs best on the development set and evaluate it on the test set.

5.2 Evaluation

We adopt overall accuracy (Overall Acc) to evaluate model's performance, measuring the ratio of sample for which both QI, HI and KBI are predicted correctly. Furthermore, to give more detailed analysis, we also calculate F1 score on the QI, HI and KBI labels.

5.3 Human Performance

To measure human performance on the CI-ToD dataset, we ask three experts to judge each sample from dataset. Only if the results of the three experts are consistent, we consider this sample is predicted correctly by human. The human performance is shown in the last row of Table 3.

5.4 Main Results

Table 3 shows the results of the models discussed in the previous section.

From the results, we have the two interesting observations: (1) The human performance is 93.2%. In contrast, all of the non pre-trained and pre-trained models perform significantly worse than humans, demonstrating that there is ample room for improving consistency ability in the task-oriented dialogue; (2) Pretrained models outperform all non-pre-trained models in CI-ToD, which is consistent with results in other literature (Talmor et al., 2019). We think that knowledge learned from pre-training can benefitial to consistency identification.

5.5 Qualitative Analysis

5.5.1 Performance Across Different Consistency Types

We compare human performance and model performance across different consistency types. The results are shown in Table 3. We can observe that humans are good at deciding the all consistency types, indicating that it's easy for human to detect whether a dialogue is consistent because human have a strong reasoning ability. In contrast, we find that the best pre-trained model (`BART`) obtains the worst results on HI type compared with other types (*QI and KBI*). This is because that correctly detecting HI rely on the dialogue context information which faces the challenges of coreference resolution. We will discuss it in details in Section 5.7.

5.5.2 Context Ablation Study

In this section, we analyze the impact of context on final performance. More specifically, we conduct experiments by removing the corresponding dialogue contextual information and only keeping the final user query. Figure 6 shows the results of `BART` without contextual information. We observe that our model drops in all consistency types. This is because dialogue context can help model to understand the whole dialogue topic, which is useful to the consistency detection.

5.5.3 Multi-Task Training vs. Separate Task Training

In this section, we explore the effectiveness of the proposed multi-task framework. In particular, we conduct separate training setting where we use the `BART` to perform each task prediction (QI, HI and KBI) separately. The comparison results are shown in Figure 7, we can observe that model with multi-task training outperforms separate task training paradigm in all metrics, which indicates that QI, HI and KBI tasks are correlated, and thus modeling the correlation across tasks can improve performance.

5.5.4 Using CI-ToD as an Automatic Metric

In this section, we want to further investigate whether it can judge the quality of the utterances by different task-oriented dialogues and be used

Baseline category	Baseline method	QI F1	HI F1	KBI F1	Overall Acc
Non pre-trained Model	ESIM (Chen et al., 2017)	0.512	0.164	0.543	0.432
	InferSent (Conneau et al., 2017)	0.557	0.031	0.336	0.356
	RE2 (Yang et al., 2019)	0.655	0.244	0.739	0.481
Pre-trained Model	BERT (Devlin et al., 2019)	0.691	**0.555**	0.740	0.500
	RoBERTa (Liu et al., 2019)	0.715	0.472	0.715	0.500
	XLNet (Yang et al., 2020)	0.725	0.487	0.736	0.509
	Longformer (Beltagy et al., 2020)	0.717	0.500	0.710	0.497
	BART (Lewis et al., 2020)	**0.744**	0.510	**0.761**	**0.513**
Human	Human Performance	0.962	0.805	0.920	0.932

Table 3: Comparison of varying approaches on CI-ToD dataset.

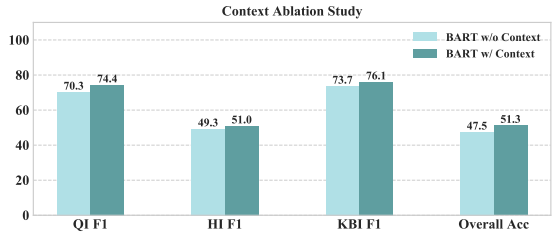

Figure 6: Context Ablation Study

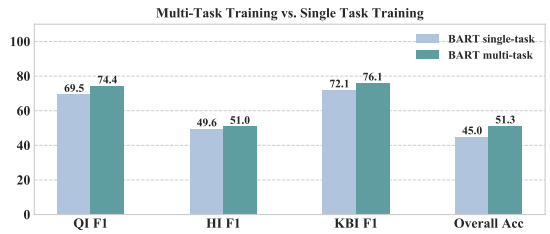

Figure 7: Multi-Task Training vs. Single Task Training

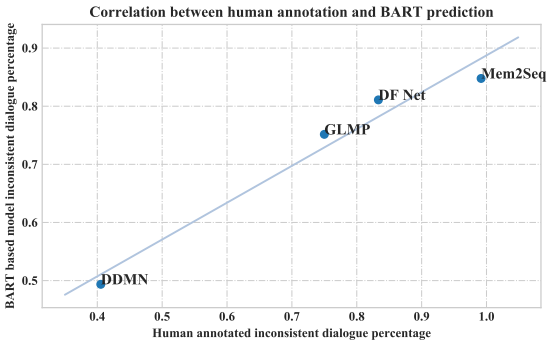

Figure 8: Correlation between human annotation and BART prediction

as an automatic metric checking generation consistency. We compare the overall accuracy of the well-trained best model `BART` with the contradiction rate by human judgements on the utterances generated by different models. In particular, we explore the state-of-the-art end-to-end task-oriented dialogue models (`Mem2seq` (Madotto et al., 2018), `GLMP` (Wu et al., 2019b), `DF-Net` (Qin et al., 2020b), `DDMN` (Wang et al., 2020)). The results are shown in Figure 8 and we can see that the scores are positively correlated with human judgments, with a Pearson correlation coefficient of 0.9. This demonstrates the proposed consistency identification model can be used as a automatic metric to evaluate consistency in task-oriented dialogues.

5.6 Error Analysis

In this section, we empirically divide all the error samples generated with `BART` into three categories, which are shown in Figure 9.

Long KB. When the KB is relatively large, it contains a lot of redundant information which is irrelevant to the current conversation. This redundant information will become noise in the process of model learning and simply flatting the KB into a sequence can not effectively modeling the relevant information. For example, as shown in Figure 9, when the KB is large (49 rows), `BART` predicts the KBI as 1 incorrectly.

Long Dialog History. When the dialogue history is too long, it may contain some noise information. As shown in Figure 9, there are three rounds of dialogues at this time, the system expresses *"palo_alto_medical_foundation is located 2_miles away"* at first round of dialogue while describes *"palo_alto_medical_foundation is located 3_miles away"* at last turn, causing the HI due to the irrelevant middle context.

Coreference Resolution. When there are some implicit or explicit references in the dialogue, it is necessary to resolve the references to restore the intention of the conversation, which greatly increases the difficulty of the model to predict the types of inconsistencies. For example, in Figure 9, the last round of the user's query *"can i have the address"* does not clearly indicate a specific

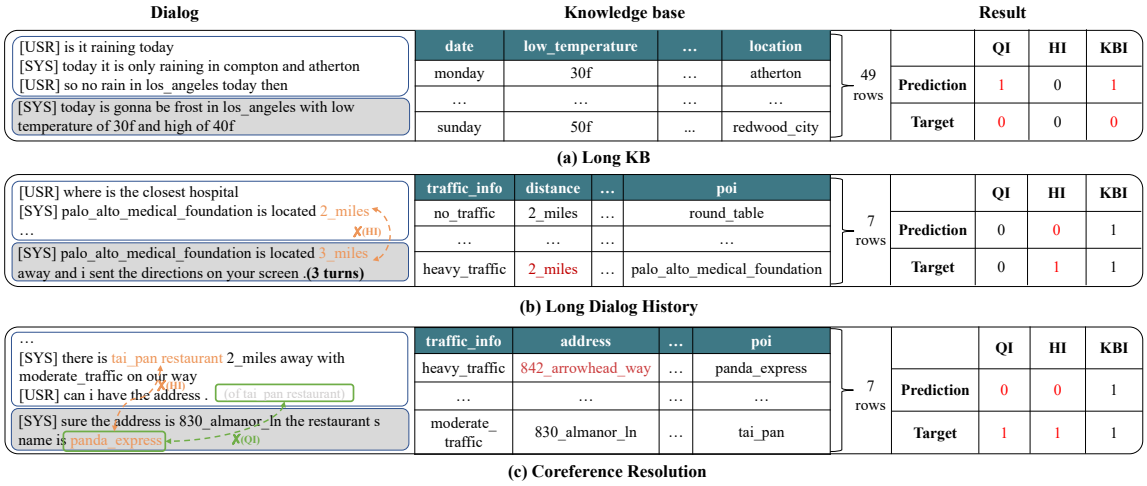

Figure 9: Error type in CI-ToD baseline model's prediction.

object, which confuses model to predict the QI and HI as 0 incorrectly. Actually, by resolving the implicit reference according to the dialogue history, we can know that the reference object of the user's current problem is *"tai_pan restaurant"*, which helps model to obtain correct results.

5.7 Challenges

Based on above analysis, we summarize the current challenges faced by the consistency detection task:

KB Representation. The corresponding Knowledge base is the relational database, which has high-order structure information presented in the original knowledge graph. How to modeling the structure information in the relational knowledge base rather than simply flattening the KB is an interesting research question to investigate. In addition, since the size of KB is relatively big, how to effectively modeling relevant KB information rather than injecting noisy is another challenge to explore.

Effectively Context Modeling. Since some dialogue has extreme long histories, not all context information have a positive influence for the final performance. How to effectively model the long-distance dialogue history and filter irrelevant information is an interesting research topic.

Coreference Resolution. There are multiple coreference resolution in a dialogue, which will result in ambiguity in the user's query, making it difficult for model to predict the consistency label correctly. Thus, how to explicitly conduct coreference resolution to help the consistency detection is an important research question.

Explicit Joint Learning. Though achieving promising performance based on the multi-task training paradigm, the prior work did not "explicitly" model the relationships between the different tasks (HI, QI and KBI task); instead, it adopted shared parameters to "implicitly" model the correlation. However, simply relying on a set of shared parameters cannot make a full interaction to achieve desirable results (Qin et al., 2019a, 2020a). Thus, how to explicitly modeling the correlation between HI, QI and KBI to directly control information flow still deserves to be explored.

6 Related Work

This work is related to the considering consistency in open-domain dialogue. In recent years, several personalized dialogue datasets have been introduced, such as PersonaChat (Zhang et al., 2018) and PersonalDialog (Zheng et al., 2020). These datasets are able to implicitly consider the consistency in dialogue generation, but fail to explicitly teach the model to judge whether the generated system response is consistent.

Another series of related work explicity improve consistency in dialogue. To this end, some benchmarks have been proposed to promote the research. Welleck et al. (2019) made an early step towards reducing the dialogue consistency identification to natural language inference (NLI). Dziri et al. (2019) presented a novel paradigm for evaluating the coherence of dialogue systems by using state-of-the-art entailment techniques and build a synthesized dataset InferConvAI geared toward evaluating consistency in dialogue systems. Nie et al. (2021) introduced the DialoguE COntradiction DE-

tection task (DECODE) and a new conversational dataset containing contradictory dialogues, aiming to evaluate the ability to detect contradictory. Song et al. (2020) proposed a KvPI dataset and profile consistency identification task for open-domain dialogue agents to further evaluate whether the system response is inconsistent with the corresponding profile information. Compared with their work that mainly focus on the open-domain dialogue direction, we aim to fill the gap of consistency identification in task-oriented dialogue systems. Furthermore, we introduce a human-annotated dataset to this end. Besides, we provide some key challenges and future directions to facilitate further research.

7 Conclusion

We studied consistency identification in task-oriented dialogue and introduced a new human-annotated dataset CI-ToD. Further, we analyzed the problems of CI-ToD through extensive experiments and highlight the key challenges of the task. We hope CI-ToD can facilitate future research on consistency identification in task-oriented dialogue.

Acknowledgements

This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 61772153. This work was also supported by the Zhejiang Lab's International Talent Fund for Young Professionals.

Ethical Considerations

Data Access. We collected our data from KVRET dataset (Eric et al., 2017). This dataset is an open-source dataset free for academic research.

Annotation Platform Construction. The annotation interface is built by authors based on open-resource JAVA framework and HTML. The server for collecting and storing annotation was rent from Alibaba Cloud using our funding.

Dataset Collection Process. For the annotation, we first launch interviews of the task introduction with 100 example questions, which is paid as $20, for them to try a few examples to get informed and familiar with the task on onboarding test process. Then during annotation process, annotators were paid $15.0 per hour, and the total human-hours we cost are about 300 hours. After annotation, the authors re-check those examples with mismatched tags, which cost about another 20 hours

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 146–148, Florence, Italy. Association for Computational Linguistics.

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.

Chuang Fan, Chaofa Yuan, Lin Gui, Yue Zhang, and Ruifeng Xu. 2021. Multi-task sequence tagging

for emotion-cause pair extraction via tag distribution refinement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu. 2020. Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 97–106, Online. Association for Computational Linguistics.

Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, Yulan He, and Ruifeng Xu. 2020. Aspect-invariant sentiment features learning: Adversarial multi-task learning for aspect-based sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 825–834.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020a. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8665–8672.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019a. A stack-propagation framework with token-level intent detection for spoken language understanding. 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 2078–2087, Hong Kong, China. Association for Computational Linguistics.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019b. 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. 2020b. 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.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6651–6662, Online. Association for Computational Linguistics.

Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 625–638, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 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*, pages 4100–4110, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Haoyang Wen, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu. 2018. Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3781–3792, Santa Fe, New Mexico, USA. 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, Florence, Italy. Association for Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019b. Global-to-local memory pointer networks for task-oriented dialogue.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 896–907, Online. Association for Computational Linguistics.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and effective text matching with richer alignment features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4699–4709, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2020. Personalized dialogue generation with diversified traits.