**BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling**

Zhaojiang Lin\(^1\), Andrea Madotto\(^1\), Genta Indra Winata\(^1\), Peng Xu\(^1\), Feijun Jiang\(^2\), Yuxiang Hu\(^2\), Chen Shi\(^2\), Pascale Fung\(^1\)

\(^1\)Center for Artificial Intelligence Research (CAiRE)  
\(^2\)The Hong Kong University of Science and Technology  
\(^2\)Alibaba Group  
{zlinao, amadotto, giwinata, pxuab}@connect.ust.hk

Abstract

Task-oriented dialogue (ToD) benchmarks provide an important avenue to measure progress and develop better conversational agents. However, existing datasets for end-to-end ToD modeling are limited to a single language, hindering the development of robust end-to-end ToD systems for multilingual countries and regions. Here we introduce BiToD, the first bilingual multi-domain dataset for end-to-end task-oriented dialogue modeling. BiToD contains over 7k multi-domain dialogues (144k utterances) with a large and realistic bilingual knowledge base. It serves as an effective benchmark for evaluating bilingual ToD systems and cross-lingual transfer learning approaches. We provide state-of-the-art baselines under three evaluation settings (monolingual, bilingual, and cross-lingual). The analysis of our baselines in different settings highlights 1) the effectiveness of training a bilingual ToD system compared to two independent monolingual ToD systems, and 2) the potential of leveraging a bilingual knowledge base and cross-lingual transfer learning to improve the system performance under low resource conditions.

1 Introduction

Task-oriented dialogue (ToD) systems are designed to assist humans in performing daily activities, such as ticket booking, travel planning, and online shopping. These systems are the core modules of virtual assistants (e.g., Apple Siri and Amazon Alexa), and they provide natural language interfaces for online services [1]. Recently, there has been growing interest in developing deep learning-based end-to-end ToD systems [2-4, 5-6, 7-8, 9-10, 11-12, 13-14, 15-16] because they can handle complex dialogue patterns with minimal hand-crafted rules. To advance the existing state-of-the-art, large-scale datasets [17, 18, 19, 20] have been proposed for training and evaluating such data-driven systems.

However, existing datasets for end-to-end ToD modelling are limited to a single language, such as English [17, 18, 19], or Chinese [19, 20]. The absence of bilingual or multilingual datasets not only limits the research on cross-lingual transfer learning [21] but also hinders the development of robust end-to-end ToD systems for multilingual countries and regions.

To tackle the challenge mentioned above, we introduce BiToD, a bilingual multi-domain dataset for task-oriented dialogue modelling. BiToD has 7,232 bilingual dialogues (in English and Chinese), spanning seven services within five domains, where each dialogue is annotated with dialogue states.

\(^{\ast}\) Equal contribution

\(^{2}\) Data and code are available in [https://github.com/HLTCHKUST/BiToD](https://github.com/HLTCHKUST/BiToD)

Hi, can you help me find a place to eat?

Sure! How much do you want to spend and how high of a rating would you prefer?

I’d like to eat at an expensive restaurant rated at least 9.

Got it. What kind of food do you want?

Any type of food is fine, but I want a place with Vegan Options.

RESTAURANTS_SEARCH(dietary=Vegan, rating≥9, price=expensive)

name: Sushi Tsubomi
location: Wan Chai
cuisine: Japanese
price: expensive
rating: 10
dietary: Vegan
available options: 51

Okay. Out of 51 restaurants, I recommend you go to Sushi Tsubomi. They are a Japanese restaurant rated at a 10.

I’d like to visit an attraction nearby as well, rated at an 8 minimum.

Sure, I can help with that. What type of attraction would you like?

I’d like to visit some famous landmarks.

ATTR ACTIONS_SEARCH(rating≥8, type=Landmarks, location=Wan Chai)

name: Hong Kong Convention... location: Wan Chai
type: Sights & Landmarks
address: 1 Expo Dr., Wanchai
phone: 2582 8888
available options: 5

The contribution of this work is three-fold. 1) We propose the first bilingual dataset (BiToD) with a total of 7,232 dialogues for end-to-end ToD modeling. BiToD serves as an effective benchmark for evaluating bilingual ToD systems and cross-lingual transfer learning approaches. 2) We provide novel baselines under the three evaluation settings, i.e., monolingual, bilingual, and cross-lingual. 3) We show the effectiveness of training a bilingual ToD system compared to two independent monolingual ToD systems as well as the potential of leveraging a bilingual knowledge base and cross-lingual transfer learning to improve the system performance under low resource condition.

The paper is organized as follows: We next describe the BiToD data collection methods in Section 2. We then describe our proposed tasks in section 3. Section 4 introduces our baselines, and we finally present and discuss results in Section 5.
2 BiToD Dataset

In this paper, we focus on one-to-one conversations that only involve two speakers (e.g., a user $U$ and a system $S$). A dialogue is a sequence of utterances $(U_1, S_1, U_2, S_2, \text{and so on})$, each a single contribution from one speaker to the dialogue [22]. In ToD, there can be additional API calls in system turn when the user requests the system to search about a certain information. In addition to aforementioned concept, there are several frequently used terms in ToD. Domains are the topics of the current conversation, for example, restaurant domain is about restaurant reservation and metro domain is about taking a metro to somewhere. Multi-domain dialogues are referred as dialogues that involves more than one domain. As shown in Figure 1, the English dialogue involves both restaurant and attraction domains.

BiToD is designed to develop virtual assistants in multilingual cities, regions, or countries (e.g., Singapore, Hong Kong, India, Switzerland, etc.). For the BiToD data collection, we chose Hong Kong since it is home to plenty of attractions, restaurants and more, and is one of the most visited cities globally, especially by English and Chinese speakers. In BiToD, most of Chinese dialogues are collected in simplified Chinese by native mandarin speakers (due to lack of Cantonese annotators). The hiring process is available in Appendix A.2.3 This section describes the knowledge base construction and provides detailed descriptions of the dialogue collection.

2.1 Knowledge Base Collection

We collect publicly available Hong Kong tourism information from the Web, to create a knowledge base that includes 98 metro stations, 305 attractions, 699 hotels, and 1,218 restaurants. For the weather domain, we synthetically generate the weather information on different dates. Then, we implement seven service APIs (Restaurant_Searching, Restaurant_Booking, Hotel_Searching, Hotel_Booking, Attraction_Searching, MTR_info, Weather_info) to query our knowledge base. The knowledge base statistics are shown in Table 3. Although we aim to collect a fully parallel knowledge base, we observe that some items do not include bilingual information. For example, several traditional Cantonese restaurants do not have English names, and similarly, some restaurants do not provide addresses in Chinese. This lack of parallel information reflects the real-world challenges that databases are often incomplete and noisy.

2.2 Dialogue Data Collection

The dialogues are collected through a four-phase pipeline, as shown in Figure 2. We first design a schema, as a flowchart, for each service API, to specify the possible API queries and expected system actions after the API call. Then, user goals are sampled from the knowledge base according to the pre-defined schemas. Based on the user goals, the dialogue simulator interacts with the APIs to generate dialogue outlines. Finally, the dialogue outlines are converted into natural conversations through crowdsourcing. Our data collection methodology extends the Machine-to-Machine (M2M) approaches 23 to bilingual settings to minimize the annotation overhead (time and cost).

Schemas and APIs. The dialogue schema shown as a flowchart (Restaurant_Searching) in Figure 2a specifies the input and output options of the API and the desired system behaviours. To elaborate, the user searches a restaurant by name, location, cuisine, etc. Then the system calls the API and informs the user of the restaurant name and other requested information. If the user is not satisfied with the search results, the system continues searching and provides other options. To ensure the provided services are realistic, we impose a few restrictions, as in [11]. Firstly, each API has a list of required slots, and the system is not allowed to hit the API without specifying values for these slots. For example, the system needs to obtain departure and destination locations before calling the metro-info API. Secondly, the system must confirm the booking information with the user before making any reservations (e.g., restaurant booking).

User Goals. A user goal consists of a list of intents and a set of constraints under each intent. Figure 2b shows a single domain (intent) example where the user’s intent is Restaurant_Search. A constraint is defined with a triple (slot, relation, value) (e.g., (Rating, at_least, 4)). Different from previous work, which defined user constraints as slot-value pairs, we impose slot-value relations 13 (listed in Figure 5b) to promote more diverse user goals. To generate a user goal, we first sample
Figure 2: Illustration of the bilingual dialogues collection pipeline: a) Design a schema for each service API; b) Sample user goals from the bilingual knowledge base (KB) according to schema; c) Based on one of the user goals, the dialogue simulator generates the dialogue outlines while interacting with the APIs; d) Convert the dialogue outlines to natural conversations via crowdsourcing. Note that English and Chinese user goals are sampled independently.

Dialogue Outline Generation. Dialogue outlines are generated by a bilingual dialogue simulator that accepts user goals in both languages as inputs. The dialogue simulator consists of a user agent and a system agent. Both agents interact with each other using a finite set of actions specified by speech acts over a probabilistic automaton designed to capture varied dialogue trajectories [1]. Each speech act takes a slot or slot-relation-value triple as an argument. When the conversation starts, the user agent is assigned a goal, while the system agent is initialized with a set of requests related to the services. During the conversation, the user informs constraints according to the user goal, and the system responds to the user queries while interacting with the service APIs. For some services, the system needs to request all the required slots before querying the APIs. After the API call, the system either informs the search result or searches for other options until the user intents are fulfilled. Following [1], we also augment the value entities during the dialogue outlines generation process, e.g., Tsim Sha Tsui can be replaced with its abbreviation TST, as shown in Figure 2.c. After the user goal is fulfilled by a series of user and system actions, we convert all the actions into natural language using templates. In this phase, we obtain the dialogue states annotations and speech acts automatically for both the user and system sides.

Dialogue Paraphrase. The dialogue outlines are converted to natural dialogues via crowdsourcing. Figure 6 and 7 show the interface for Chinese and English paraphrasing, where workers see the full dialogue and rewrite the dialogue turn by turn. Before the task, workers are asked to read the
instructions, shown in Figure 8. In the instructions, we specify that the paraphrased dialogue should retain the same meaning as the dialogue outline but sound like a real conversation between a user and a professional assistant. The user utterances are expected to be creative and diverse, while the system utterances are expected to be formal and correct. To ensure all the essential information is presented in the new dialogue, we highlight all the entities with bold text. In the user utterances, the highlighted entities are allowed to be paraphrased without losing their original meaning; e.g., “The restaurant should provide Vegan Options” is allowed to be rewritten as “I would like to find a vegan-friendly restaurant”. In contrast, all the entities in the system utterances are required to be unchanged.

Quality Verification. After the dialogue paraphrasing, workers are asked to read through the new dialogue and answer the following questions, as in [23]: 1) Does it seem like a conversation between a user that sounds like you and an assistant that sounds formal? 2) Does it have the same meaning as the original conversation, while still making sense on its own? The first question is for examining whether the new conversation is realistic, and the second question is for verifying whether the dialogue outline and the paraphrased dialogue are valid. Given the two answer options: 1) Yes, 2) No, but I cannot make it better, 97.56% of annotators chose the first option for the first question and 98.89% of them chose the first option for the second question. After the dialogue paraphrasing, we randomly sampled around 3000 dialogues, and asked different sets of annotators to check the quality paraphrased dialogues (in terms of naturalness of the language and dialogue flow). We found around 103 (3.43%) low quality dialogues, and all of them have been annotated.

2.3 Dataset Statistics

We collected 7,232 dialogues with 144,798 utterances, in which 3,689 dialogues are in English, and 3,543 dialogues are in Chinese. We split the data into 80% training, 8% validation, and 12% testing, resulting in 5,787 training dialogues, 542 validation dialogues, and 902 testing dialogues. In Figure 3, we show the main data statistics of the BiToD corpus. As shown in Figure 3a, the lengths of the dialogues vary from 10 turns to more than 50 turns. Multi-domain dialogues, in both English and Chinese, have many more turns compared to single-domains. The most used relation in user goals is equal_to (Figure 3b), and the most common speech-acts (Figure 3c) for users and systems are inform and offer, respectively. Finally, in Table 4 in the Appendix, we list all the informative and requestable slots per domain.

2.4 Dataset Features

Table 1 shows the comparison of the BiToD training set to previous ToD datasets. Prior work for end-to-end ToD modelling only focuses on a single language. Our BiToD is the first bilingual ToD corpus with comparable data size. In addition to its bilingualism, BiToD also provides the following unique features:
<table>
<thead>
<tr>
<th>Language(s)</th>
<th>MultiWoZ</th>
<th>FRAMES</th>
<th>TM-1</th>
<th>SGD</th>
<th>STAR</th>
<th>RiSAWOZ</th>
<th>CrossWoZ</th>
<th>BiToD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dialogues</td>
<td>8,438</td>
<td>1,369</td>
<td>13,215</td>
<td>16,142</td>
<td>10,000</td>
<td>5,012</td>
<td>5,787</td>
<td></td>
</tr>
<tr>
<td>Number of domains</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Total number of turns</td>
<td>115,434</td>
<td>19,986</td>
<td>274,647</td>
<td>329,964</td>
<td>127,833</td>
<td>134,580</td>
<td>84,692</td>
<td>115,638</td>
</tr>
<tr>
<td>Slots</td>
<td>25</td>
<td>6</td>
<td>-</td>
<td>214</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>4,510</td>
<td>3,871</td>
<td>-</td>
<td>14,139</td>
<td>-</td>
<td>4,061</td>
<td>7,871</td>
<td>8,206*</td>
</tr>
<tr>
<td>Deterministic API</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Complex User Goal</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Mixed-Language Context</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 1: Comparison of BiToD to previous ToD datasets. The numbers are provided for the training set except for FRAMES and STAR. *We consider entities in different language as different slots and values.

**Deterministic API.** Given an API query for recommendation services (e.g., restaurant searching and hotel searching), there is typically more than one matched item. Previous works [17][19] randomly sampled one or two items as API results and returned them to users. However, in real-world applications, the system should recommend items according to certain criteria (e.g., user rating). Moreover, the randomness of the API also increases the difficulty of evaluating the models. Indeed, the evaluation metrics in [17][19] rely on delexicalized response templates, which are not compatible with knowledge-grounded generation approaches [10][12]. To address these issues, we implement deterministic APIs by ranking the matched items according to user ratings.

**Complex User Goal.** To simulate more diverse user goals, we impose different relations for slot-value pairs. For example, in the restaurant searching scenarios, a user might want to eat Chinese food (cuisine, equal_to, Chinese), or do not want Chinese food (cuisine, not, Chinese). Figure 3b shows the distribution of different relations in user goals.

**Mixed-Language Context.** Our corpus contains code-switching utterances as some of the items in the knowledge base have mixed-language information. In the example in Figure 1b, the system first recommends a restaurant called *ChocoDuck Bistro* and the user asks for other options. Then the system searches other restaurants with an additional constraint (restaurant_name, not, ChocoDuck Bistro). In this example, both restaurants only have English names, which is a common phenomenon in multilingual regions like Hong Kong. Thus, ToD systems need to handle the mixed-language context to make correct API calls. Note that the code-switching utterances in our dataset are mostly restricted to the interchange between Chinese and English name entities, such as hotel names, attraction names.

**Cross-API Entity Carry-Over** Our corpus includes scenarios where the value of a slot is not presented in the conversation, and the system needs to carry over values from previous API results. In the example in Figure 1a, the user first finds and books a restaurant without specifying the location; then she (她) wants an attraction nearby the restaurant. In this case, the system needs to infer the attraction location (Wan Chai) from the restaurant search result.

3 Tasks & Evaluations

3.1 Dialogue State Tracking

Dialogue state tracking (DST), an essential task for ToD modelling, tracks the users’ requirements over multi-turn conversations. DST labels provide sufficient information for a ToD system to issue APIs and carry out dialogue policies. In this work, we formulate a dialogue state as a set of slot-relation-value triples. We use Joint Goal Accuracy (JGA) to evaluate the performance of the DST. The model outputs are correct when all of the predicted slot-relation-value triples exactly match the oracle triples.
3.2 End-to-End Task Completion

A user’s requests are fulfilled when the dialogue system makes correct API calls and correctly displays the requested information. We use the following automatic metrics to evaluate the performance of end-to-end task completion: 1) Task Success Rate (TSR): whether the system provides the correct entity and answers all the requested information of a given task, 2) Dialogue Success Rate (DSR): whether the system completes all the tasks in the dialogue, 3) API Call Accuracy (APIAcc): whether the system generates a correct API call, and 4) BLEU [24]: measuring the fluency of the generated response.

3.3 Evaluation Settings

**Monolingual.** Under the monolingual setting, models are trained and tested on either English or Chinese dialogues.

**Bilingual.** Under the bilingual setting, models are trained on bilingual dialogues (full training set), and in the testing phase, the trained models are expected to handle dialogues in both languages simultaneously without any language identifiers.

**Cross-lingual.** This setting simulates the condition of lacking data in a certain language, and we study how to transfer the knowledge from a high resource language to a low resource language. Models have full access to the source language in this setting but limited access to the target language (10%).

4 Proposed Baselines

Our proposed baselines are based on the recent state-of-the-art end-to-end ToD modeling approach MinTL [15] and cross-lingual transfer approach MTL [25]. We report the hyper-parameters and training details in the Appendix A.5.

**Notations.** We define a dialogue $D = \{U_1, S_1, \ldots, U_T, S_T\}$ as an alternating set of utterances from user and systems. At turn $t$, we denote a dialogue history as $H_t = \{U_{t-w}, S_{t-w}, \ldots, S_{t-1}, U_t\}$, where $w$ is the context window size. We denote the dialogue state and knowledge state at turn $t$ as $B_t$ and $K_t$, respectively.

4.1 ToD Modeling

Figure 4 describes the workflow of our baseline model. We initialize the dialogue state $B_0$ and knowledge state $K_0$ as empty strings. At turn $t$, the input of our model is the current dialogue history $H_t$, previous dialogue state $B_{t-1}$ and knowledge state $K_{t-1}$. Similar to the text-to-text transfer learning approach [26], we add a prompt $P_B = \text{"TrackDialogueState : "}$ to indicate the generation task. Then, a multilingual sequence-to-sequence (mSeq2Seq) model takes the flattened dialogue history $H_t$ and generates a new dialogue state $B_t$ and knowledge state $K_t$. The model then checks if the generated response is an API or a Response. If the output is an API, the system queries the KB and updates the knowledge $K$; otherwise, the Response is shown to the user.
input sequence and outputs the Levenshtein Belief Spans \((\text{Lev}_t)\) \cite{lev}: 

\begin{equation}
\text{(1)} : \text{Lev}_t = \text{mSeq2Seq}(P_{tB}, H_t, B_{t-1}, K_{t-1}).
\end{equation}

The \(\text{Lev}_t\) is a text span that contains the information for updating the dialogue state from \(B_{t-1}\) to \(B_t\). The updated dialogue state \(B_t\) and a response generation prompt, \(P_{Rt} = "\text{Response} : "\), are used as input. Then, the model will either generate an API name (2a) when an API call is needed at the current turn, or a plain text response directly returned to the user (2b). If the model generates an API name, it is

\begin{equation}
\text{(2a)} : \text{API} = \text{mSeq2Seq}(P_{R_t}, H_t, B_t, K_{t-1}),
\end{equation}

the system will query the API with the constraints in the dialogue state and update the knowledge state \(K_{t-1} \rightarrow K_t\). The updated knowledge state and API name are incorporated into the model to generate the next turn response generation.

\begin{equation}
\text{(2b)} : \text{Response} = \text{mSeq2Seq}(P_{R_t}, H_t, B_t, K_t, \text{API}).
\end{equation}

All the aforementioned generation process are based on a single \text{mSeq2Seq}, and we initialized our model with two pre-trained models, \text{mT5} \cite{mt5} and \text{mBART} \cite{mbart}. 

4.2 Cross-lingual Transfer

Based on the modelling strategy mentioned above, we propose three baselines for the cross-lingual setting.

\text{mSeq2seq}. Directly finetune the pre-trained mSeq2seq models like mBART and mT5 on the 10% dialogue data in the target language.

\text{Cross-lingual Pre-training (CPT)}). First, pre-train the mBART and mT5 models on the source language, then finetune the models on the 10% target language data.

\text{Mixed-Language Pre-training (MLT)}. To leverage the fact that our knowledge base contains the bilingual parallel information for most of the entities, we replace the entities in the source language data (both input sequence and output sequence) with their target language counterpart in our parallel knowledge base to generate the mixed-language training data. We first pre-train the mSeq2seq models with the generated mixed-language data, then finetune the models on the 10% target language data.

\text{Translate Train (TT)}. We study applying the state-of-the-art pre-trained neural machine translation model M2M-100 \cite{m2m100} to our cross-lingual transfer setting. We first translate the dataset of source language to target language, then train the model on the translated dataset. Finally, we finetune the trained model with the 10% target language data. Note that this approach can be applied only when we have additional resource to train a machine translation model from the source language to the target language.

5 Results & Discussion

The main results for DST and end-to-end task completion are reported in Table 2. Note that the API\(\text{Acc}\) is highly correlated with the JGA because the dialogue states contain constraints for issuing APIs. And, the DSR is a more challenging metric compared to TSR because the dialogue might contain 2-5 tasks.

\text{Monolingual vs Bilingual}. Comparing the models that are trained under monolingual and bilingual setting, the latter can leverage more training data and handle tasks in both languages simultaneously without a language identifier. We observe that \text{mT5} achieves better results in the bilingual settings, while \text{mBART} performs better with monolingual training. The underlying reason might be the different pre-training strategies of the two mSeq2seq models. mBART is pre-trained with language tokens in both the encoder and decoder, but in our bilingual setting, we do not provide any language information. Such a discrepancy does not exist in the \text{mT5} model, as it is pre-trained without language tokens.
<table>
<thead>
<tr>
<th>Models</th>
<th>TSR</th>
<th>DSR</th>
<th>API\textsubscript{Acc}</th>
<th>BLEU</th>
<th>JGA</th>
<th>TSR</th>
<th>DSR</th>
<th>API\textsubscript{Acc}</th>
<th>BLEU</th>
<th>JGA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monolingual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MinTL(mBART)</td>
<td>56.00</td>
<td>33.71</td>
<td>57.03</td>
<td>35.34</td>
<td>67.36</td>
<td>56.82</td>
<td>29.35</td>
<td>71.89</td>
<td>20.06</td>
<td>72.18</td>
</tr>
<tr>
<td>MinTL(mT5)</td>
<td>69.13</td>
<td>47.51</td>
<td>67.92</td>
<td>38.48</td>
<td>69.19</td>
<td>53.77</td>
<td>31.09</td>
<td>63.25</td>
<td>19.03</td>
<td>67.35</td>
</tr>
<tr>
<td><strong>Bilingual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MinTL(mBART)</td>
<td>42.45</td>
<td>17.87</td>
<td>65.35</td>
<td>28.76</td>
<td>69.37</td>
<td>40.39</td>
<td>16.96</td>
<td>65.37</td>
<td>5.23</td>
<td>69.50</td>
</tr>
<tr>
<td>MinTL(mT5)</td>
<td>71.18</td>
<td>51.13</td>
<td>71.87</td>
<td>40.71</td>
<td>72.16</td>
<td>57.24</td>
<td>34.78</td>
<td>65.54</td>
<td>22.45</td>
<td>68.70</td>
</tr>
<tr>
<td><strong>Cross-lingual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZH \rightarrow EN (10%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models</td>
<td>TSR</td>
<td>DSR</td>
<td>API\textsubscript{Acc}</td>
<td>BLEU</td>
<td>JGA</td>
<td>TSR</td>
<td>DSR</td>
<td>API\textsubscript{Acc}</td>
<td>BLEU</td>
<td>JGA</td>
</tr>
<tr>
<td>MinTL(mBART)</td>
<td>1.11</td>
<td>0.23</td>
<td>0.60</td>
<td>3.17</td>
<td>4.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>2.14</td>
</tr>
<tr>
<td>+ CPT</td>
<td>36.19</td>
<td>16.06</td>
<td>41.51</td>
<td>22.50</td>
<td>42.84</td>
<td>24.64</td>
<td>11.96</td>
<td>29.04</td>
<td>8.29</td>
<td>28.57</td>
</tr>
<tr>
<td>+ MLT</td>
<td>33.62</td>
<td>11.99</td>
<td>41.08</td>
<td>20.01</td>
<td>55.39</td>
<td>44.71</td>
<td>21.96</td>
<td>54.87</td>
<td>14.19</td>
<td>60.71</td>
</tr>
<tr>
<td>+ TT*</td>
<td>21.61</td>
<td>10.18</td>
<td>27.44</td>
<td>17.86</td>
<td>37.5</td>
<td>43.86</td>
<td>19.78</td>
<td>50.71</td>
<td>14.46</td>
<td>56.76</td>
</tr>
<tr>
<td>MinTL(mT5)</td>
<td>6.78</td>
<td>1.36</td>
<td>17.75</td>
<td>10.35</td>
<td>19.86</td>
<td>4.16</td>
<td>2.20</td>
<td>6.67</td>
<td>3.30</td>
<td>12.63</td>
</tr>
<tr>
<td>+ CPT</td>
<td>44.94</td>
<td>24.66</td>
<td>47.60</td>
<td>29.53</td>
<td>48.77</td>
<td>43.27</td>
<td>23.70</td>
<td>49.70</td>
<td>13.89</td>
<td>51.40</td>
</tr>
<tr>
<td>+ MLT</td>
<td>56.78</td>
<td>33.71</td>
<td>56.78</td>
<td>32.43</td>
<td>58.31</td>
<td>49.20</td>
<td>27.17</td>
<td>50.55</td>
<td>14.44</td>
<td>55.05</td>
</tr>
<tr>
<td>+ TT*</td>
<td>56.43</td>
<td>34.16</td>
<td>57.54</td>
<td>31.2</td>
<td>58.85</td>
<td>47.67</td>
<td>26.08</td>
<td>50.88</td>
<td>14.46</td>
<td>54.01</td>
</tr>
</tbody>
</table>

Table 2: Dialogue state tracking and end-to-end task completion results in monolingual, bilingual, and cross-lingual settings.

**Cross-lingual.** We observe that it is difficult for the baseline models to converge with minimal training data (10%) due to the complex ontology and diverse user goals. Interestingly, pre-training the mSeq2seq models on the source language improves both DST and task completion performance. Such results indicate the excellent cross-lingual transferability of multilingual language models. Furthermore, the mixed-language training strategy further improves the cross-lingual few shot performance, especially the JGA, which suggests that the bilingual knowledge base can facilitate the cross-lingual knowledge transfer in the low resource scenario. Figure 5 shows that, by using 30%-50% target language training data, mT5+MLT is able to achieve comparable results to monolingual full training. We also observe that using a translation model does not further improve the results because many name entities are mistakenly translated.

**Limitations and Future Work.** The main limitation of this work is the low number of languages in the corpus due to the difficulty of collecting the knowledge base in languages other than English and Chinese in Hong Kong. In future work, we plan to extend the dataset to more languages including low resource languages in dialogue research (e.g., Cantonese, Indonesian), to better examine the cross-lingual transferability of end-to-end ToD systems. Another limitation is that the M2M data collection might not cover rare and unexpected user behaviours (e.g., non-collaborative dialogues), as dialogue simulators generate the dialogue outlines. We also note that the dialogues in our dataset do not have common features found in spontaneous dialogue [29, 30], such as ellipses, inversions, repairs [31, 32], and split utterances [33], because the utterances in the dataset are collected by paraphrasing (writing). However, we see BiToD as a necessary step for building robust multilingual ToD systems before tackling even more complex scenarios.

# Related Work

**Datasets for End-to-End ToD Modeling.** Many datasets have been proposed in the past to support various assistant scenarios. In English, Wen et al. [34] collected a single domain dataset with a Wizard-of-Oz (Woz) setup, which was latter extended to multi-domain by many follow-up works [17, 18, 35, 36]. Despite its effectiveness, Woz data collection method is expensive since two annotators need to be synchronized to conduct a conversation, and the other set of annotators need to annotate speech-act and dialogue states further. To reduce the annotation overhead (time and cost), Byrne et al. [23] proposed a Machines Talking To Machines (M2M) self-chat annotations schema. Similarly, Rastogi et al. [4] applied M2M to collect a large-scale schema-guided ToD dataset, and Kottur et
Figure 5: Cross-lingual transfer results (by using mT5+MLT) on ZH→EN (left) and EN→ZH (right) with different ratio (%) of target language data. mT5+MLT achieves comparable results to monolingual full training by using 30%-50% target language data.

Extended Multilingual Datasets for Spoken Language Understanding (SLU). Different from aforementioned datasets focus on dialogue level modeling, there are many SLU datasets have been proposed for utterance level semantic parsing. They serve as effective benchmarks for SLU tasks, such as intent classification, slot filling and MTOP are multilingual datasets designed to study cross-lingual transfer learning for SLU.

7 Conclusion

We present BiToD, the first bilingual multi-domain dataset for end-to-end task-oriented dialogue modeling. BiToD contains over 7k multi-domain dialogues (144k utterances) with a large and realistic knowledge base. It serves as an effective benchmark for evaluating bilingual ToD systems and cross-lingual transfer learning approaches. We provide state-of-the-art baselines under three evaluation settings (monolingual, bilingual and cross-lingual). The analysis of our baselines in different settings highlights 1) the effectiveness of training a bilingual ToD system compared to two independent monolingual ToD systems, and 2) the potential of leveraging a bilingual knowledge base and cross-lingual transfer learning to improve the system performance under low resource conditions.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section A.1 for more information.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section A.5 for more information.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]
In this paper, we propose a new bilingual dataset for end-to-end task-oriented dialogue systems training and evaluation. In the dialogue outlines, we sample user names from 100 most common male names and female names uniformly. However, after the dialogue paraphrase, we found that, out of 144,000 utterances, the assistant call the user “Sir” 73 times and “Madam” 8 times in the English dialogues, while in Chinese dialogues, the assistant call the user “先生 (Sir)” 162 times and “女士 (Madam)” 2 times. Nevertheless, this social bias issue has been addressed by replacing these with gender neutral terms (Sir/Madam). We do not foresee any other direct social consequences or ethical issues. Furthermore, our proposed dataset encourages research in the cross-lingual few shot setting, where fewer data and resources are needed, rendering it energy-efficient models.

We follow datasheets [51] for datasets guideline to document the following:

**A.2.1 Motivation**

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - BiToD is created to benchmark the multilingual ability of end-to-end task oriented dialogue systems. Existing end-to-end benchmarks are limited to a single language (e.g., English or Chinese), thus BiToD fills the need of having a dataset for training and evaluating end-to-end task-oriented dialogue systems in the multilingual and cross-lingual settings.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - HKUST CaiRE team and Alibaba team work together to create this dataset.
- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
  - Alibaba team funded the creation of the dataset.

**A.2.2 Composition**

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  - BiToD is made of conversations (text) between two speakers (user and assistant) and the textual knowledge in return from the API-call (tuple in a DB). BiToD also includes speech-acts for both user and systems, and dialogue state annotations.
- How many instances are there in total (of each type, if appropriate)?
  - BiToD has 7,232 dialogues with 144,798 utterances, in which 3,689 dialogues are in English and 3,543 dialogues are in Chinese.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  - BiToD has been designed from scratch and thus contains all possible instances.
- What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  - Each sample has raw text of conversations, speech-acts for both user and systems, dialogue state annotations, query, and knowledge bases return.
- Is there a label or target associated with each instance? If so, please provide a description.
Each response is annotated with its speech-acts and the response itself is target label.

• Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  – No, we included all the information we had.

• Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.
  – No.

• Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
  – Yes, we split the data into 80% training, 8% validation, and 12% testing, resulting in 5,787 training dialogues, 542 validation dialogues, and 902 testing dialogues.

• Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
  – In 2.44% of the dialogues, the annotators reported that the conversation did not sound formal enough, and in 1.11% of the dialogues, the annotators reported that the dialogues are not valid – did not sound coherent.

• Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
  – Yes, BiToD is self-contained.

• Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.
  – No.

• Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  – No.

• Does the dataset relate to people? If not, you may skip the remaining questions in this section.
  – No.

• Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
  – N/A

• Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
  – N/A

• Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
  – N/A.
A.2.3 Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
   - See main paper.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
   - A dialogue simulator generates dialogue outlines. Then, each utterance in the dialogue outline is paraphrased by Amazon Mechanical Turk (for the English instances) and COREDATA CO.,LTD (for the Chinese instances).

3. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
   - We hire crowdworkers from Amazon Mechanical Turk (for English annotation) and CORE-DATA CO. (for simplified Chinese annotation). For English annotators, we require workers to have Mechanical Turk Masters qualification and have HIT approval rate higher than 95%. For Chinese annotators, we hire only native mandarin speakers that have passed our internal qualification test. We paid them roughly $10-12 per hour, calculated by the average time to write the paraphrase which is approximately 8 minutes.

4. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
   - The data was collected during February 2021 to May 2021.

5. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
   - We have conducted an internal ethical review process by the HKUST ethical team.

6. Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
   - No.

7. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
   - N/A.

8. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
   - Yes, the workers knew the data collection procedure. Screenshots are shown in Figure 6, Figure 7, Figure 8 and Figure 9 in the Appendix.

9. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
   - AMT has its own data policy (https://www.mturk.com/acceptable-use-policy) and COREDATA CO.,LTD (http://www.core-dt.com/).

10. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

3http://www.core-dt.com/
• Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
  – N/A

A.2.4 Preprocessing/cleaning/labeling
• Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
  – No data cleaning or preprocessing is done for the released dataset since the dialogue data were generated by a simulator and only paraphrased by the workers.
• Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.
  – N/A.
• Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
  – N/A.

A.2.5 Uses
• Has the dataset been used for any tasks already? If so, please provide a description.
  – BiToD is a new dataset we collected for end-to-end task-oriented modeling and dialogue state tracking tasks. In this work, we build baseline models on BiToD for the mentioned tasks as a benchmark for future research.
• Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
  – Yes, we release our dataset, code, and baseline models at https://github.com/HLTCHKUST/BiToD
• What (other) tasks could the dataset be used for?
  – BiToD could be used for training dialogue policy by using the speech-act annotation, natural language generation modules, and user simulators.
• Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks)? If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
  – No.
• Are there tasks for which the dataset should not be used? If so, please provide a description.
  – No.

A.2.6 Distribution
• Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
  – No.
• How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
  – It is released on Github at https://github.com/HLTCHKUST/BiToD No DOI.
• When will the dataset be distributed?
It is released at our repository.

• Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
  
  – Apache License 2.0.
  
  [https://github.com/HLTCHKUST/BiToD/blob/main/LICENSE](https://github.com/HLTCHKUST/BiToD/blob/main/LICENSE)

• Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
  
  – No.

• Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
  
  – No.

A.2.7 Maintenance

• Who is supporting/hosting/maintaining the dataset?
  
  – HKUST CAiRE research team.

• How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  
  – Create an open issue on our Github repository or contact the authors (check author list email).

• Is there an erratum? If so, please provide a link or other access point.
  
  – No.

• Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
  
  – No. If we plan to update in the future, we will indicate the information on our Github repository.

• If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
  
  – No.

• Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
  
  – Yes. If we plan to update the data, we will keep the original version available and then release the follow-up version, for example, BiToD-2.0

• If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
  
  – Yes, they can submit a Github pull request or contact us privately.

A.3 Accessibility

1. Links to access the dataset and its metadata.
  
  [https://github.com/HLTCHKUST/BiToD](https://github.com/HLTCHKUST/BiToD)

2. The data is saved in a json format, where an example is shown in the README.md file.

3. HKUST CAiRE team will maintain this dataset on the official company Github account.

4. Apache License 2.0.
  
  [https://github.com/HLTCHKUST/BiToD/blob/main/LICENSE](https://github.com/HLTCHKUST/BiToD/blob/main/LICENSE)
Table 3: Bilingual knowledge base statistics.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Attraction</th>
<th>Hotel</th>
<th>Restaurant</th>
<th>Weather</th>
<th>Metro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>EN</td>
<td>ZH</td>
<td>EN</td>
<td>ZH</td>
<td>EN</td>
</tr>
<tr>
<td># Slots</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td># Entities</td>
<td>1,079</td>
<td>1,118</td>
<td>2,642</td>
<td>2,652</td>
<td>5,489</td>
</tr>
</tbody>
</table>

Table A.4 Data Usage

The authors bear all responsibility in case of violation of rights.

A.5 Training Details

We implement our baselines based on the huggingface Transformers [52]. In all of our experiments, we set the dialogue context window size $w = 2$ and we use the pre-trained model mT5-small and mBART-large. They are trained with batch size 128 using an AdamW [53] optimizer with the initial learning rate of 0.0005 and 0.0001 respectively. In monolingual and bilingual settings, all the models are trained for 8 epochs, while in cross-lingual setting, the models are first trained on source language dialogues for 8 epochs and then fine tune the model on target language for 10 epochs. We use 2 NVIDIA V100 GPUs for mBART training and 2 1080Ti for mT5 training. All the trainings take less than 10 hours. We use greedy decoding in test time. More training information is available in [https://github.com/HLTCHKUST/BiToD](https://github.com/HLTCHKUST/BiToD).

A.6 Error Analysis

In Table 5, 6, 7, 8 and 9 we analyze the typical errors made by our baseline models under different settings. In Table 5, the model fails to predict the value of dietary_restriction, which makes the generated response display the wrong information. In Table 6, the model carry over the wrong slot values from previous turns, and this leads to the same wrong recommendation as in the previous case. In Table 7, the model generates an API call rather than a plain text response because the output of the API call is empty since destination and departure are the same. Then, Table 8 shows an example of the wrong predicted API call name from the cross-lingual model. Finally, Table 9 shows a further example of a monolingual model that hallucinate a slot (location), leading to a wrong API call.
Figure 6: Interface of the Chinese dialogue paraphrasing. Crowd workers are asked to read the instructions before starting the task. All the system entities need to be reserved in the paraphrased dialogues.
Figure 7: Interface of the English dialogue paraphrasing. Crowd workers are asked to read the instructions before starting the task. All the system entities need to be reserved in the paraphrased dialogues.
### Table 4: BiToD English ontology.

<table>
<thead>
<tr>
<th>Domain</th>
<th>API</th>
<th>Informable Slots</th>
<th>Requestable Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant</td>
<td>search</td>
<td>dietary_restrictions, cuisine, name, price_level, location, rating</td>
<td>cuisine, name, phone_number, available_options, address, rating</td>
</tr>
<tr>
<td>booking</td>
<td></td>
<td>name, date, user_name, time, number_of_people,</td>
<td>ref_number</td>
</tr>
<tr>
<td>Attraction</td>
<td>search</td>
<td>type, name, location, rating</td>
<td>name, phone_number, available_options, address, rating</td>
</tr>
<tr>
<td>Metro</td>
<td>MTR</td>
<td>departure, destination</td>
<td>shortest_path, estimated_time, price</td>
</tr>
<tr>
<td>Weather</td>
<td>search</td>
<td>day, city</td>
<td>day, max_temp, weather, city, min_temp</td>
</tr>
<tr>
<td>Hotel</td>
<td>search</td>
<td>name, stars, price_level, location, rating</td>
<td>available_options, name, price_per_night, rating</td>
</tr>
<tr>
<td>booking</td>
<td></td>
<td>name, number_of_rooms, start_day, user_name, start_month, number_of_nights</td>
<td>ref_number</td>
</tr>
</tbody>
</table>

| Act type   | affirm, offer, request_more, notify_fail, confirm, inform_intent, goodbye, negate, inform, request, notify_success, thank_you, greeting, request_update |
| Rel. type  | equal_to, not, less_than, at_least, one_of |

### Dialogue History

**User:** Hello
**System:** Hello, how shall I assist you?
**User:** Could you please suggest a restaurant with moderate price range.
**System:** Do you have any dietary restrictions and preferring high rating restaurant?
**User:** It should have Gluten Vegan Options with at least 3 rating.
**System:** Do you have any other preference?
**User:** That is what I want.

**Gold:** I have searched and found 210 restaurants for you and I will recommend ChocoDuck Bistro that serves Italian food with 10 rating.
**Pred:** I have 135 available restaurants, I recommend Chullschick. It serve Peruvian food, and its rating is 10

### Dialogue State Tracking

<table>
<thead>
<tr>
<th>Intents</th>
<th>Gold Slot-Values</th>
<th>Pred Slot-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>restaurants_search</td>
<td>price_level = moderate</td>
<td>price_level = moderate</td>
</tr>
<tr>
<td></td>
<td>dietary_restriction = Vegan Options rating &gt; 3</td>
<td>dietary_restriction = Gluten Free Options rating &gt; 3</td>
</tr>
</tbody>
</table>

Table 5: Example of MT5 generation under bilingual setting. In this turn the model generates the wrong value for one of the slots (i.e., dietary_restriction), which is then reflected to the generated response in the next turn.
**Instructions**

You’ll be shown an unnatural computer generated conversation between a **USER** and a **SYSTEM**. Your task is to paraphrase the messages in the **Dialogue Outline** in order to create a new conversation that has the exact same meaning but sounds like a real conversation between a user and a professional assistant.

Feel free to be creative with your paraphrased messages, as long as they meet the following important requirements:

- User messages need to look like something you would say to a smart assistant.
- Assistant messages need to look formal -- just like chat replies from a human customer service agent.
- All of your paraphrased messages must have the same meaning as the original message.
- Your paraphrased conversation needs to make sense on its own.
- Make sure to ret.ain all the named entities (e.g., number, names, address etc.) in system utterances.

When done rewording/rephrasing the original phrases, you’ll be asked to re-read the entire dialog as a whole and verify it makes sense as a stand alone dialog, as opposed to paraphrasing each sentence without looking at the result as a whole.

**Check a full example here.**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Dialogue Outline</th>
<th>Conversational Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER</td>
<td>I want to find a restaurant to eat.</td>
<td>Hi! I am looking for a restaurant. Do you have any recommendations?</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>Do you have any requirements for restaurant rating?</td>
<td>Delighted to be of service. To get started, can you tell me what rating level are you looking for?</td>
</tr>
<tr>
<td>USER</td>
<td>Any ratings are fine.</td>
<td>I don’t really care about ratings.</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>Do you have any dietary restrictions?</td>
<td>Okay, how about dietary restrictions?</td>
</tr>
<tr>
<td>USER</td>
<td>The restaurant should be/ have Vegan Options.</td>
<td>Oh yes, I am actually Vegan!</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>What kind of food do you like? Do you have any preferences for price level?</td>
<td>Which cuisine do you like and do you have any price range in mind?</td>
</tr>
<tr>
<td>USER</td>
<td>The restaurant should have Australian food. Any prices are fine for me.</td>
<td>Yes, I would like to have Australian food. I don’t care much about the price.</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>Any requirements about the location?</td>
<td>Where the restaurant should be?</td>
</tr>
<tr>
<td>USER</td>
<td>The restaurant should be located in Central District.</td>
<td>Somewhere near central.</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>There are 4 available restaurants, I recommend Elixir. It serve Australian food, and its rating is 10.</td>
<td>Ok, there are 4 restaurants that match your requirements. I would recommend the Elixir, an Australian restaurant with 10 star rating.</td>
</tr>
<tr>
<td>USER</td>
<td>That’s all I need. Byebye.</td>
<td>Perfect, that’s all thanks</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>goodbye</td>
<td>Always at your service. Have a good day.</td>
</tr>
</tbody>
</table>

Figure 8: Detailed instructions for English dialogue paraphrasing.
**Figure 9:** Examples for English dialogue paraphrasing.

<table>
<thead>
<tr>
<th>Good examples</th>
<th>Bad examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary</strong></td>
<td><strong>Examples</strong></td>
</tr>
<tr>
<td>This example is good because it includes a very user diverse response.</td>
<td>This example miss some of the original entities. In this case the</td>
</tr>
<tr>
<td>• Original: I want to know the weather information. On <strong>saturday</strong>, in <strong>Melbourne</strong>.</td>
<td>Paraphrase misses the phone number. If you forget some of the</td>
</tr>
<tr>
<td>• Paraphrased: Hi, I am in <strong>Melbourne</strong> this week. How’s the weather like this <strong>saturday</strong>?</td>
<td>entities, the interface will hint you the missing entities.</td>
</tr>
<tr>
<td>This example is good because include all (no more, no less) the entities</td>
<td>• Original: I found a attraction called <strong>Lamma Island</strong>, its rating</td>
</tr>
<tr>
<td>from the original dialogue.</td>
<td>is 9. Its address is <strong>Lamma Island</strong>, <strong>Islands District</strong> and</td>
</tr>
<tr>
<td>• Original: There are 7 available restaurants, I recommend <strong>Enchanted Garden Restaurant</strong>, its rating is 9.</td>
<td>phone number is +602 2962 8100.</td>
</tr>
<tr>
<td>• Paraphrased: There are 7 restaurant that covers your description.</td>
<td>• Paraphrased: There is <strong>Lamma Island</strong>, it has a very high rating, like 9, and it’s in the <strong>Lamma Island</strong>, <strong>Islands District</strong></td>
</tr>
<tr>
<td>I would recommended the <strong>Enchanted Garden Restaurant</strong> which has a high rating of 9.</td>
<td>This example provide more information than the one present in the original</td>
</tr>
<tr>
<td>in the <strong>USER</strong> turns, the paraphrase can un-normal and some of the</td>
<td>entities. Please try to stick to the entities in the Original text.</td>
</tr>
<tr>
<td>entities can be paraphrased. For example, <strong>cheap</strong> can become <strong>economic</strong>,</td>
<td>• Original: I would like to make a reservation for a hotel.</td>
</tr>
<tr>
<td>but please keep the numerical entity in the paraphrase.</td>
<td>• Paraphrased: I would like to make a reservation for a 3 stars hotel.</td>
</tr>
<tr>
<td>• Original: I am looking for a hotel of <strong>cheap</strong> price level. I would like a 1 star hotel. I am good with all locations.</td>
<td>Avoid non-sensical paraphrases and do not copy directly from original. Try to be creative.</td>
</tr>
<tr>
<td>• Paraphrased: Hi, I would like an <strong>economic</strong> hotel. I don’t care about location, but it needs to be 1 star.</td>
<td>• Original: Any requirements about the location? Do you have any preferences for price level?</td>
</tr>
<tr>
<td>• Paraphrased: <strong>The user</strong> is a hotel enthusiast, looking for high ratings, but no particular price range.</td>
<td>• Paraphrased: <strong>The user</strong> is a hotel enthusiast, looking for high ratings, but no particular price range.</td>
</tr>
</tbody>
</table>
**Dialogue History**

*User:* Can you indicate a hotel of at least 2 rating?

*System:* Of 699 available hotels, I recommend the rated 10 Cordis, Hong Kong.

*User:* Is there other options? I want one near Hennessy Road

*System:* Of 6 available hotels, I recommend The St. Regis Hong Kong, rated 10.

*User:* How much is the fee?

*System:* 2147 HKD per night.

*User:* Can you please book it for me?

*System:* How many rooms would you like to book?

*User:* I only need 1.

*System:* When will the check in be?

*User:* It will be May 3, book it in the name of Linda.

*System:* How many nights will you be staying?

*User:* You can book it for 11 nights.

*System:* Confirmation: The St. Regis Hong Kong hotel will be booked for 11 nights starting from May 3, 1 single room.

*User:* Affirmative.

*System:* Ok. ZQQZIQPV is your reserve number.

*User:* Can you select any Arabic or Latin restaurants in the area?

---

**Gold:**
A good pick is Zahrabel, it serves Arabic food and its rating is 9. The address is 25/F, 239 Hennessy Road, Wanchai and phone number is +852 5503 9181.

**Pred:**
Ok. I have 24 available restaurants, I recommend Lamees. It serve Arabic food, and its rating is 10.

---

**Dialogue State Tracking**

<table>
<thead>
<tr>
<th>Intents</th>
<th>Gold Slot-Values</th>
<th>Pred Slot-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotels_search</td>
<td>rating &gt; 2</td>
<td>rating &gt; 2</td>
</tr>
<tr>
<td></td>
<td>location = Hennessy Road</td>
<td>location = Hennessy Road</td>
</tr>
<tr>
<td>hotels_booking</td>
<td>name = The St. Regis Hong Kong</td>
<td>name = The St. Regis Hong Kong</td>
</tr>
<tr>
<td></td>
<td>number_of_rooms = 1</td>
<td>number_of_rooms = 1</td>
</tr>
<tr>
<td></td>
<td>start_day = 3</td>
<td>start_day = 3</td>
</tr>
<tr>
<td></td>
<td>start_month = 5</td>
<td>start_month = 5</td>
</tr>
<tr>
<td></td>
<td>user_name = Linda</td>
<td>user_name = Linda</td>
</tr>
<tr>
<td></td>
<td>number_of_nights = 11</td>
<td>number_of_nights = 11</td>
</tr>
<tr>
<td>restaurants_search</td>
<td>location = Hennessy Road</td>
<td>cuisine ∈ {Arabic, Latin}</td>
</tr>
<tr>
<td></td>
<td>cuisine ∈ {Arabic, Latin}</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Example of MT5 generation under bilingual setting. In this turn the model curry over the wrong value for one slot (i.e., restaurant_search.location), which is then reflected to the generated response in the next turn.
Dialogue History

**User:** Hello. Can you help me find a place to eat?

**System:** Would you like to choose the rating value for the restaurant?

**User:** Not really. I just want to make sure they have Vegan Options.

**System:** Did you have a price range or cuisine type in mind?

**User:** Any cuisine will work, but it needs to be moderately priced.

**System:** I’ve found 210 options with those criteria. May I recommend an Italian place with a rating of 10 called ChocoDuck Bistro?

**User:** That sounds good. Can I make a reservation?

**System:** Definitely! How many will be in your party and what time would you like to book?

**User:** There will be 14 of us, and we can be there by 4:10 in the afternoon on Sunday.

**System:** Wonderful, and what name should I put that under?

**User:** Put it under Danielle, please.

**System:** Great! Just to confirm your reservation for Sunday at 4:10 in the afternoon at ChocoDuck Bistro for a party of 14, correct?

**User:** That’s right. Confirm the reservation.

**System:** Your reservation is booked and the reference number is TA82102K.

**User:** Alright, now can you help me find a tour of the area?

**System:** Sure. You can use the 9 rated attraction called Peak Tram Fast-Track Guided Tour located at Central MTR Exit K. They can be reached by phone at +852 3462 6208.

**User:** Ok, and how to I get on the MTR from the restaurant to the tour?

**Gold:** Of course, you’ll need to Take the brown line of the Kam Sheung Road station. Then change at Nam Cheong station from brown line to orange line. Get off the train at Hong Kong station. That will take approximately 24 mins and cost 10.44 HKD.

**Pred:** HKMTR en

Dialogue State Tracking

<table>
<thead>
<tr>
<th>Intents</th>
<th>Gold Slot-Values</th>
<th>Pred Slot-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>restaurants_search</td>
<td>rating &gt; don’t care, dietary_restrictions = Vegan Options, price_level = moderate, cuisine = don’t care</td>
<td>rating &gt; don’t care, dietary_restrictions = Vegan Options, price_level = moderate, cuisine = don’t care</td>
</tr>
<tr>
<td></td>
<td>name = ChocoDuck Bistro, number_of_people = 14, time = 4:10 in the afternoon, date = Sunday, user_name = Danielle</td>
<td>name = ChocoDuck Bistro, number_of_people = 14, time = 4:10 in the afternoon, date = Sunday, user_name = Danielle</td>
</tr>
<tr>
<td>attractions_search</td>
<td>type = Tours</td>
<td>location = New Territories</td>
</tr>
<tr>
<td>mtr</td>
<td>destination = Peak, departure = New Territories</td>
<td>destination = New Territories, departure = New Territories</td>
</tr>
</tbody>
</table>

Table 7: Example of MT5 generation under bilingual setting. In this turn the model predicts several wrong slot values, and this lead to generate and API-call (HKMTR en) rather than a plain response.

Dialogue History

**User:** I need to find a place to eat.

**System:** Alright, what type of restaurant are you looking for? Any dietary needs?

**User:** I don’t care about either cuisine or dietary needs.

**Gold:** How about rating?

**Pred:** What sort of cuisine do you prefer?

Dialogue State Tracking

<table>
<thead>
<tr>
<th>Intents</th>
<th>Gold Slot-Values</th>
<th>Pred Slot-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>restaurants_search</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>attractions_search</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8: Example of MT5+MLT generation under cross-lingual setting. In this case the model predicts the wrong API-Call (attractions_search instead of restaurants_search). This lead to a fluent but repetitive response.
I need to find a place to eat.

Alright, what type of restaurant are you looking for? Any dietary needs?

I don’t care about either cuisine or dietary needs.

How about rating?

I want to eat at an expensive restaurant that’s rated at least 2.

Okay, I found 184 restaurants. I recommend the Seafood restaurant The Ocean by Olivier Bellin, which is rated 10.

Can you find something near Eastern District?

Yes, I found 1 European restaurant, Le 188° Restaurant & Lounge, rated at 9.

Okay, and what’s their phone and address?

It’s located at 41/F, Harbour Grand Hong Kong, MTR Fortress Hill Station, Exit A, 23 Oil Street, North Point and their number is +852 2121 2693.

Alright, I want to make a reservation there for 7:20 pm on Sunday.

Sure, I can do that. Can you tell me how many people will be there and which name I should reserve it under?

Yes, book a table for twenty one under my name Ryan.

Alright, to confirm: you’re reserving a table for twenty one at Le 188° Restaurant & Lounge on Sunday at 7:20 pm. Is that correct?

Yes that’s right.

Okay, it is reserved. Your reference number is 1GKOTNWV.

Thanks, are there any Water & Amusement Parks nearby?

Yes, there are 4, and I recommend Ocean Park with a rating of 9.

Where is that, and what’s their number?

It’s located at No.180 Wong Chuk Hang Road, Aberdeen and their number is +852 3923 2323.

Table 9: Example of MT5 generation under monolingual setting. In this turn the model hallucinate one slot (i.e., attraction_search.location), which leads to a different query result and thus the generated response is different from the gold.