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)
[1]The Hong Kong University of Science and Technology
[2]Alibaba Group
`{zlinao, amadotto, giwinata, pxuab}@connect.ust.hk`

A Appendix

A.1 Ethics Statements

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.

A.2 Dataset documentation and intended uses

We follow datasheets [1] 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.

---

* Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

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 it-self 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

- 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.
- 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 [2] (for the Chinese instances).
- 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 COREDATA CO. (for simplified Chinese annotation). For English annotators, we require workers have Mechanical Turk Masters qulification 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.
- 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.
- 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.
- Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

---

[2] http://www.core-dt.com/

  - No.
- 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.
- 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 1, Figure 2, Figure 3 and Figure 4 in the Appendix.
- 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/`).
- 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).
  - `https://www.mturk.com/acceptable-use-policy` and `http://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 will 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`
- 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.

Domain	**Attraction**		**Hotel**		**Restaurant**		**Weather**		**Metro**	
Language	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
# Slots	7	7	9	9	13	13	5	5	5	5
# Entities	1,079	1,118	2,642	2,652	5,489	5,035	77	77	161	161

Table 1: Bilingual knowledge base statistics.

  - 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`
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`

A.4 Data Usage

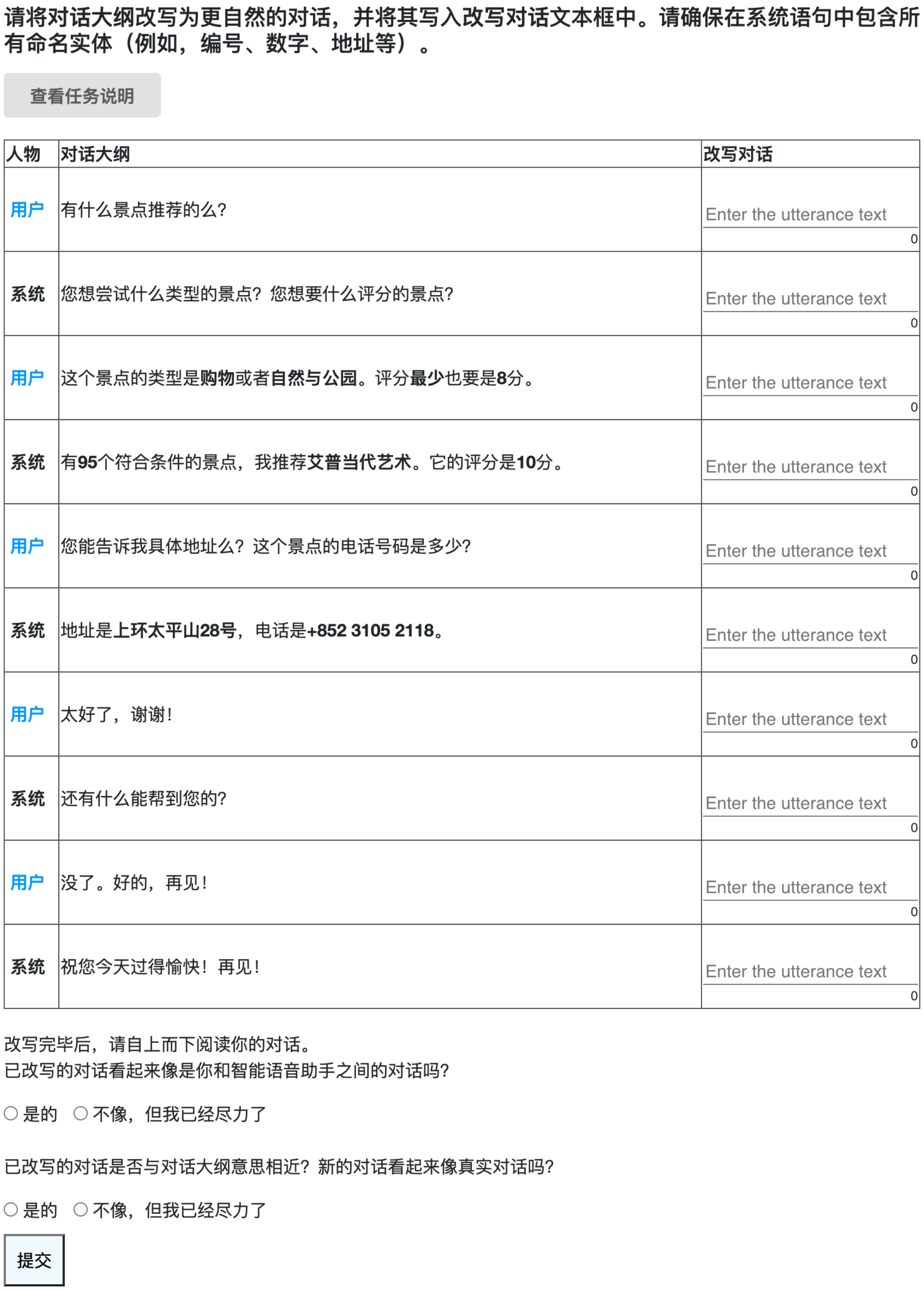
The authors bear all responsibility in case of violation of rights.

A.5 Training Details

We implement our baselines based on the huggingface Transformers [2]. 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 [3] 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`.

A.6 Error Analysis

In Table 3, 4, 5, 6 and 7 we analyze the typical errors made by our baseline models under different settings. In Table 3, the model fails to predict the value of dietary_restriction, which makes the generated response display the wrong information. In Table 4, 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 5 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 6 shows an example of the wrong predicted API call name from the cross-lingual model. Finally, Table 7 shows a further example of a monolingual model that hallucinate a slot (location), leading to a wrong API call.



请将对话大纲改写为更自然的对话，并将其写入改写对话文本框中。请确保在系统语句中包含所有命名实体（例如，编号、数字、地址等）。

人物	对话大纲	改写对话
用户	有什么景点推荐的么?	Enter the utterance text 0
系统	您想尝试什么类型的景点?您想要什么评分的景点?	Enter the utterance text 0
用户	这个景点的类型是**购物**或者**自然与公园**。评分**最少**也要是**8**分。	Enter the utterance text 0
系统	有**95**个符合条件的景点，我推荐**艾普当代艺术**。它的评分是**10**分。	Enter the utterance text 0
用户	您能告诉我具体地址么?这个景点的电话号码是多少?	Enter the utterance text 0
系统	地址是**上环太平山28号**，电话是**+852 3105 2118**。	Enter the utterance text 0
用户	太好了，谢谢!	Enter the utterance text 0
系统	还有什么能帮到您的?	Enter the utterance text 0
用户	没了。好的，再见!	Enter the utterance text 0
系统	祝您今天过得愉快！再见!	Enter the utterance text 0

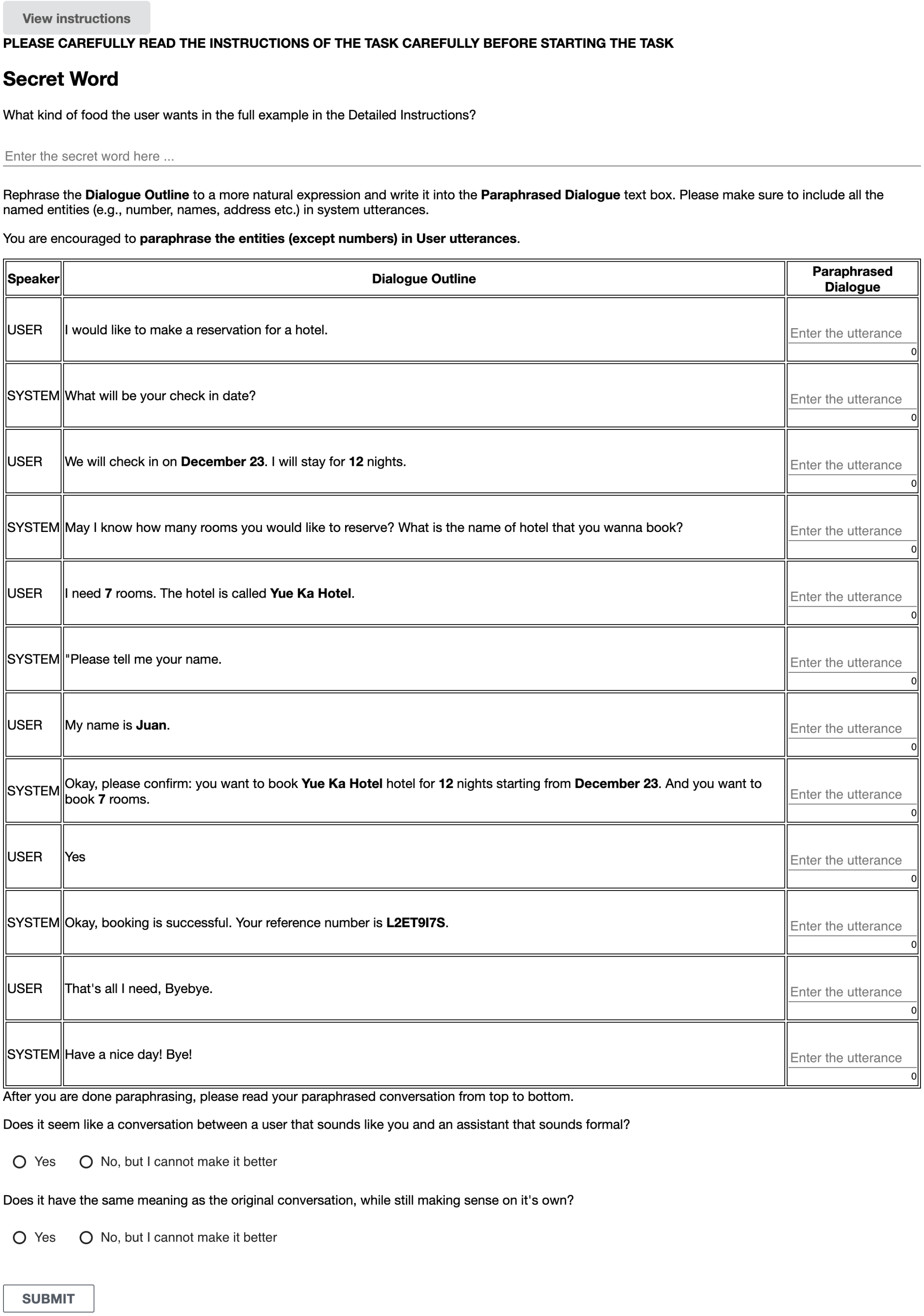
改写完毕后，请自上而下阅读你的对话。
已改写的对话看起来像是你和智能语音助手之间的对话吗?

○ 是的 ○ 不像，但我已经尽力了

已改写的对话是否与对话大纲意思相近?新的对话看起来像真实对话吗?

○ 是的 ○ 不像，但我已经尽力了

Figure 1: 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.



PLEASE CAREFULLY READ THE INSTRUCTIONS OF THE TASK CAREFULLY BEFORE STARTING THE TASK

Secret Word

What kind of food the user wants in the full example in the Detailed Instructions?

Enter the secret word here ...

Rephrase the **Dialogue Outline** to a more natural expression and write it into the **Paraphrased Dialogue** text box. Please make sure to include all the named entities (e.g., number, names, address etc.) in system utterances.

You are encouraged to **paraphrase the entities (except numbers) in User utterances**.

Speaker	Dialogue Outline	Paraphrased Dialogue
USER	I would like to make a reservation for a hotel.	Enter the utterance 0
SYSTEM	What will be your check in date?	Enter the utterance 0
USER	We will check in on **December 23**. I will stay for **12** nights.	Enter the utterance 0
SYSTEM	May I know how many rooms you would like to reserve? What is the name of hotel that you wanna book?	Enter the utterance 0
USER	I need **7** rooms. The hotel is called **Yue Ka Hotel**.	Enter the utterance 0
SYSTEM	"Please tell me your name.	Enter the utterance 0
USER	My name is **Juan**.	Enter the utterance 0
SYSTEM	Okay, please confirm: you want to book **Yue Ka Hotel** hotel for **12** nights starting from **December 23**. And you want to book **7** rooms.	Enter the utterance 0
USER	Yes	Enter the utterance 0
SYSTEM	Okay, booking is successful. Your reference number is **L2ET9I7S**.	Enter the utterance 0
USER	That's all I need, Byebye.	Enter the utterance 0
SYSTEM	Have a nice day! Bye!	Enter the utterance 0

After you are done paraphrasing, please read your paraphrased conversation from top to bottom.

Does it seem like a conversation between a user that sounds like you and an assistant that sounds formal?

○ Yes ○ No, but I cannot make it better

Does it have the same meaning as the original conversation, while still making sense on it's own?

○ Yes ○ No, but I cannot make it better

Figure 2: 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

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

Table 2: BiToD English ontology.

	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

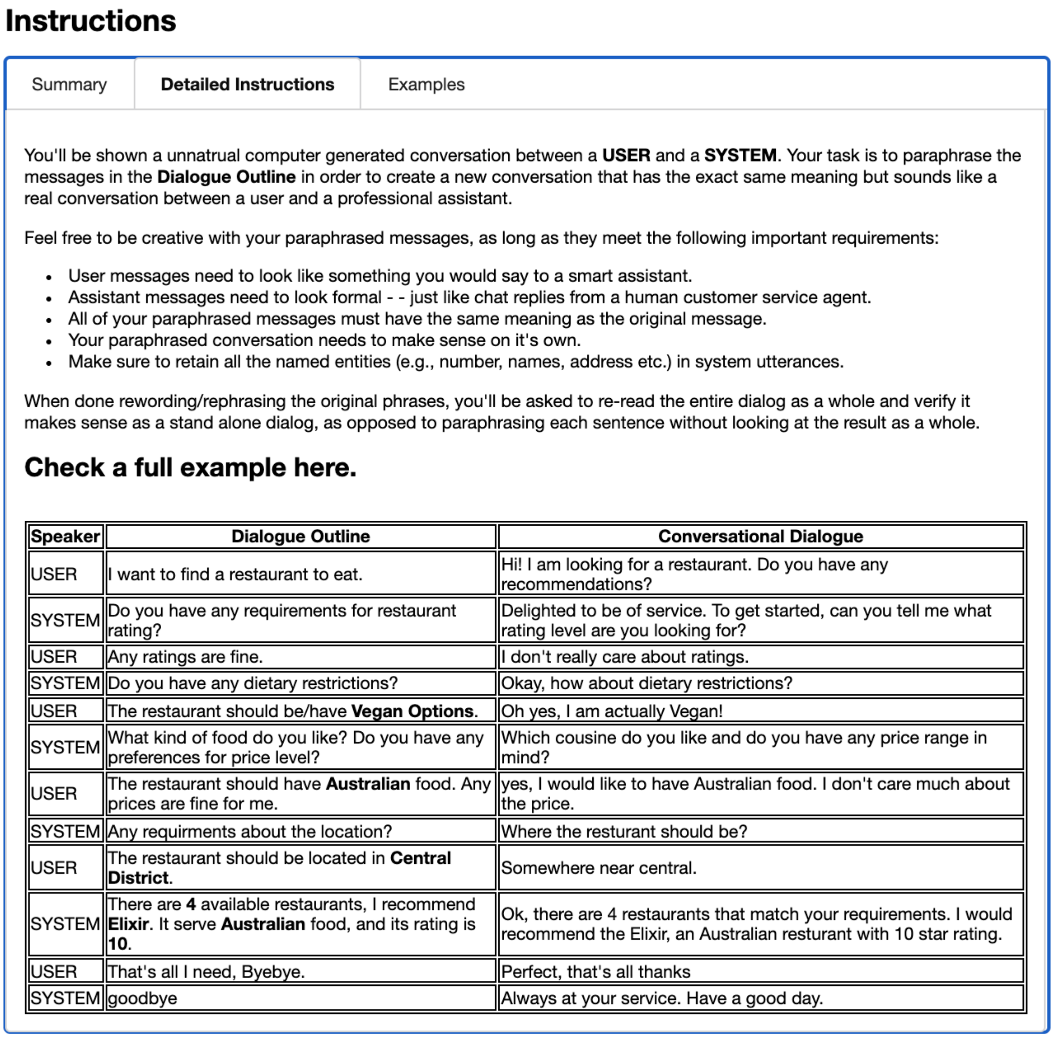
Dialogue State Tracking

Intents	Gold Slot-Values	Pred Slot-Values
restaurants_search	`price_level` = moderate `dietary_restriction` = **Vegan Options** `rating` > 3	`price_level` = moderate `dietary_restriction` = **Gluten Free Options** `rating` > 3

Table 3: 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.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.



Instructions

You'll be shown a unnatrual 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 it's own.
- Make sure to retain 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.

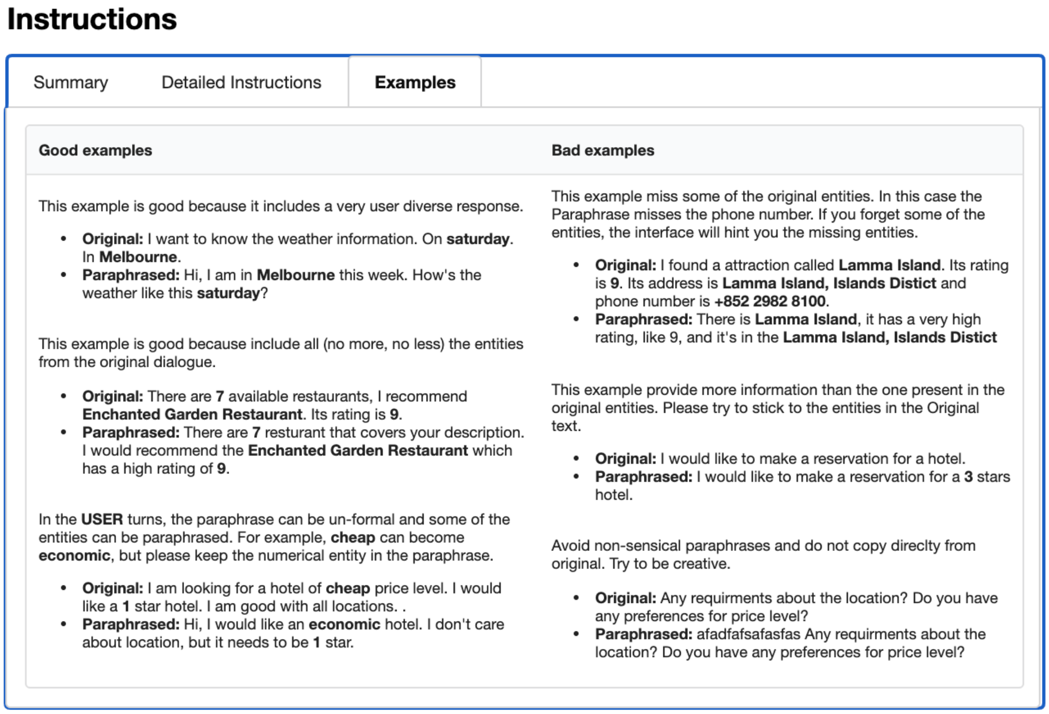
Check a full example here.

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

Figure 3: Detailed instructions for English dialogue paraphrasing.

[2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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



Instructions

Good examples

This example is good because it includes a very user diverse response.

- **Original:** I want to know the weather information. On **saturday**. In **Melbourne**.
- **Paraphrased:** Hi, I am in **Melbourne** this week. How's the weather like this **saturday**?

This example is good because include all (no more, no less) the entities from the original dialogue.

- **Original:** There are **7** available restaurants, I recommend **Enchanted Garden Restaurant**. Its rating is **9**.
- **Paraphrased:** There are **7** resturant that covers your description. I would recommend the **Enchanted Garden Restaurant** which has a high rating of **9**.

In the **USER** turns, the paraphrase can be un-formal and some of the entities can be paraphrased. For example, **cheap** can become **economic**, but please keep the numerical entity in the paraphrase.

- **Original:** I am looking for a hotel of **cheap** price level. I would like a **1** star hotel. I am good with all locations. .
- **Paraphrased:** Hi, I would like an **economic** hotel. I don't care about location, but it needs to be **1** star.

Bad examples

This example miss some of the original entities. In this case the Paraphrase misses the phone number. If you forget some of the entities, the interface will hint you the missing entities.

- **Original:** I found a attraction called **Lamma Island**. Its rating is **9**. Its address is **Lamma Island, Islands Distict** and phone number is **+852 2982 8100**.
- **Paraphrased:** There is **Lamma Island**, it has a very high rating, like 9, and it's in the **Lamma Island, Islands Distict**

This example provide more information than the one present in the original entities. Please try to stick to the entities in the Original text.

- **Original:** I would like to make a reservation for a hotel.
- **Paraphrased:** I would like to make a reservation for a **3** stars hotel.

Avoid non-sensical paraphrases and do not copy direclty from original. Try to be creative.

- **Original:** Any requirments about the location? Do you have any preferences for price level?
- **Paraphrased:** afadfafsafasfas Any requirments about the location? Do you have any preferences for price level?

Figure 4: Examples for English dialogue paraphrasing.

	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

Intents	Gold Slot-Values	Pred Slot-Values
hotels_search	`rating` > 2 `location` = Hennessy Road	`rating` > 2 `location` = Hennessy Road
hotels_booking	`name` = The St. Regis Hong Kong `number_of_rooms` = 1 `start_day` = 3 `start_month` = 5 `user_name` = Linda `number_of_nights` = 11	`name` = The St. Regis Hong Kong `number_of_rooms` = 1 `start_day` = 3 `start_month` = 5 `user_name` = Linda `number_of_nights` = 11
restaurants_search	`location` **= Hennessy Road** `cuisine` $\in$ {Arabic, Latin}	`cuisine` $\in$ {Arabic, Latin}

Table 4: 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

Intents	Gold Slot-Values	Pred Slot-Values
restaurants_search	`rating` > don't care `dietary_restrictions` = Vegan Options `price_level` = moderate `cuisine` = don't care	`rating` > don't care `dietary_restrictions` = Vegan Options `price_level` = moderate `cuisine` = don't care
restaurants_booking	`name` = ChocoDuck Bistro `number_of_people` = 14 `time` = 4:10 in the afternoon `date` = Sunday `user_name` = Danielle	`name` = ChocoDuck Bistro `number_of_people` = 14 `time` = 4:10 in the afternoon `date` = Sunday `user_name` = Danielle
attractions_search	`type` = **Tours**	`location` = **New Territories**
mtr	`destination` = **Peak** `departure` = New Territories	`destination` = **New Territories** `departure` = New Territories

Table 5: 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

Intents	Gold Slot-Values	Pred Slot-Values
restaurants_search	-	-
attractions_search	-	-

Table 6: 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.

	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.
System:	How about rating?
User:	I want to eat at an expensive restaurant that's rated at least 2.
System:	Okay, I found 184 restaurants. I recommend the Seafood restaurant The Ocean by Olivier Bellin , which is rated 10.
User:	Can you find something near Eastern District?
System:	Yes, I found 1 European restaurant, Le 188° Restaurant & Lounge, rated at 9.
User:	Okay, and what's their phone and address?
System:	It's located at 41/F, Harbour Grand Hong Kong, MTR Fortress Hill Station, Exit A, 23
User:	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.
System:	Sure, I can do that. Can you tell me how many people will be there and which name I should
User:	reserve it under?
	Yes, book a table for twenty one under my name Ryan.
System:	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?
User:	Yes thats right
System:	Okay, it is reserved. Your reference number is 1GKOTNWV.
User:	Thanks, are there any Water & Amusement Parks nearby?
System:	Yes, there are 4, and I recommend Ocean Park with a rating of 9.
User:	Where is that, and what's their number?
Gold:	It's located at No.180 Wong Chuk Hang Road, Aberdeen and their number is +852 3923 2323.
Pred:	They are located at 7:20 pm on Le 188° Restaurant & Lounge.

	Dialogue State Tracking	
Intents	Gold Slot-Values	Pred Slot-Values
restaurants_search	`cuisine` = don't care `dietary_restrictions` = don't care `rating` > 2 `price_level` = expensive `location` = Eastern District `cuisine` = don't care `cuisine` = don't care	`cuisine` = don't care `dietary_restrictions` = don't care `rating` > 2 `price_level` = expensive `location` = Eastern District `cuisine` = don't care `cuisine` = don't care
restaurants_booking	`name` = Le 188° Restaurant & Lounge `time` = 7:20 pm `date` = Sunday `user_name` = Ryan `number_of_people` = twenty one	`time` = 7:20 pm `date` = Sunday `user_name` = Ryan `number_of_people` = twenty one
attractions_search	`type` = Water & Amusement Parks	`type` = Water & Amusement Parks `location` = **Eastern District**

Table 7: 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.