Information Extraction and Human-Robot Dialogue towards Real-life Tasks: A Baseline Study with the MobileCS Dataset

Hong Liu[1,3,†]**, Hao Peng**[1,3,†]**, Zhijian Ou**[1,3∗]**, Juanzi Li**[1,3]**, Yi Huang**[2,3]**, Junlan Feng**[2,3]

[1]Tsinghua University, Beijing, China

[2]China Mobile Research Institute, Beijing, China

[3]Tsinghua University-China Mobile Communications Group Co., Ltd. Joint Institute, Beijing, China

{liuhong21,peng-h21}@mails.tsinghua.edu.cn,
{ozj,lijuanzi}@tsinghua.edu.cn,
{huangyi,fengjunlan}@chinamobile.com

Abstract

Recently, there have merged a class of task-oriented dialogue (TOD) datasets collected through Wizard-of-Oz simulated games. However, the Wizard-of-Oz data are in fact simulated data and thus are fundamentally different from real-life conversations, which are more noisy and casual. Recently, the SereTOD challenge is organized and releases the MobileCS dataset, which consists of real-world dialog transcripts between real users and customer-service staffs from China Mobile. Based on the MobileCS dataset, the SereTOD challenge has two tasks, not only evaluating the construction of the dialogue system itself, but also examining information extraction from dialog transcripts, which is crucial for building the knowledge base for TOD. This paper mainly presents a baseline study of the two tasks with the MobileCS dataset. We introduce how the two baselines are constructed, the problems encountered, and the results. We anticipate that the baselines can facilitate exciting future research to build human-robot dialogue systems for real-life tasks.

1 Introduction

Building human-robot dialogue systems is an important research question not only for artificial intelligence applications but also for artificial intelligence itself. In the Turing test, if the human evaluator finds that human-robot dialogue and human-human dialogue are indistinguishable, the robot would be said to exhibit intelligent behaviour and pass the test (Turing, 1950). So presumably, the best strategy to build an intelligent dialogue system may be to train the system over a large amount of real human-to-human conversations to mimic human behaviors. This approach was once pursued and several human-human dialogue datasets have been released, such as the Twitter dataset (Ritter et al., 2010), the Reddit conversations (Schrad-

ing et al., 2015), and the Ubuntu technical support corpus (Lowe et al., 2015). It is argued in (Budzianowski et al., 2018) that the lack of grounding conversations onto an existing knowledge base (KB) limits the usability of the systems developed over these human-human dialogue datasets.

So a class of Wizard-of-Oz simulated games have emerged to collect human-human conversations (Wen et al., 2017b; El Asri et al., 2017; Budzianowski et al., 2018; Zhu et al., 2020; Quan et al., 2020), particularly for task-oriented dialogue (TOD) systems which help users accomplish specific goals such as finding restaurants or booking flights and usually require a task-related KB. In the Wizard-of-Oz set-up, through random sampling based on an ontology and a KB (both are pre-defined), a task template is created for each dialogue session between two crowd workers. One worker acts as the role of a user and the other performs the role of a clerk (i.e. the system side). In practice, multiple workers may contribute to one dialogue session. In this way, annotations of belief states and systems acts become easy, and grounding conversations onto the KB is realized.

However, dialogue data collected in the Wizard-of-Oz set-up are in fact simulated data and thus are fundamentally different from real-life conversations. During the Wizard-of-Oz collection, specific instructions (e.g., goal descriptions for the user side and task descriptions for the system side) are provided for crowd workers to follow. In contrast, real-life dialogues are more casual and free-style, without instructions. Even with some goals in mind, chit-chat or redundant turns are often exist in real-life conversations, e.g., asking for repeating or confirming key information. In some sense, we could say that real-life dialogues are more *noisy*. Moreover, spoken conversations in real-world have a distinct style with those well-written conversations and are full of extra noise from grammatical errors, influences or barge-ins (Kim et al., 2021).

For building dialogue systems that are more applicable to real-life tasks, real human-human dialogue datasets with grounding annotations to KBs are highly desirable.

Recently, the SereTOD challenge is organized (Ou et al., 2022) and releases a new human-human dialogue dataset, called the MobileCS (Mobile Customer Service) dataset. It consists of real-world dialog transcripts between real users and customer-service staffs from China Mobile. Based on the observation and analysis of those dialogue transcripts, a schema is summarized to our best[1], according to which about 10,000 dialogues are annotated with entities, attribute triples and speaker's intents for every turn. The annotated part of the MobileCS dataset is randomly split into a train, development and test set, which consists of 8975, 1025 and 962 dialogues, respectively.

Based on the MobileCS dataset, the SereTOD challenge not only evaluates the construction of the dialogue system itself (Task 2), but also examines information extraction from dialog transcripts (Task 1), which is crucial for building the KB. The MobileCS data are more noisy and challenging, as compared to previous Wizard-of-Oz data. It is non-trivial to establish baseline systems on such dataset. This paper mainly presents a baseline study of the two tasks with the MobileCS dataset. Two baseline systems are constructed for the two tasks respectively, which both are released as open source[2] and provided to the participating teams in the SereTOD challenge. We introduce how the two baselines are constructed, the problems encountered, and the results. The results clearly show the challenge for information extraction and human-robot dialogue, when trained and tested on real human-human data. We anticipate that the baselines can facilitate exciting future research to build human-robot dialogue systems for real-life tasks.

2 Related Work

2.1 Dialogue Datasets

According to Budzianowski et al. (2018), existing dialog datasets (whether task-oriented or not) can be grouped into three categories: machine-to-machine, human-to-machine, and human-to-human. The machine-to-machine datasets may en-

sure full coverage of all possible dialogue outcomes within a certain domain, but they do not consider noisy conditions in real life, which poses a risk of a mismatch between training data and real interactions. The human-to-machine datasets, however, depend on the provision of an existing working dialogue system, which limits the practicality of the datasets. The human-to-human datasets address the problems in the above two classes of datasets. However, previous human-to-human datasets lack knowledge base and explicit goal in the conversation, making that systems trained with these corpus struggle in generating consistent and diverse responses (Li et al., 2016).

It is non-trivial to collect a TOD dataset with knowledge base and user goals. Previous TOD datasets are either collected through Wizard-of-Oz simulated games (Wen et al., 2017b; El Asri et al., 2017; Budzianowski et al., 2018; Zhu et al., 2020; Quan et al., 2020), or collected by converting machine-generated outlines to natural languages using crowd workers (Shah et al., 2018; Rastogi et al., 2020; Lee et al., 2022). However, during the collection of these previous datasets, specific instructions are provided for crowd workers, which is different from real-life conversation scenarios and leads to a gap between collected data and dialogues in real-life. The MobileCS dataset, introduced in SereTOD Challenge, comes from real-world dialogue transcripts and represents a step advancing to remedy the above deficiencies.

2.2 Dialogue Information Extraction

Dialogue information extraction is the task of extracting structured information, e.g., entities and attributes, from dialogue transcripts. Different from the traditional information extraction in general domain text (Sarawagi et al., 2008; Li et al., 2020b; Han et al., 2020), dialogue transcirpts are more verbalized and loose with more irregular expressions and grammar errors. Previous works have explored how to extract user information (Catizone et al., 2010; Hirano et al., 2015; Wu et al., 2019), clinical information (Kannan et al., 2018; Peng et al., 2021), and relations between speakers and mentioned entities in dialogues (Yu et al., 2020; Jia et al., 2021). However, there is no previous work focusing on extracting information on real-world dialogue transcripts between real users and customer-service staffs. In the paper, we develop a modern dialogue information extraction baseline, based on the Mo-

---

bileCS dataset, which contains dialogue transcripts from China Mobile.

2.3 Task-oriented Dialogue System

The methodology for building TOD systems is gradually advancing from separate training of individual modules (Williams et al., 2016; Mrkšić et al., 2017; Dai et al., 2018) to the end-to-end (E2E) trainable approach (Wen et al., 2017a; Liu and Lane, 2017; Lei et al., 2018; Shu et al., 2019; Zhang et al., 2020; Gao et al., 2020; Zhang et al., 2020). In early E2E methods, the sequential turns of a dialog are modeled with LSTM-based backbones. Recently, self-attention based Transformer neural networks (Vaswani et al., 2017) have shown their superiority in capturing long-term dependencies over LSTM based networks. Transformer based pretrained language models (PLM), such as GPT2 (Radford et al., 2019) and T5 (Raffel et al., 2020), have been leveraged to build generative E2E TOD systems in the pretraining-and-finetuning paradigm, which have shown improved performance over LSTM-based ones. Examples include GPT2-based SimpleTOD (Hosseini-Asl et al., 2020), SOLOIST (Li et al., 2020a), AuGPT (Kulhánek et al., 2021) and UBAR (Yang et al., 2021), and T5-based PPTOD (Su et al., 2021) and MT-TOD (Lee, 2021), among others. However, these previous TOD systems are mainly examined on simulated data collected by crowd workers. It is not clear what the potential performance of the current methodology of building TOD systems is in real-life tasks. In this paper, we present our effort to answer this question, by developing a TOD system on the MobileCS data, which are from real-life customer-service.

3 MobileCS Dialogue Dataset

The MobileCS dialogue dataset contains 10,000 dialogue labeled by crowd-sourcing and around 90,000 unlabeled dialogues. For each dialogue turn, the annotations consist of entities, attribute triples, and speaker's intents within the scope of the schema. Another around 1000 dialogs are put aside as the test data. More detailed information about the MobileCS dataset can be found in the challenge description paper for the SereTOD challenge (Ou et al., 2022).

The two tasks defined over the MobileCS dataset for the SereTOD challenge require different annotations. For information extraction (Task 1), the annotations of entities and attribute triples are needed for training and evaluating the system. For TOD system construction (Task 2), user intents, system intents and a local knowledge base (*local KB*, which covers personal information and relevant public knowledge in a dialogue) are required. A *global KB*, which covers and fuses all public knowledge and all personal information in the MobileCS domain, is difficult to obtain during the research phase. Thus, the SereTOD challenge introduces the concept of a local KB, which could be viewed as being composed of the relevant local snapshots from the global KB for each dialog. The local KB is obtained automatically by integrating all the annotations of entities and attributes into a sequence of entities[3]. Besides, user goals are needed for evaluating the performance of TOD systems in Task 2. Similarly, user goals are obtained automatically by integrating user intents and all the entities and attributes mentioned by the user. The examples of local KB and user goal can be found in Listing 1 in the challenge description paper (Ou et al., 2022).

Data Quality The MobileCS data were annotated by two professional data labeling teams according to well documented guidelines as described in (Ou et al., 2022). Quality control was enforced by sampling the annotated data and performing crossing check of the annotations by the two teams. Nevertheless, there still exist annotation errors in such a large dataset. Some annotation errors can be corrected by rules. A typical example of errors is the granularity error of entity types. In the schema, entity types have inheritance relationships, for example, "main package" inherits from "package" and contains all its properties. Therefore, there are quite a few annotation confusions between parent types and child types in the data. To correct those type errors, the most fine-grained type for each entity was selected according to the attributes held by the entity. By combining the schema with manual rules, more annotation errors can be corrected. The updated MobileCS dataset is called v1.1, which is released in the SereTOD challenge and used in the experiments in this paper.

---

[3]An interesting future problem is to study the quality of the local KBs constructed in such a way and their influence on the performance of the dialogue system.

4 Tasks

4.1 Task 1: Information Extraction from Dialog Transcripts

Task 1 aims to extract structured information from real-world dialogue transcripts for constructing KB for the TOD system. This task consists of two sub-tasks: entity extraction and slot filling. The entity extraction task aims to extract entities, involving named entity recognition and entity coreference resolution. And the slot filling task aims to extract the attributes and values of entities, and the status of user accounts. Compared to the information extraction tasks on general domain texts, this task poses more challenges. First, dialogue transcripts are more verbalized, loose and noisy, which requires more robust models. Second, dialogue transcripts contain more pronouns and referents, some of which even span several rounds. This requires coreference resolution and long context modeling.

4.2 Task 2: Task-Oriented Dialog Systems

The basic task for the TOD system is, for each dialog turn, given the dialog history, the user utterance and the local KB, to predict the user intent, query the local KB and generate appropriate system intent and response according to the queried information. Compared with previous work, this task has the following characteristics. First, there is no global KB but only a local KB for each dialog, containing all the information in entity and attribute annotations and representing the unique information for each user, e.g., the user's package plan and remaining phone charges. Second, the user's constraints on entities are relatively simple, e.g., 38M data package, so the customer service system usually can confirm the entities that the user refers to in one dialogue turn, without the need of dialogue state accumulation.

5 Baseline Models

5.1 Task 1 Baseline

Task 1 involves two challenging sub-tasks: entity extraction and slot filling. Therefore, we design a pipeline method to extract information from dialogue transcripts. For entity extraction, the pipeline is two-step: named entity recognition and entity coreference resolution. For slot filling, the pipeline is also two-step: slot recognition and entity slot alignment. For each step, we first utilize a text encoder backbone to encode utterances and then a task-specific module to extract specific information based on the encoded representations of the utterances. In our experiments, we adopt three text encoders: LSTM (Lai et al., 2015), BERT (Kenton and Toutanova, 2019), and RoBERTa (Liu et al., 2019). The overall model architecture is illustrated in Figure 1. The hyper-parameters are shown in Table 1. The details of each step are as follows.

Named Entity Recognition First, we utilize a sequence labeling method to extract entity mentions in dialogue transcripts as in Yamada et al. (2020a). Specifically, after encoding utterances, we adopt conditional random field (Lafferty et al., 2001) on the top of hidden representations to label entity mentions from each utterance of the dialogue transcripts.

Entity Coreference Resolution After extracting entity mentions from dialogue transcripts, we utilize an entity coreference resolution method to group the mentions that refer to the same entity, as the local KB organizes knowledge in entity level instead of mention level. Specifically, after encoding dialogues, we adopt the dot product between the representation vectors of the two entity mentions as the metric to assess whether two mentions refer to the same entity. The representation vector of an entity mention is defined by the mean pooling of the representations of the tokens of the entity mention, as did in Yao et al. (2019). We then utilize the binary cross entropy loss as the objective to fine-tune the backbone encoders.

Slot Recognition Slot recognition aims to recognize slots from plain texts, regardless of which entity the slot belongs to. We utilize a sequence labeling method to recognize the slots, i.e., to label certain spans in the utterance as slots, which are the attributes of entities and the status of users. Specifically, we utilize the same model architecture as in Named Entity Recognition to label slots from each utterance of the dialogue transcripts.

Entity Slot Alignment To construct a local KB, the final procedure is to link slots to the corresponding entities. We formulate the task as a sequence classification task. Specifically, we highlight an entity and a slot using special markers and then encode the text to the contextual representation, which is inspired by Soares et al. (2019). We adopt a linear classification head to classify whether the
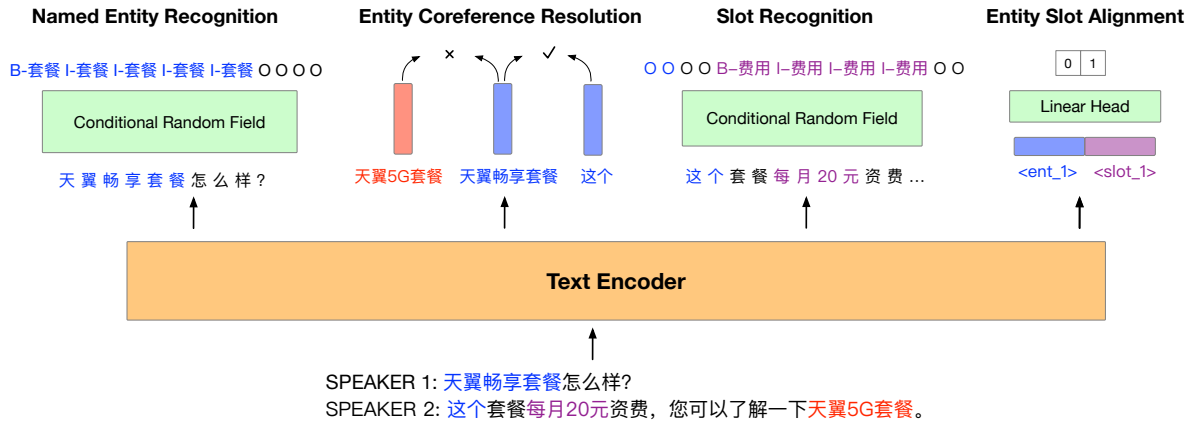
Figure 1: The overall model architecture of the pipeline model for Task 1. For the sub-task entity slot alignment, we utilize marker (e.g., `<ent_1>`*entity mention*`<\ent_1>`) to highlight entities and slots in the original text input.

Hyper-parameter	LSTM	PLMs
Learning Rate	1×10^{-3}	3×10^{-5}
Batch Size	64	64
Epoch	20	5

Table 1: Hyper-parameters of fine-tuning LSTM and PLMs (BERT, RoBERTa) on Track1 task.

slot corresponds to the entity.

5.2 Task 2 Baseline

KB Query We need to design a KB query function to help the TOD system access information from the local KB. After observing the dataset, we find that user queries can be divided into three different types. We encapsulate all query scenarios into one function and list their inputs (i.e. the arguments of the query function) and outputs as follows.

- Query the attribute of a specified entity. The input is the entity name and the attribute to be queried, the output is the attribute value in the local KB.

- Query entities of specified types. The input is entity type, the output is the entity names of this type.

- Query the attribute for users. The input is the attribute to be queried, the output is the queried attribute value in the local user profile (part of the local KB).

With the above query function, the TOD system can use the predicted user intent to access information from the local KB.

Baseline Architecture We divide the TOD system into several sub-tasks. For every dialog turn, the system needs to perform the following steps in order: 1) predict the entity name mentioned or referred to by the user; 2) predict the user intent (including the arguments of the query function); 3) query the local KB using the predicted user intent and obtain the KB result; 4) predict the system intent; 5) predict the system response. Note that there are many pronouns and co-references of entity names, so that the system may not be able to predict correct entity name with only the user utterance in current turn. To solve this problem, dialogue history information is needed. However, in real-life dialogues, the dialogue history is particularly long and contains plenty of characters, which will seriously hurt the training efficiency of the model (Liu et al., 2022). Therefore, we maintain a list of entity names mentioned by the user in all previous turns (entity name history) to replace the dialogue history. The entity name history and user utterance are fed into the model as the conditioning input to complete the above sub-tasks. Similar to Hosseini-Asl et al. (2020); Li et al. (2020a); Kulhánek et al. (2021); Yang et al. (2021); Liu et al. (2022), we employ a sequence generation architecture based on Chinese GPT-2 (Du, 2019) to implement the dialog system, which is depicted in Figure 2.

Data Analysis As described in Section 1, there are chit-chat or redundant turns in real-life dialogues. As observed from MobileCS, we find that these redundant turns can be divided into three cases: 1) one speaker asks for repetition and the other repeats what he/she said before; 2) one speaker confirms information and the other re-
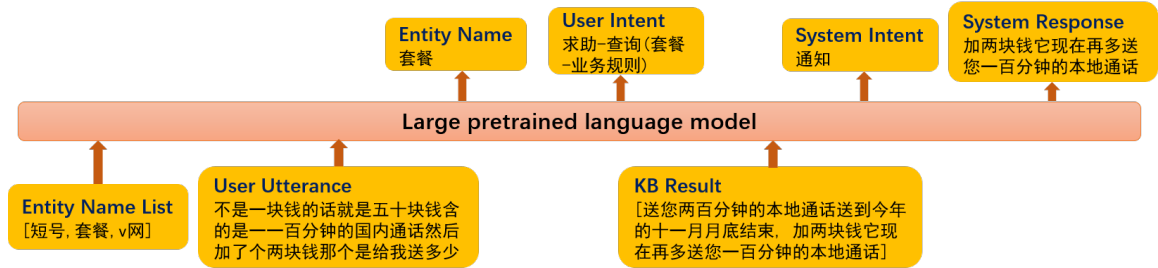
Figure 2: The baseline model architecture for Task 2. Examples are provided under the title of each box.

sponds to it passively; 3) the user interrupts the agent with some simple interjections, and then the agent continues to speak. Three examples corresponding to the three cases are shown in Table 2.

These redundant turns are interesting new phenomena revealed from the MobileCS data, which are transcribed from spoken conversations. Remarkably, the repetition and confirmation may be caused by that the staff did not hear clearly due to the accent or low quality of the user speech. The interjection is a special feature for spoken dialogues. However, after transcription of speech, the speech modality is missing, since only text is remained. Thus, the system in textual dialogues receives no relevant input from the user and is thus unable to respond properly. We leave further study of this problem for future work. In this work, we perform some pre-processing on the data to reduce the noise brought by the three cases. Specifically, for the first two cases, we simply delete the whole redundant turn (including utterances on both user and system sides) in the dialogue. For the third case, we merge the redundant turn with its previous turn by deleting the user utterance and merge the agent response with the previous one. Finally, we obtain a cleaned dataset with 15% fewer turns than the original one.

6 Evaluation

6.1 Task 1 Results

Metrics The evaluation metrics are two-fold. The metric for entity extraction is the span-level F1, following previous named entity recognition work (Yamada et al., 2020b). The metric for slot filling is the triple-level F1: a predicted *entity-slot-value* triple is correct if and only if the entity, slot and value are all correct. The evaluation for slot filling is a combinatorial optimization problem, as the entity is also predicted. We hence utilize the Hungarian algorithm (Kuhn, 1955) to find a best entity matching between predictions and golden

user	成那改成那最便宜那是打打那个长途是多少钱呢
system	呃呃您再说一下
user	我说改成那种你说那个便宜的是打打那个长途是多少钱一分钟呢
user	沈那中学里面
system	沈那中学是吗
user	对
system	三十八我看这边是流量送您六百兆通话送您两百分钟
user	嗯
system	前三个月每个月还送您二十块钱话费和一g的流量

Table 2: Examples of three types of redundant turns in MobileCS. The redundant utterances are marked in blue.

labels before calculating the metric for slot filling.

Results The models are trained on the training set for a certain number of epochs (as shown in Table 1), selected according to performance over the dev set, and evaluated on the official test set[4]. The results are shown in Table 3. It can be seen that even with powerful pre-trained language models as text encoders, the performance of the baseline model is poorer on the MobileCS dataset, especially for the named entity recognition and slot recognition sub-tasks, as compared to results on other datasets reported in the literature (Yamada et al., 2020a). These results demonstrate how demanding the MobileCS dataset is, and indicate that extracting structured information from long and loose texts, e.g. dialogue transcripts, remains challenging for existing models, which urges more powerful and robust models.

6.2 Task 2 Results

Metrics In order to measure the performance of TOD systems, both automatic evaluation and human evaluation are conducted. For automatic evaluation, metrics include **Precision/Recall/F1 score, Success rate and BLEU score**. P/R/F1 are calculated for both predicted user intents and system

---

[4]Notably, the challenge leaderboad for Track 1 are ranked by the results tested over 500 dialogues, which is only a subset of the official test set and were held out by the Challenge Organizers and not sent to the teams.

Backbone	F1 (NER)	Golden Labels			Pipeline
		B^3 (ECR)	F1 (SR)	Acc. (ESA)	F1 (SF)
LSTM (Lai et al., 2015)	35.02	85.84	43.89	76.84	31.37
BERT$_{\text{BASE}}$ (Kenton and Toutanova, 2019)	34.21	88.09	46.46	76.50	33.24
RoBERTa$_{\text{BASE}}$ (Liu et al., 2019)	33.74	88.02	45.59	77.32	33.28
RoBERTa$_{\text{LARGE}}$ (Liu et al., 2019)	**35.06**	**89.42**	**46.89**	**78.07**	**34.95**

Table 3: Experimental results of Task 1 on the official test set, with different text encoder backbones. "Golden Labels" means using golden prerequisite labels (e.g. golden entities for entity coreference resolution) for each pipeline step. "Pipeline" represents using previous predictions for each pipeline step. The evaluation metric is micro F1 for named entity recognition and entity slot alignment, B-cubed metric (Bagga and Baldwin, 1998) for entity coreference resolution, and accuracy for entity slot alignment. NER: Named Entity Recognition. ECR: Entity Coreference Resolution. SR: Slot Recognition. ESA: Entity Slot Alignment. SF: Slot Filling.

Dataset	U-P/R/F1	S-P/R/F1	BLEU	Success
Original	0.681/0.569/0.620	0.635/0.501/0.502	3.79	0.268
Cleaned	0.686/0.595/0.637	0.656/0.547/0.596	4.13	0.279

Table 4: The results of Task 2 baseline on the official dev set. U-P/R/F1 and S-P/R/F1 denote P/R/F1 for the user side and the system side, respectively.

Fluency	Coherency	Success
2.76	2.18	2.24

Table 5: Human evaluation of the Task 2 baseline system (trained on the cleaned dataset).

intents. Success rate is the percentage of generated dialogs that achieve user goals. Specifically, for each dialogue, we extract the information requested in the user goal from the local KB, then we regard this dialogue as a success if the generated responses contain all the requested information. BLEU score evaluates the fluency of generated responses by comparing them with oracle responses. For human evaluation, 5 testers (staffs from China Mobile) interacted with the system, and each tester should interact for at least 10 dialogues with the system. The tester would score the system on a 5-point scale (1 to 5) by the following 3 metrics. Success measures if the system successfully completes the user goal by interacting with the user. Coherency measures whether the system's response is logically coherent with the dialogue context. Fluency measures the fluency of the system's response.

Results Based on the analysis in Section 5.2, we conduct experiments on the original dataset and the cleaned dataset, respectively. The models are trained on the official training set for 40 epochs, and tested on the official dev set. The results are shown in Table 4. It can be seen that the model trained on the cleaned dataset outperforms the model trained on the original dataset in all metrics, which demonstrates the benefit of cleaning up redundant conversations. Nevertheless, the results on the cleaned MobileCS still fall behind by a large margin in comparison to the results on other

Wizard-of-Oz datasets. For example, the Success rate of state-of-the-art models on MultiWOZ2.1 is around 75%, while it is lower than 30% on MobileCS. The BLEU score on MobileCS is much lower than that on CrossWOZ (Liu et al., 2021). Note that both TOD systems on MobileCS and CrossWOZ are based on Chinese GPT-2, though not strictly comparable. These results demonstrate how challenging of building TOD systems for real-life tasks is. The agent responses from real-life are much more difficult to be modeled, as compared those in the Wizard-of-Oz scenarios.

We further perform human evaluation for the best baseline model (i.e. the model trained on the cleaned dataset) and the average scores of all tested dialogues are shown in Table 5. The scores of the three metrics are relatively low (lower than 3), which shows that in most cases, responses generated by the baseline system are neither fluent nor coherent enough, and can not provide requested information satisfactorily. In a word, building a TOD system that can perform well on real-life dialogues is very challenging, and there is much room for the baseline TOD system to be improved. The MobileCS dataset offers a valuable and challenging testbed for future research of building human-robot dialogue systems for real-life tasks.

7 Discussion and Conclusion

The performance of task-oriented dialogue systems on Wizard-of-Oz datasets have been improved con-

tinuously to a high level, for example, as shown in MultiWOZ[5]. However, Wizard-of-Oz dialogue data are in fact simulated data and thus are fundamentally different from real-life conversations, which are more noisy and casual. For further advancement of human-robot dialogue technology, real human-human dialogue data with grounding annotations to KBs are highly desirable. Further, noting that the KB is an indispensable part for TOD systems and usually is not readily available for real-life tasks, it is very important to investigate not only the dialogue system itself but also information extraction to construct the KB.

With the MobileCS dataset released by the Sere-TOD challenge, this paper presents a baseline study of both information extraction (Task 1) and human-robot dialogue (Task 2) over real human-human dialogue data. We introduce how the baselines for the two tasks are constructed, the problems encountered, and the results. It is found that the MobileCS dataset offers a challenging testbed for both tasks, with interesting open problems. Our baselines provide an easy entry to investigate the new dataset, and we anticipate that the baselines can facilitate exciting future research to build human-robot dialogue systems for real-life tasks.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Roberta Catizone, Alexiei Dingli, and Robert J. Gaizauskas. 2010. Using dialogue corpora to extend information extraction patterns for natural language understanding of dialogue. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Yinpei Dai, Zhijian Ou, Dawei Ren, and Pengfei Yu. 2018. Tracking of enriched dialog states for flexible conversational information access. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6139–6143. IEEE.

Zeyao Du. 2019. GPT2-chinese: Tools for training GPT2 model in chinese language. https://github.com/Morizeyao/GPT2-Chinese.

Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

S. Gao, Y. Zhang, Z. Ou, and Z. Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. *arXiv preprint arXiv:2004.03186*.

Toru Hirano, Nozomi Kobayashi, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. User information extraction for personalized dialogue systems. *SEMDIAL 2015 goDIAL*, page 67.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Qi Jia, Hongru Huang, and Kenny Q Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13125–13133.

Anjuli Kannan, Kai Chen, Diana Jaunzeikare, and Alvin Rajkomar. 2018. Semi-supervised learning for information extraction from dialogue. In *Interspeech*, pages 2077–2081.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*.

---

[5]https://github.com/budzianowski/multiwoz

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10938–10946.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Baolin Peng Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020a. SOLOIST: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics (TACL), 2021*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020b. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Bing Liu and Ian Lane. 2017. An end-to-end trainable neural network model with belief tracking for task-oriented dialog. *Proc. Interspeech 2017*, pages 2506–2510.

Hong Liu, Yucheng Cai, Zhenru Lin, Zhijian Ou, Yi Huang, and Junlan Feng. 2021. Variational latent-state GPT for semi-supervised task-oriented dialog systems. *arXiv preprint arXiv:2109.04314*.

Hong Liu, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. Revisiting markovian generative architectures for efficient task-oriented dialog systems. *ArXiv preprint arXiv:2204.06452*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhijian Ou, Junlan Feng, Juanzi Li, Yakun Li, Hong Liu, Hao Peng, Yi Huang, and Jiangjiang Zhao. 2022. A challenge on semi-supervised and reinforced task-oriented dialog systems. *arXiv preprint arXiv:2207.02657*.

Shuang Peng, Mengdi Zhou, Minghui Yang, Haitao Mi, Shaosheng Cao, Zujie Wen, Teng Xu, Hongbin Wang, and Lei Liu. 2021. A dialogue-based information extraction system for medical insurance assessment. *arXiv preprint arXiv:2107.05866*.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Alan Ritter, Colin Cherry, and William B. Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond W. Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *EMNLP*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Lei Shu, Piero Molino, Mahdi Namazifar, Hu Xu, Bing Liu, Huaixiu Zheng, and Gökhan Tür. 2019. Flexibly-structured model for task-oriented dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *CoRR*, abs/2109.14739.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 49:433–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017a. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017b. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Getting to know you: User attribute extraction from dialogues. *arXiv preprint arXiv:1908.04621*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020a. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020b. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6442–6454. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards fully end-to-end task-oriented dialog system with GPT-2. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. *arXiv preprint arXiv:2004.08056*.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*.