7-2021

MMConv: An environment for multimodal conversational search across multiple domains

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Le Hong LONG

Zheng ZHANG

Minlie HUANG

Tat-Seng CHUA

Citation

LIAO, Lizi; LONG, Le Hong; ZHANG, Zheng; HUANG, Minlie; and CHUA, Tat-Seng. MMConv: An environment for multimodal conversational search across multiple domains. (2021). *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11-15, Virtual*. 675-684.
Available at: https://ink.library.smu.edu.sg/sis_research/7286

MMConv: An Environment for Multimodal Conversational Search across Multiple Domains

Lizi Liao[1,2], Le Hong Long[2], Zheng Zhang[3], Minlie Huang[3], Tat-Seng Chua[2]
[1]Sea-NExT Joint Lab, Singapore
[2]School of Computing, National University of Singapore
[3]Department of Computer Science and Technology, Tsinghua University
liaolizi.llz@gmail.com,lehonglong@u.nus.edu,zhangz.goal@gmail.com,aihuang@tsinghua.edu.cn,dcscts@nus.edu.sg

ABSTRACT

Although conversational search has become a hot topic in both dialogue research and IR community, the real breakthrough has been limited by the scale and quality of datasets available. To address this fundamental obstacle, we introduce the Multimodal Multi-domain Conversational dataset (MMConv), a fully annotated collection of human-to-human role-playing dialogues spanning over multiple domains and tasks. The contribution is two-fold. First, beyond the task-oriented multimodal dialogues among user and agent pairs, dialogues are fully annotated with dialogue belief states and dialogue acts. More importantly, we create a relatively comprehensive environment for conducting multimodal conversational search with real user settings, structured venue database, annotated image repository as well as crowd-sourced knowledge database. A detailed description of the data collection procedure along with a summary of data structure and analysis is provided. Second, a set of benchmark results for dialogue state tracking, conversational recommendation, response generation as well as a unified model for multiple tasks are reported. We adopt the state-of-the-art methods for these tasks respectively to demonstrate the usability of the data, discuss limitations of current methods and set baselines for future studies.

CCS CONCEPTS

• **Computing methodologies → Intelligent agents**; **Artificial intelligence**.

KEYWORDS

datasets, multimodal dialogue, conversational search

ACM Reference Format:
Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462970

1 INTRODUCTION

The ever-increasing variety of information and products leads to information overload problem. Search and recommendation systems are developed to help people sift through information easier and make better decisions. The information search paradigm has been evolving from mostly unidirectional and text-based to interactive and multimodal [39]. Recently, there is also a growing interest in all matters conversational. By incorporating multimodal conversation, it offers users a natural way to query the system by combining information in various modalities. It also helps to tackle the basic asymmetric problem in search by injecting conversation to resolve ambiguities in search and recommendation.

However, the evolution from traditional IR to conversational IR faces many challenges. Among them are the need to develop new models and framework to: handle the ambiguity when understanding human language; model multimodal context and history, integrate domain knowledge and user models in decision-making; conduct interactive IR and QA, develop intervention strategy to incorporate conversation into search; and integrate conversation with third-party services such as recommendation, database search and Web search. Moreover, there are the issues of resources, methodologies and biases in evaluating (multi-turn) conversational search systems. These difficulties have pointed to the possible solution of using statistical framework and machine learning techniques. Therefore, inspired from the progress in dialogue research community, one may adapt and develop similar components, such as natural language understanding, dialogue management, language generation, and even end-to-end conversation modelling *etc.* However, the real breakthrough has largely been blocked by a comprehensive multimodal conversational search environment for facilitating the corresponding research tasks.

To drive the progress of building conversational search and recommendation systems using data-driven approaches, there are some corpora proposed recently. In general, existing corpora are either machine synthesized or collected via crowd-sourcing online. For example, [8, 32, 42] heavily rely on existing recommendation datasets and manually created templates to mimic conversations. Focusing on the recommendation part, these datasets lack the essential naturalness of conversations and oversimplify the conversation flow [17]. To make the interaction more realistic, there are also datasets such as [17, 21] that recruit crowd-sourced workers to interact in real-time under pre-defined search or recommendation settings. However, they either work on a single domain, rely on a single modality, or without pairing with any belief state and agent act annotations. None of them provide a comprehensive base to study various multimodal conversational search tasks.

SIGIR ’21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462970

Table 1: Comparison of our dataset MMConv to existing task-oriented dialogue datasets across domain, modality and tasks. 'Conv.' and ' Rec.' stand for 'conversational' and 'recommendation' respectively.

Datasets	# Dialogues	# Utters	Types	Domains	User Data	Modality	State Label
Facebook Rec [8]	1M	6M	Conv. Rec.	Movie	×	Text	×
REDIAL [17]	10K	163K	Conv. Rec.	Movie	×	Text	×
TG-ReDial [44]	10K	129K	Conv. Rec.	Movie	√	Text	×
OpenDialKG [23]	15K	143K	Conv. Rec.	Movie, book	×	Text	×
DuRecDial [21]	10K	156K	Conv. Rec.	Movie, music, news etc.	√	Text	×
MGConvRex [40]	7K	73K	Conv. Rec.	Restaurant	√	Text	√
WOZ 2.0[25]	1.2K	12K	Conv. Search	Restaurant	×	Text	√
DSTC2 [38]	1.6K	23K	Conv. Search	Restaurant	×	Text	√
FRAMES [9]	1.3K	20K	Conv. Search	Flight, hotel, budget	×	Text	√
KVRET [10]	3K	15K	Conv. Search	In-car assistant	×	Text	×
MultiWOZ [3]	8K	115K	Conv. Search	Hotel, restaurant etc.	×	Text	√
VisDial [5]	123K	2.4M	Image-based QAs	Concepts in image	×	Multi.	×
GuessWhat [6]	155K	1.6M	Image-based QAs	Concepts in image	×	Multi.	×
IGC [24]	4K	25K	Image-based QAs	Concepts in image	×	Multi.	×
MMD [29]	150K	6M	Fashion Search	Fashion	×	Multi.	×
MMConv	5.1K	39.7K	Conv. Search	5 domains in travel	√	Multi.	√

This paper introduces a Multimodal Multi-domain Conversational search (MMConv [1]) environment. It provides a large-scale multi-turn conversational corpus with dialogues spanning across several domains and modalities. Along which, there are also paired real user settings, structured venue database, annotated image repository as well as crowd-sourced knowledge database. More importantly, each dialogue is fully annotated with a sequence of dialogue belief states and corresponding system dialogue acts which is scarce in existing multimodal conversation corpora. Hence, MMConv can be used to develop individual system modules for conversational search following task-oriented dialogue research. On the other hand, with over 5k fully annotated dialogues, MMConv also enables researchers to carry on end-to-end conversational modelling experiments. Accordingly, we provide a set of bench-marking results using current SOTA methods for various tasks, which may facilitate a lot of exciting ongoing research in the area.

2 RELATED WORK

2.1 Dialogue Data Collection Paradigms

Based on different ways of collecting dialogue data, existing corpora can roughly be divided into three categories: machine synthesized, human-to-machine and human-to-human. At the very beginning, due to the interactive nature of conversation and the tremendous human labor required, many datasets are machine synthesized especially for those large-scale ones like [8, 29]. Such approach relies on simulated participants and exhaustive templates. Templates are then mapped to a natural language by either pre-defined rules [2] or crowd-sourced workers [31]. However, the dialogue flow is predefined, and it often does not take into account noisy conditions experienced in real interactions [1]. To improve the naturalness of dialogues collected while save human labor to some extend, many task-oriented dialogue corpora are fostered based on human-to-machine interactions. They rely on an existing dialogue system instead of collecting dialogue corpus from scratch. For example, the dataset for the first Dialogue State Tracking Challenge [37] is created via human machine interaction for live bus schedule information over the phone. Later, the second and third DSTCs [13, 38] have produced bootstrapped human-machine datasets for restaurant search. Although it seems to be a solution, it is only possible with an existing working system available.

The most direct and natural way is to collect human-to-human dialogues. Therefore, based on real user interaction logs on the Internet, several dialogue corpora such as [22, 28, 30] are released. However, these are mostly open-domain chit-chats, and the lack of an explicit goal limits their applicability in task-oriented scenarios. Another way to collect human-to-human data is to follow the Wizard-of-Oz framework (WOZ) [15]. One of the earliest trial is the ATIS corpus [12]. Recently, the original WOZ framework is modified to suit for crowd-sourcing. For example, Wen et al. [36] collected hundreds of dialogues via Amazon Mechanical Turk and later extended a second version WOZ 2.0 [25]. Similarly, El Asri et al. [9] collected the Frame corpus in a more complex travel booking domain. More recently, a larger MultiWOZ corpus spanning multiple domains became popular [3]. However, during collection of these corpora, in order to enable parallel collection and avoid the distracting latencies in conventional WOZ scenarios, users and wizards are asked to participate in multiple dialogues concurrently. They contribute just a single turn to each dialogue while need to review all previous turns contributed by others. It thus might hinder the coherence and quality of collected dialogues. On the contrary, our multimodal conversational search scenario involves multimodality data and requires frequent interactions with large back-end database. These largely undermine the reliability of leveraging similar WOZ setting as them. Hence, we resort to the human-to-human role playing paradigm where each dialogue is completed by a fixed pair of annotators. With no real-life multimodal conversational search data publicly available, we use real-world database and real user settings, together with strict quality control process to make our dialogue data as real as possible. Many studies [5, 21, 35, 40] have shown that this approach can be applied to collect high-quality conversations where a machine learning system can learn from.

[1] https://github.com/lizi-git/MMConv

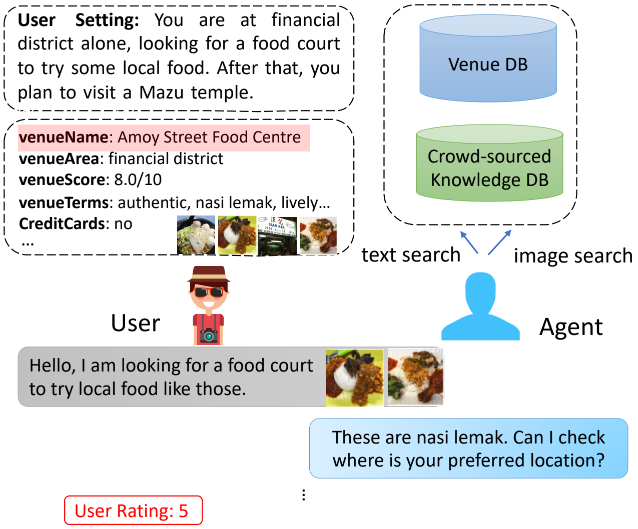
2.2 Across Modalities and Domains

As partly listed in Table 1, there are many multi-turn task-oriented conversational datasets contributed recently. We can clearly observe three trends: (1) a shift from pure text modality to cross modality; (2) an expansion from single domain to handling multiple domains at the same time; and (3) an emphasis on search and recommendation scenarios on task-oriented dialogue systems. Dialogue research starts from the natural language processing community. Correspondingly most of existing datasets are based on text modality. Recently, due to convenience of using visual modality in search and the rapid progress at the intersection of vision and language – in particular, in image captioning and visual question answering (VQA), there emerge multimodal dialogue datasets such as [5, 6, 24]. However, as pointed out in [18, 24], the problem setting for these works actually belongs to image-grounded QA. It is far from the task-oriented conversational search scenario which involves dynamical contexts such as to intersect with an external database to recommend restaurants. Consequently, Saha et al. [29] proposed a large scale multimodal conversational search dataset in fashion domain. However, it largely relies on utterance templates. More importantly, none of these multimodal dialogue datasets provide either dialogue state annotation or structured database. These are hard to support conversational search studies from various angles.

Dialogue analysis in early days tends to work on single domain setting with a small fixed ontology [37], as automatic speech recognition and spoken language understanding errors are common. However, handling tasks across different domains has become a more and more prominent requirement for building conversational agents [27]. Accordingly, datasets spanning multiple domains like [3, 9, 21] come into play. Here, we are concerned with five domains such as food and hotel that are closely related with each other under travel scenarios. This provides a good base for developing systems that are capable of handling multiple tasks at the same time.

3 DATA COLLECTION AND ANNOTATION

The MMConv dataset is collected by enabling multimodal conversations between human-to-human role-playing pairs under real life travel scenarios. We collect multi-domain conversations where an agent helps user to complete multiple tasks such as recommend venue or check reservation. As illustrated in Figure 1, the user needs to role-play a traveler under specific user setting. The target venue(s) information is provided but the user cannot expose unique information such as name, address or telephone number. To guide the conversation while allow flexibility, we provide structured preference list for user to express in each user setting. The agent observes the whole venue database and crowd-sourced knowledge database to find venues, provide recommendations and complete tasks. At the end of conversation, the user will give a feedback rating score to evaluate the agent's performance. To realize the whole collection process illustrated in Figure 1, we created our own databases, website and search engines. A total of 87 students are recruited in the data collection process. As discussed in Section 2.1, we apply the human-to-human role-playing scheme where one fixed pair completes the whole conversation. Detailed pre-collection training and various strict quality control checkup processes help to ensure high quality of the collected corpora.



User Setting: You are at financial district alone, looking for a food court to try some local food. After that, you plan to visit a Mazu temple.

venueName: Amoy Street Food Centre
venueArea: financial district
venueScore: 8.0/10
venueTerms: authentic, nasi lemak, lively...
CreditCards: no
...

Hello, I am looking for a food court to try local food like those.

These are nasi lemak. Can I check where is your preferred location?

User Rating: 5

Figure 1: The multimodal conversation collection setting. We collect 39.7K multimodal turns between pairs of human players. The user must follow the user setting and deliver venue information correctly; the agent must make use of the text & image search engines to respond and make appropriate venue recommendations for good user rating.

3.1 Task Setting

To the best of our knowledge, there is no real-life multimodal conversational search data publicly available. To make our data as close to real life settings as possible, we collect realistic databases, harvest real user settings from social sites, design various conversation scenarios that agent might encounter and explicitly define the targets of conversations. Together they ensure the appropriateness and naturalness of our overall task setting.

3.1.1 Real databases and user settings. We construct two real-world databases (venue DB and crowd-sourced knowledge DB) and an associated image repository. The venue DB is collected from Foursquare City Guide based on Singapore. We focus on five domains: food, hotel, nightlife, sightseeing and shopping mall, which are closely related to each other under the travel scenario. Initially, there are 14,671 venues obtained. Later filtering by rating scores, number of reviews (> 5) and images (> 10), we finalize 1,771 venues in our venue DB as in Table 2. For these venues, detailed information such as score, price range, location, whether have wifi or parking service *etc.* are included. Foursquare City Guide also provides abundant user ratings, reviews, posted photos for these venues [2]. The reviews are often commented based on existing user experiences, hence would offer detailed information such as popular dishes or dining environment. The posted photos largely come from professional or smart phone camera shooting in location, which gives direct feeling of the venue and serves as a convenient way of expression. We thus construct a crowd-sourced knowledge DB containing user reviews about these venues. Each review is associated with meta information such as number of up-votes. We also organize the

[2] Sensitive user information is removed from the data.

Table 2: Venue distribution over different domains.

Domain	food	hotel	nightlife	mall	sightseeing
# Venues	1,162	83	128	96	302
# Reviews	27,303	2,661	3,064	3,803	6,019
# Images	67,058	2,268	12,692	851	31,084

Table 3: Statistics of different conversation scenarios.

Scenarios	# conversations
Search venue by image	104
Recognize concept by image	214
Find venue by preferences	4,924
Cross-domain venue recommendation	3,007
Subsequent venue substitution	683
Venue comparison	182
Find specific shop in mall	53

images about these venues into an image repository and annotate them with image labels beside the venue association.

To collect real user settings, we crawl reviews and forum threads from tripadvisor.com and lonelyplanet.com. Venues appeared in the same thread are regarded as a venue bunch that can be used as targets in the same conversation, such as "Marina Bay Sands" and "Gardens by the Bay". The former is an integrated landmark resort while the later is a famous garden nearby. They are frequently co-visited by travelers to Singapore. To eliminate possible noises in crawled data, we do manual filtering by local undergraduate students and finally harvest 386 venue bunches for goal setting. Based on the contexts in review or forum threads for these venue bunches, we further enrich the goal settings with matching details such as "travel alone", "with kids" or "pub lover" *etc.*

3.1.2 Various conversation scenarios. We further analyze the contexts of user settings and harvest seven broad conversation scenarios as in Table 3. We then enrich them with more details. Generally speaking, we follow the rule of naturalness and utility of things. For instance, when some distinctive local food is associated with a venue such as "ayam buah keluak", "bak chor mee" or "chee cheong fun", we will apply the recognizing concept by image scenario to simulate user wanting to try local food without knowing the exact name. When the target venue got unique appearance, we set the searching venue by image scenario to provide a convenient way for search. If the target venues in a bunch are of same type with minor differences only, we will activate the subsequent venue substitution scenario to encourage user's goal change situation during conversation. At the end, there are seven broad scenarios captured in the corpus. The statistics are listed in Table 3. Note that one conversation may involve several different scenarios. For example, the agent may first recognize a dish by image and find the venue by uttered preferences, then do cross-domain venue recommendation to find another venue as illustrated in Figure 1.

3.1.3 Targets of conversations. Based on user setting and conversation scenario, we explicitly define structured preferences for user as additional target beyond the venue(s) target. In which we list attributes with preference level to inform, tasks to complete, and special settings such as provide image to get food name. It helps to regularize the conversation flow while allow expression flexibility. Beyond existing task oriented dialogues that over-emphasize on task completion, we also encourage to understand how people naturally express preferences. At the end of conversation, the user is required to rate agent's overall performance. The score ranges from 1 to 5, and the higher the better. The score 3 sets a baseline for successful conversations where venues are found and tasks are completed. The score 5 refers to successful, responsive, informative conversations that satisfy users the most. In our final corpus, only those dialogues with score 3 or higher are kept.

3.2 User Side

To facilitate conversation collection, three sections of information are provided to user: user setting, structured preferences and venue information. As illustrated above, the user setting sets a background for the conversation. In structured preferences, we list several preferred attribute values with preference levels for the user to express and assign several tasks such as "check outdoor seating" or "get phone number" to complete. For detailed venue information, it is closely related to domain slots structure. The domain slots of a task-oriented dialogue system is often defined by an ontology, a structured representation of the back-end database. The ontology defines all entity attributes called slots such as *venueArea* or *venueScore*, and all possible values for each slot. The general ontology structure is shown in Table 4. In general, the slots can be divided into *informable slots* and *requestable slots*. The *informable slots* are the attributes that the user can provide to agent for narrowing down the search space (*e.g. venueArea* or *price-range*). The *requestable slots* refer to some unique information such as *venueAddress* or *phone-number* that the user cannot expose. Beyond these fixed slots, we notice that users tend to mention terms such as "family friendly", "rooftop garden" or "mini bars" *etc.* to describe the target venues. These are hard to be grouped into any specific slot. We thus keep a special open span slot to contain all these salient terms. It allows our dialogue to be more flexible and rich in details. We also list some images for venues in the venue information section. The user can directly drag the image to input box as part of the utterance. There is no overlap between images in goals and in databases to avoid exact match.

3.3 Agent Side

The agent player responds freely to user based on conversation history and back-end databases. Considering the large amount of attributes, reviews and images associated with venues, it is unrealistic for agent player to go through all of them. We thus provide a text search engine and image search engine to the agent. The text search engine is built using Elasticsearch on the venue DB, while image search engine is built on ResNet-50 extracted image features. To make it easier for agent player, the text search results are further grouped by attributes and the image search results are ranked by similarity. All venue names appeared in the results can be clicked and detailed venue information will be shown in a pop-up page. Unlike the crowd-sourcing setting in MultiWOZ [3] where the agent and user first need to go through the dialogue history then contribute only one turn, the agent and user pair in our setting needs to complete the whole conversation and fulfill all pre-set tasks to ensure coherence and consistency of generated dialogues.

Table 4: Full ontology of all domains in our corpus. The upper script indicates which domain it belongs to. $\star$: universal, 1: food, 2: hotel, 3:nightlife, 4:mall, 5:sightseeing.

Action	inform / request/ recommend / negate / do not care/ confirm / show image/ greet / bye / others
Slots	drinks^{1,3} / music^{1,3} / reservations^{1,2,3,5} / dining options^{1,3} / stores^{4} / wifi* / menus^{1,2,3} / outdoor seating^{1,3} / venue domain* venue neighborhood* / wheelchair accessible ^{1,3}/ smoking ^{1}/ parking ^{1,3}/ restroom ^{1,2,3}/ credit cards * / pricerange ^{1,3}/ venuename* / venue score* / tips* / telephone* / venue address*

3.4 Annotation

The process of annotating belief states and dialogue acts is widely treated as the most challenging and time consuming part of dialogue data collection. It is usually done after the dialogue is generated as [3]. However, our preliminary trials show that asking the players to annotate their generated contents at the same time not only saves time for annotation but also provides a good base for quality checking. Therefore, our annotation of conversations is divided into four stages: during collection, during modification, slot mapping and manual correction. During the first stage, both user and agent players are asked to select out value for slots or salient terms (such as *expensive* or *cozy*) and pair them with intent acts (such as *inform* or *request* in Table 4). Then the generated dialogue with annotations will be manually checked for quality verification. If it does not pass, we will ask the pair to modify the dialogue and annotation until it meets our criteria. Then in the third stage, we organize these selected terms or phrases into slot value pairs with the help of venue database. After that, we further do manual correction to eliminate some arbitrary errors.

We also annotate all the images in our repository regardless of whether they are used in conversations or not. It would facilitate detailed multimodal research in various aspects. For this part, we combine automatic annotation with human annotation in an iterative fashion. EfficientNet [33] is leveraged to extract image features. We use K-Means clustering to cluster these images, manually select some obvious clusters out and assign label to them. We then try different parameter settings to do clustering and selection on the remaining images. Finally, 315 classes are generated on the image repository and the remaining noisy images are removed. We observe that a large portion of the removed images are Selfie images and menu photos.

3.5 Data Quality

To ensure the quality of collected conversation data, we apply the 5-step scheme of training, collection, checking, modification, and re-checking. Before the collection of data, we carry out training for all participants for about half an hour. When the dialogues are generated successfully during collection, we add a manual checking procedure. If a dialogue does not pass the checking, the participants will be notified and reasons will be given. Then, the participants need to modify the conversation and annotations. To save human labor, only one time modification is allowed and no pay will be credited to conversations failed to pass the last check.

Furthermore, to ensure the quality of annotations, we estimate the inter-annotator agreement for the last manual correction stage of dialogue act annotation. We calculate the averaged weighted kappa value [11] for all dialogue acts over 300 random sampled turns. The high score of $\kappa = 0.82$ demonstrates good agreements between annotators.

4 THE MMCONV CORPUS

4.1 Data Structure

The main goal of the data collection is to acquire highly natural conversations that cover a wide variety of styles and scenarios. In total, the presented corpus consists of five domains: *Food*, *Hotel*, *Nightlife*, *Shopping mall* and *Sightseeing*. Controlled by our various task settings, the collected dialogues cover between one to four domains per dialogue, and are thus of greatly varying length and complexity. There are 808 single-task dialogues that contains a single venue target and 4, 298 multi-task dialogues consisting of at least two to four venue targets. These different venues vary in domains most of the times. For ease of illustration, we name as single domain dialogues and multi-domain dialogues respectively.

According to the information modalities involved in dialogues, we can also group the dialogues into 751 single-modality dialogues and 4, 355 multi-modality dialogues. The corpus was randomly split on goals into a train, validation and test set to enforce reproducibility of results. Only those successful dialogues (all venues are found and tasks are completed) are included in our corpus. Each dialogue consists of a goal, multiple turn utterances as well as a set of belief states and dialogue acts with slots (values) annotations.

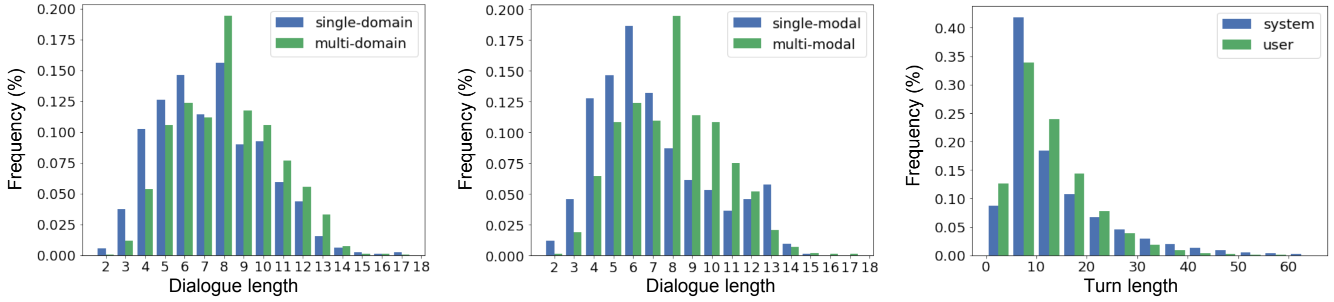
4.2 Data Statistics

Table 5: The general statistics of the MMConv corpus.

Entry	Number
# dialogues	5,106
# turns	39,759
# single domain v.s. multi-domain	808 v.s. 4,298
# single modality v.s. multi-modality	751 v.s. 4,355
# goals	386
# total venues in DB	1,771
# total images	113,953
# total reviews	42,850
# average user ratings	4.67

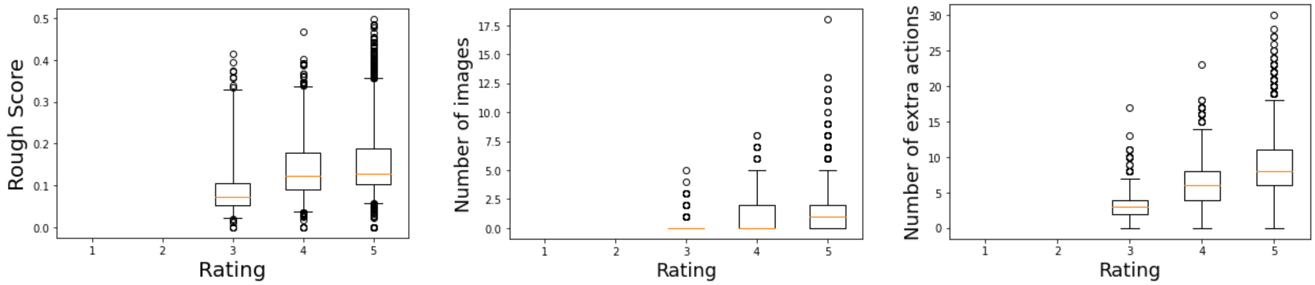
The general statistics of the MMConv corpus are listed in Table 5 [3]. Following data collection process from the previous section, a total of 5, 106 successful conversations were collected. Figure 2 (a) shows the distribution of dialogue length grouped by single and multiple domain dialogues. The average number of turns are 7.4 and 8.1 respectively. Figure 2 (b) presents the distribution of single and multi modality dialogues, in which the average turn numbers are 7.1 and 7.9 accordingly. Note that most of dialogues with multiple venue targets involve multiple modalities, thus multi-modal dialogues are more frequent than single-modal ones when the dialogue length exceeds seven. We also plot the distribution over turn length (*i.e.* number of tokens in turn) for user and system

[3] Only dialogues with user rating 3 or higher are kept in MMConv.



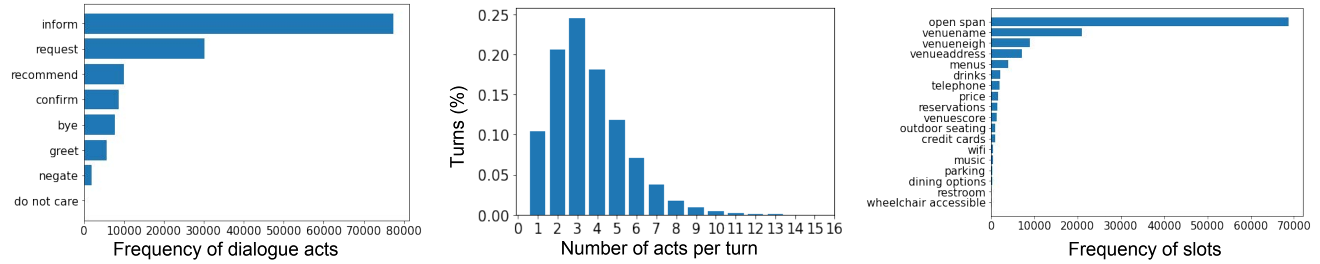
(a) Dialogue length distribution for domain (b) Dialogue length distribution for modality (c) Turn length distribution for user and system

Figure 2: Dialogue length (*number of turns in dialogue*) and turn length (*number of tokens in turn utterance*) distributions.



(a) Leveraging crowd-sourced tips (b) Making use of images (c) Performing extra dialogue acts

Figure 3: Correlations between user rating score and agent's behaviors.



(a) Dialogue act frequency (b) Distribution of dialogue act number per turn (c) Slot frequency

Figure 4: Various frequency statistics for dialogue acts and slots in the corpus.

in Figure 2 (c). The agent responses are much longer (13.7) than users utterances (12.0).

We believe that the more diverse agent responses would facilitate the training of more advanced generation models. As the user provides rating score for agent's responses at the end of each dialogue, we plot the correlations between user rating score and various agent behaviors in Figure 3. Specifically, Figure 3 (a) shows the correlation between user rating and the overlapping ratio of generated responses and crowd-sourced user review tips. As expected, when the agent makes good use of reviews of venues when making recommendation or suggestions to the user, the user tends to be more satisfied. Figure 3 (b) presents the correlation between user rating and the number of images involved in agent's responses. Intuitively, providing images to user would give them a more direct feeling of the recommended place thus improve user satisfaction. Moreover, we plot the correlation between user rating and agent's initiative in Figure 3 (c). Besides answering user's requests, the number of new dialogue acts generated by the agent reflects its activeness and involvement in the conversation. This also affects user ratings.

We group the statistical distributions of annotations in Figure 4. We first show the distribution of annotated dialogue acts of dialogues in Figure 4 (a). It is a summarized list of actions and corresponding frequencies. Different dialogue acts like *request(venueArea)* and *request(price-range)* for the same action *request* are grouped together. The Figure 4 (b) presents the distribution of number of acts per turn, where over 80% of dialogue turns have more than one dialogue act for either user or agent; this again shows the complexity and richness of the collected dialogues. As a corpora with more realistic and complex conversational search behaviours like these, MMConv creates new challenges for reinforcement learning-based algorithms as concurrent actions are common. Figure 4 (c) shows the distribution of slots annotated in the corpus. Due to the space limitation, only the most frequent slots are listed in the figure. It gives an overview of the information diversity during conversation procedure.

5 BENCHMARK RESULTS AND DISCUSSIONS

MMConv can serve as a benchmark resource for a range of conversational tasks. We implement several SOTA methods for various tasks and adapt to MMConv. We report results and discuss their limitations and the new challenges introduced by the MMConv.

5.1 Multimodal Dialogue State Tracking

As a bottleneck problem, dialogue state tracking (DST) interprets user goals and feeds downstream policy learning. Currently, multimodal DST is still in its infancy and the purely textual counterpart has been well researched [20, 43]. In general, they classify over fixed ontology or identify text span in utterances to extract or generate slot values. Since the ontology of our corpus contains both fixed slots and flexible salient terms, we adopt the state-of-the-art method DS-DST [41] from textual DST and make use of image information via predicted labels. As shown in Figure 5, it adapts a single BERT [7] question answering model to jointly handle both the categorical slots and non-categorical ones (*e.g.* flexible salient terms). For the categorical slots, it selects the most plausible values from the picklists based on the contextual representation (the left part). For the non-categorical slots, it utilizes a two-way linear mapping to find text spans (the right part). We train EfficientNet [33] separately for image label classification, then concatenate the predicted labels of images to its corresponding turn utterance.

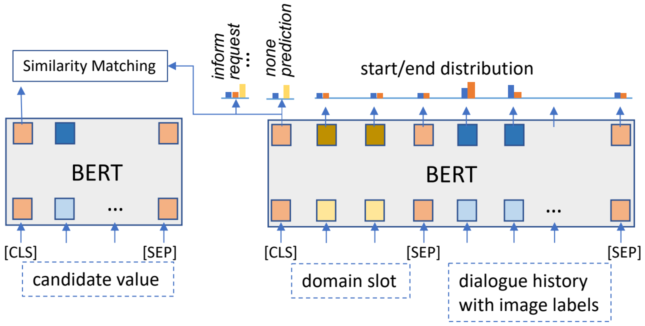


Figure 5: The adapted DS-DST model [41] to jointly handle categorical and non-categorical slots.

Table 6: The joint accuracy scores for multimodal dialogue state tracking using the adapted DS-DST model [41].

Datasets	Categorical	Non-categorical	Overall
WOZ 2.0	0.93	0.99	0.86
DSTC2	0.94	0.95	0.81
MultiWOZ 2.0	0.70	0.71	0.52
MultiWOZ 2.1	0.69	0.68	0.51
MMConv	0.91	0.23	0.18[1]

[1] Intent action is also correct for all slots.

We report the joint accuracy results over five datasets in Table 6. It measures the accuracy of the generated belief states as compared to oracle belief states. Model outputs are only counted as correct when all the predicted slot values exactly match the oracle values. The same model was trained across five datasets separately. Although not directly comparable, Table 6 shows that the model consistently performs worse on the new dataset than on the others. It shows that the new dataset is more demanding as the conversations are richer and more complicated. Specifically, there are two major observations. First, existing dialogue ontology over-simplifies conversational search scenarios. For example, WOZ 2.0 only contains three categorical slots: *area, price-range* and *food-type* for finding restaurants, and provides two non-categorical slots: *postcode* and *phone number* with obvious appearance patterns. Together with a limited number of value candidates for these slots, it results in virtual-high performance of 0.86 in joint accuracy. The DSTC2 dataset has more dialogues but the shared ontology with WOZ 2.0 is still rather simple. The later datasets MultiWOZ 2.0 and 2.1 provide rich ontology with over 20 slots across seven domains. Thus the DST performance drops dramatically (from around 0.8 to 0.5). Moreover, the intent action for slots are all *inform* for these datasets. However, in MMConv, we observe that users have different intentions, and tend to provide useful information such as 'good for date', 'great value' *etc.* which are hard to be grouped into slots but essential for finding venue. We therefore include action prediction, and organize these salient terms into the non-categorical part. The relatively low performance calls for more advanced mechanisms to handle them. Second, images in multimodal conversational search play important roles. For example, besides recognizing the exact food concepts from food images, it can further convey detailed information such as environment or be used to search for similar images thus narrow down the venue candidates. But current DST models largely lack good ways to handle images. These annotated datasets such as DSTC2 and MultiWOZ are all purely text-based, while the multimodal ones such as VisDial, IGC or MMD shown in Table 1 all lack annotations. Our fully annotated MMConv dataset provides a good base for carrying out such studies.

5.2 Recommendation in Conversational Search

Under the conversational search setting, given the belief state of dialogue history, the agent decides what to perform next – whether to request for more information, check a specific value or recommend a place (or an item). Most of existing conversational recommendation studies over-simplify the task [19]. They either work on simulated data [16, 32, 42], or simplify the slot value structure into one-hot attributes [4, 17]). Hence, we adopt the most recent state-of-the-art model UMGR [40] that is closest to our more complicated conversational setting. As shown in Figure 6, it introduces user memory graph to holistically represent the knowledge about users and associated items. Based on updating and reasoning over the graph, it predicts action for the agent first, and then rank slots, values or items based on the predicted action. In this way, policies that contain users/items unseen during training can also be generated.

The results are shown in Table 7. Act Accuracy is reported for all predicted dialogue acts. EMR stands for turn-level entity matching rate, which compares predicted entities like slots, values, venues against annotated ones when the dialogue act is predicted correctly; IMR stands for item matching rate, which evaluates the predicted venues against the ground-truth across all turns in a dialogue.

As expected, the same model performs badly on MMConv as compared to the other datasets. First, the act accuracy of MMConv is almost 1.74 times lower than that of MGConvRex. It is also inferior to that of TG-ReDial. This is mainly because the responses in MMConv often have multiple actions. We observe 34.6% of agent

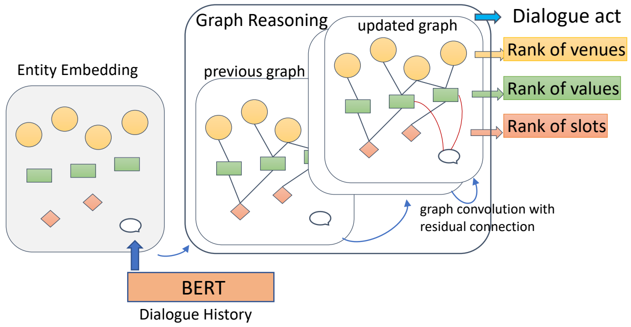


Figure 6: The adapted UMGR model [40] for conversation recommendation modeling.

Table 7: The results for conversational recommendation using the adapted UMGR model [40] across datasets. Images are transformed into predicted labels in text.

Metrics	MGConvRex	TG-ReDial	MMConv
Act Accuracy	65.70	32.65	23.96
EMR@1	33.92	19.45	10.01
EMR@3	48.47	22.32	11.80
EMR@5	52.54	26.49	15.91
IMR	67.93	17.60	9.58

responses containing more than one action, even when purely greeting or goodbye turns with single action are counted. However, most of the current datasets and models (even for UMGR) on conversational recommendation assume a single action per turn for agent [16, 32, 42]. Such simplification makes the learning easier but hinders the applicability of the developed methods. Second, the EMR and IMR on MMConv are also the lowest. It shows the complexity of our dataset. MGConvRex focuses on restaurant domain with only 10 slots and a total of 470+ values. However, MMConv covers five domains with 22 different slots and a much larger number of values. The number of venue targets (1,771) also exceeds that of MGConvRex. All these call for more advanced models to handle. Note that although TG-ReDial contains over 33K movies, it has about 2.5K topics. With 13.16 movies per topic on average, the topic representation largely narrows down the recommendation space.

Moreover, current researches on multimodal conversational recommendation modelling are rather limited. Most of current works recognize concepts from image first and then use the recognized concepts as text labels or vector representations [18]. We observe in MMConv that images are associated with concept labels as well as specific venues. Moreover, during conversation, users may express their intentions via images which also calls for better model to capture and model.

5.3 Response Generation

Generating appropriate responses for satisfactory task completion is the ultimate goal of task-oriented dialogue agents. Existing pipeline approaches generally predict multiple dialogue acts first and use them to assist response generation. To capture inherent structures of multi-domain dialogue acts and consider the semantic associations between acts and responses, the state-of-the-art model MARCO [34] generates dialogue acts and responses concurrently. As illustrated in Figure 8, by attending to different acts, the response generation module can dynamically capture salient acts and produce higher-quality responses. We adapt this model to our multimodal response generation scenario as follows. We first use EfficientNet [33] to transform images in dialogue history to textual labels and append it to corresponding turn utterances. The model takes in the transformed dialogue history and belief state as input and outputs agent actions and response. Beyond the existing dialogue acts for textual dialogues, we include actions regarding image modality such as "inform image: hotel room". When such action is predicted, we choose images associated with the specific venue regarding the predicted concept as the image response part.

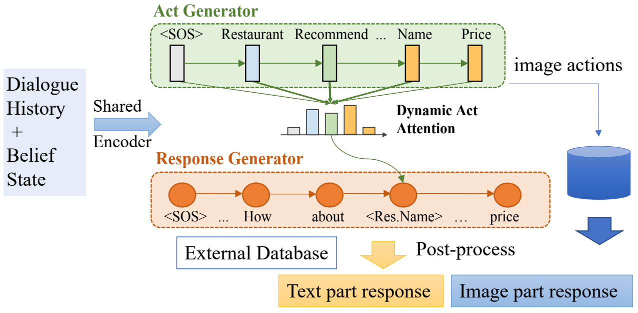


Figure 7: The adapted MARCO model [34].

Table 8: The results for response generation using the adapted MARCO model [34] across datasets.

Metrics	MultiWOZ 2.0	MultiWOZ 2.1	MMConv
Inform Rate	90.3	91.5	88.7[1]
Success Rate	75.2	76.1	82.4
BLEU	19.45	18.52	17.09
Combined	102.20	102.32	102.64
Image Match	–	–	0.17

[1] We evaluate via DB search on venue prediction turns as others.

Table 8 reports the results for the adapted MARCO model on three datasets. Image Match refers to the match rate of image concept predicted for correct venue. Since the MultiWOZ 2.0 and MultiWOZ 2.1 datasets do not contain image responses, we only report their results for the textual part. The inform rate and success rate capture how well the tasks are completed [14]. BLEU (4) score [26] measures the fluency of the generated responses. On one hand, we observe relatively good inform and success rate combination on MMConv. Similar to the evaluation on other textual datasets, we rely on pure text-based database queries. It actually inflated the performance as the returned result sets are large. Later we evaluate via directly comparing the predicted venue names to the ground truth venue names on the SimpleTOD [14] model as in Table 9, the performance drops dramatically. This actually signals a call for more appropriate way to evaluate multimodal responses. On the other hand, the BLEU score on MMConv is lower than that on others. It is probability due to the fact that the system responses in MMConv are more complicated. For example, we give more statistics for MultiWOZ 2.1: the responses in it contain 1.52 dialogue acts on average while those in MMConv contain 1.76 dialogue acts. Besides the information expressed via images, this indicates more complicated semantics expressed in MMConv responses. Also, as

Table 9: Results for DST, agent's act prediction and response generation by the adapted end-to-end SimpleTOD model [14].

Datasets	Joint Accuracy	Inform Rate	Success Rate	BLEU Score	Combined Score	Image Match
WOZ 2.0	0.81	77.2	68.8	18.79	91.79	–
MultiWOZ 2.0	0.57	84.4	70.1	15.01	92.26	–
MultiWOZ 2.1	0.56	85.0	70.5	15.23	92.98	–
MMConv	0.28[2]	14.6[1]	9.2[1]	20.30	32.20	0.02

[1] We evaluate on predicted agent's action results. At least one exact venue should be correct to be count as informed.

[2] Here we exclude the effect of flexible open span here .

shown in Figure 3, these responses further make use of external crowd-sourced knowledge, which makes it more rich in content. For the image response part, we apply a very simple and intuitive selecting mechanism. Since image responses have large potential in boosting user satisfaction as shown in Figure 3, we expect more advanced methods to be developed for this part.

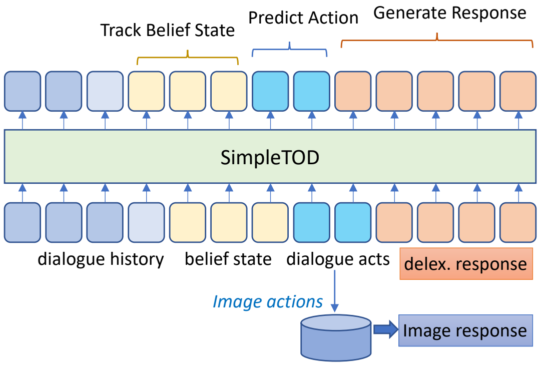


Figure 8: The adapted SimpleTOD model [14] for completing multiple tasks. It casts multiple tasks as a single sequence generation problem.

5.4 Joint Model across Multiple Tasks

Task-oriented dialogue is often decomposed into three tasks: understanding user inputs, deciding next actions, and generating responses. SimpleTOD [14] is a simple approach for all of them. It uses a single, causal language model trained on all sub-tasks recast as a single sequence prediction problem. It enables the modelling of inherent dependencies between sub-tasks, by optimizing all tasks in an end-to-end manner. We adapt it to the multimodal conversational modelling as shown in Figure 8. First of all, for images in dialogue history, we use the trained EfficientNet [33] to extract labels and append them to the original utterances. Then, the specific image slot and corresponding values are reflected in belief state and dialogue acts of agent, *e.g.* 'inform image: bar interior'. Later for response generation, besides generating textual responses sequentially, we also select images as response from the repository of the predicted venue regarding the generated image concept. Similar to the original SimpleTOD model, we delexicalize the agent response during both training and testing.

We compare the results on four datasets in Table 9. Regarding the dialogue state tracking sub-task, SimpleTOD outperforms the aforementioned DS-DST method over half of the datasets. Its superior performance derives mainly from the strong modelling and generation capability of the pre-trained GPT-2 model. For WOZ 2.0, as the ontology is simple, the mainly classification based counterpart works better. For the action prediction and response generation sub-tasks, the inform rate and success rate are related to the dialogue task completion. We observe dramatic drops regarding the inform rate and success rate on MMConv. From one aspect, MMConv contains a much larger number of target venues, slots and candidate values, which complicates the tasks. For example, there are 22 basic slots with over 6K candidate values in MMConv while MultiWOZ 2.1 only contains 19 basic slots with about 2K unique candidate values. And there are 1, 771 venues in the database of MMConv, while MultiWOZ contains only 224 venues. From another aspect, we calculate the inform rate and success rate for the other three datasets via database query results. While for MMConv, we evaluate by the exact venue name matching since it is predicted by the model. However, this exacting match condition is rather strict. Regarding the BLEU score, although MARCO reports lower score for MMConv as compared to MultiWOZ ones, SimpleTOD returns better score for it. This might be because the salient terms in belief states and actions of MMConv provide good signals for response generation and the GPT-2 model learns the patterns well. For image responses, the match rate is rather low. The main reason is due to the low inform rate. We only count as a match when both the venue name and image concept are predicted correctly. Our preliminary human evaluation show that images in responses help to improve user satisfaction as compared to the responses without it. However, the current way of incorporating it is rather simple and intuitive, more advanced methods can be investigated.

6 CONCLUSION

As multimodal conversational search is gaining more and more attention in both academia and industry, the necessity of building an entirely data-driven conversational agent becomes more apparent. Various corpora were gathered to enable data-driven approaches to conversation modelling. To date, however, the available datasets were usually constrained in linguistic variability, lacking multi-domain multi-modality use cases or unavailability of annotations. In this paper, we construct a relatively comprehensive environment for multimodal conversational search. Along with fully annotated dialogues, we also provide realistic user settings, structured venue database, crowd-sourced knowledge database as well as annotated image repository. We hope that MMConv would offer valuable training data and a new challenging test-bed for existing modular based approaches ranging from multimodal dialogue state tracking, conversational recommendation to response generation. Moreover, the scale of the data would help push forward research in the unified end-to-end multimodal conversational modelling.

ACKNOWLEDGMENT

This research is supported by the Sea-NExT Joint Lab.

REFERENCES

[1] Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *SIGDIAL*. 2–7.

[2] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683* (2016).

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*. 5016–5026.

[4] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *EMNLP-IJCNLP*. 1803–1813.

[5] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*. 326–335.

[6] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*. 5503–5512.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[8] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems.. In *ICLR*.

[9] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. In *SIGDIAL*.

[10] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *SIGDIAL*. 37–49.

[11] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[12] Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *SNL*.

[13] Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The third dialog state tracking challenge. In *SLT*. 324–329.

[14] 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* (2020).

[15] John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *TOIS* 2, 1 (1984), 26–41.

[16] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *SIGKDD*. 2073–2083.

[17] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *NeurIPS* (2018), 9725–9735.

[18] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *ACM MM*. 801–809.

[19] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2020. Topic-Guided Relational Conversational Recommender in Multiple Domains. *TKDE* (2020).

[20] Lizi Liao, Tongyao Zhu, Long Lehong, and Tat-Seng Chua. 2021. Multi-domain Dialogue State Tracking with Recursive Inference. In *The Web Conference*. To appear.

[21] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *ACL*. 1036–1049.

[22] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL*. 285–294.

[23] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*. 845–854.

[24] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In *IJCNLP*. 462–472.

[25] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *ACL*. 1777–1788.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[27] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604* (2018).

[28] Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*. 172–180.

[29] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[30] Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *EMNLP*. 2577–2583.

[31] 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* (2018).

[32] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *SIGIR*. 235–244.

[33] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*. 6105–6114.

[34] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-Domain Dialogue Acts and Response Co-Generation. In *ACL*. 7125–7134.

[35] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *ACL*. 5635–5649.

[36] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *EACL*. 438–449.

[37] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *SIGDIAL*. 404–413.

[38] Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The Dialog State Tracking Challenge Series. *AI Magazine* 35, 4 (2014).

[39] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Haitian Chen, Min Zhang, and Shaoping Ma. 2020. Preference-based Evaluation Metrics for Web Image Search. In *SIGIR*. 369–378.

[40] Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, and S Yu Philip. 2020. User Memory Reasoning for Conversational Recommendation. In *COLING*. 5288–5308.

[41] Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking. In *SEM*. 154–167.

[42] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *CIKM*. 177–186.

[43] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *The World Wide Web Conference*. 2401–2412.

[44] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *COLING*. 4128–4139.