MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce

Nolwenn Bernard
University of Stavanger
Stavanger, Norway
nolwenn.m.bernard@uis.no

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

ABSTRACT

Conversational systems can be particularly effective in supporting complex information seeking scenarios with evolving information needs. Finding the right products on an e-commerce platform is one such scenario, where a conversational agent would need to be able to provide search capabilities over the item catalog, understand and make recommendations based on the user's preferences, and answer a range of questions related to items and their usage. Yet, existing conversational datasets do not fully support the idea of mixing different conversational goals (i.e., search, recommendation, and question answering) and instead focus on a single goal. To address this, we introduce MG-ShopDial: a dataset of conversations mixing different goals in the domain of e-commerce. Specifically, we make the following contributions. First, we develop a coached human-human data collection protocol where each dialogue participant is given a set of instructions, instead of a specific script or answers to choose from. Second, we implement a data collection tool to facilitate the collection of multi-goal conversations via a web chat interface, using the above protocol. Third, we create the MG-ShopDial collection, which contains 64 high-quality dialogues with a total of 2,196 utterances for e-commerce scenarios of varying complexity. The dataset is additionally annotated with both intents and goals on the utterance level. Finally, we present an analysis of this dataset and identify multi-goal conversational patterns.

CCS CONCEPTS

• **Information systems → Search interfaces**; **Recommender systems**.

KEYWORDS

Conversational information access; Multi-goal conversations; Conversational dataset; Data collection methodology

ACM Reference Format:
Nolwenn Bernard and Krisztian Balog. 2023. MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591883

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591883

1 INTRODUCTION

How many queries would a user need to find a new pair of running shoes? One, if they knew exactly which one they wanted—however, it is not the case for the majority of users. Hence, they would most likely start an information-seeking process to explore the search space and execute several "context-free" queries to find the perfect pair of shoes [39]. For such situations, conversational systems have gained attention, as they offer several advantages over traditional means of information access. A dialogue is more natural, especially for complex information needs that might require a sequence of (co-dependent) queries. Indeed, the multi-turn structure of conversational interactions allows for making references easily to previous answers, which is not possible in a traditional search engine. Other advantages include more direct feedback, as it can be expressed in plain text, and better personalization capabilities thanks to the preferences disclosed in the dialogue [46]. Conversational information access is, in part, supported by the development of natural language processing and deep learning techniques [18, 47], and the growing adoption of conversational assistants [43]. However, advancements in the field are highly dependent on the availability of suitable conversational datasets.

Conversational information access systems support multiple conversational goals that are related to complex information seeking, exploratory information gathering, and recommendation [13]. In this work, we focus on the three main conversational goals identified in the field: *search*, *recommendation*, and *question answering (QA)* [46]. The distinction between these goals can sometimes be blurry as the same situation can be considered as search or recommendation (e.g., finding a close by hotel), or as search or question answering (e.g., agent answering a sequence of questions with passages) [46]. Crucially, in a natural conversation, the conversational goal is *dynamic*, i.e., it changes depending on the context. Therefore, a truly conversational information access system should support all these goals. Taking the example of a user looking to purchase running shoes, the agent starts by eliciting the user's preferences and makes recommendations based on them (i.e., *recommendation*). Before making a final choice, the user asks questions related to the eco-friendliness and the sole's characteristics of a suggested pair of shoes (i.e., *search* and *QA*). As the conversation progresses, the conversational goal can be updated several times based on the context. Following the example, the user might not be satisfied by the sole's characteristics of the first recommended pair of shoes and therefore asks for another recommendation. Hereinafter, we refer to similar conversations that mix goals as *multi-goal conversations*.

Most conversational datasets are created for a particular conversational goal, such as search (e.g., CAsT-19 [14] and MISC [42]) or recommendation (e.g., MultiWoZ [8] and INSPIRED [19]). Few datasets with multi-goal conversations exist, focusing on the domains of music, movie, restaurant, and news [15, 28, 29]. However, they do not support all three goals studied here. Hence, we create MG-ShopDial, a new dataset of English multi-goal conversations in the domain of e-commerce. According to Papenmeier et al. [34], e-commerce could take advantage of the conversational setting, especially for product search. Indeed, clients do not always know with precision what they are looking for, likely resulting in a multi-goal conversation (e.g., exploration of the space, disclosure of information need, clarification questions). Thus, MG-ShopDial is a resource that is particularly suited for the development of future conversational agents that can handle the evolution of a user's conversation goals as naturally as possible.

To collect the conversations, we use a coached human-human data collection protocol [37], where some participants mimic digital shopping assistants, while others play the clients. Each role comes with a set of instructions to detail what is expected with regards to conversational goals. However, we do not suggest answers or provide a specific script to follow. Instead, we emphasize on the naturalness of the conversation, the curiosity for the client, and that the shopping assistant should ask clarification questions, if necessary. In order to introduce diversity in the conversations, two types of scenarios (simple and complex) across different product categories (e.g., book, clothes, office supplies) are developed.

The creation of the dataset is facilitated by a purpose-built tool we developed to collect data in accordance with our protocol. Due to poor engagement from crowd workers, data collection is performed with volunteers to ensure that the dialogues are of high quality. As a result, MG-ShopDial contains 64 conversations with a total of 2,196 utterances. Upon analyzing MG-ShopDial, we observe a consistent conversational pattern that involves two or three distinct phases: first recommendation, then information seeking, and in some instances, a secondary recommendation. Notably, we find this pattern to be consistent across diverse scenarios and product categories. When comparing the goals in terms of associated intents, there does not appear to be a clear distinctions between *search* and *QA*, but we do observe a difference between those two goals and *recommendation*. These observations validate the usefulness of MG-ShopDial for research on multi-goal conversations.

In summary, the main contributions of this work are fourfold. First, we propose a protocol to collect realistic multi-goal conversations in the e-commerce domain. Second, we design and implement a tool to perform data collection following our protocol. Third, we collect and release a dataset of English multi-goal conversations. Finally, we present a concise analysis of the dataset. All resources developed as part of this study (MG-ShopDial, data collection tool, and annotation details) are made publicly available at: https://github.com/iai-group/MG-ShopDial.

2 RELATED WORK

We present previous work on conversational information access, data collection methodologies, and conversational datasets.

2.1 Conversational Information Access

Conversational information access systems, also referred to as *conversational information seeking* systems, are defined as agents capable of satisfying information needs through a conversation involving a sequence of interactions [46]. Conversational search, QA, and recommendation are regarded as subdomains of conversational information access, each with their own specificity. In the past few years, there has been continuous development on specific subtasks, such as query rewriting [27, 44], preference elicitation [24, 49], answer rewriting [4, 41], and user intent prediction [10, 36]. Some of these subtasks are particular to one of the subdomains mentioned above, like preference elicitation for recommendation and query rewriting for search. Hence, we notice that most studies focus on a single subdomain without considering how all these subtasks could be incorporated into a holistic conversational information access system. Our work aims to enable progress in that direction.

2.2 Data Collection Methodologies

There are various ways to collect conversational data. One approach is to get logs from an existing system, as was done, for example, for the Carnegie Mellon Communicator Corpus [7]. Another way is to perform user studies focusing on a task, typically using a crowdsourcing platform to recruit participants as in [8, 12, 48]. There are different ways to set up the crowdsourcing task, for example, Zhou et al. [48] ask crowd workers to edit utterances proposed by a neural model, while Hayati et al. [19] pair crowd workers to discuss about movies. The Wizard-of-Oz (WoZ) methodology is a popular approach to collect human-human dialogues. In this setting, there is a human intermediary acting as the conversational agent (i.e., the "wizard") and another human interacting with it not knowing that it is a human. Having a human intermediary tackles some practical limitations of conversational systems, especially regarding the understanding of natural language and tracking of conversation state. Still, some limitations remain, such as the dependence on the task studied (e.g., recommendation of hotels vs. question-answering about movies) and the tools available to the wizard [40]. To mitigate this dependence, Radlinski et al. [37] propose a new methodology derived from WoZ based on the idea of coaching the wizard rather than suggesting answers. For MG-ShopDial, we follow a similar approach, which is described in detail in Section 3.

There exist platforms to perform dialogue collection for different domains and tasks. Miller et al. [30] propose the ParlAI platform for dialogue research, with question answering and goal-oriented dialogue among the list of supported tasks, and the possibility to collect dialogues via Amazon Mechanical Turk. The CoCoA framework offers dialogue collection tools for three tasks (finding mutual friends, price negotiation, and deal or no deal) [20, 21]. Ogawa et al. [32] develop a platform based on video games to gamify dialogue collection. Their platform is presented as an alternative to crowdsourcing that has limitations to collect good quality dialogues (e.g., workers' motivation, costs). Notably, none of these tools support the task studied in this work, i.e., multi-goal conversations. Therefore, we propose our own dialogue collection application: Coached Conversation Collector (CCC), which is described in Section 4.

2.3 Conversational Datasets

Conversational information access has received a growing attention from the information retrieval, dialogue systems, and natural language processing communities. Thus, there is a large number of

Table 1: Comparison of conversational datasets selected. CS: conversational search, CR: conversational recommendation, CQA: conversational question answering, Meta: meta-communication.

Dataset	Language	Participants	CQA	CS	CR	Meta[†]	Domains	# conversations
Movie dialogue datasets [15]	EN		✓	×	✓	✓	Movies	~120,000 ... ~1M
MISC [42]	EN	Volunteers	×	✓	×	×	Open domain	88
QuAC [12]	EN	Crowd workers	✓	×	×	×	Open domain	13,594
SQuAD 2.0 [38]	EN	Crowd workers	✓	×	×	×	Open domain	151,054 (questions)
MultiWoZ [8]	EN	Crowd workers in a Wizard-of-Oz set up	×	×	✓	×	Restaurants, hotels, attractions, taxis, trains, hospitals, police	8,438
ReDial [26]	EN	Crowd workers	×	×	✓	✓	Movies	10,006
CAsT-19 [14]	EN	Experts	×	✓	×	×	Open domain	80
DoQA [11]	EN	Crowd workers in a Wizard-of-Oz set up	✓	×	×	×	Cooking, travel, movies	2,437
DuRecDial [29]	ZH	Unclear[‡]	✓	×	✓	✓	Movies, music, movie stars, food, restaurants, news, weather	10,190
TG-ReDial [48]	ZH	Crowd workers	×	×	✓	✓	Movies	10,000
INSPIRED [19]	EN	Crowd workers	×	×	✓	✓	Movies	1,001
DuRecDial 2.0 [28]	EN, ZH	Crowd workers	✓	×	✓	✓	Movies, music, movie stars, food, restaurants, news, weather	16,482
MG-ShopDial	EN	Volunteers	✓	✓	✓	✓	e-commerce	64

[†] Include sections of a conversation which do not contribute to the completion of the goal but to make the conversation fluid and natural.
[‡] Each conversation involves 2 persons, a seeker and a recommender, but their qualification is unclear.

conversational datasets available for various tasks and goals [1, 23].[1] These datasets can be classified based on the three conversational goals identified: *QA*, *search*, and *recommendation*. Note that in the context of conversational information access, we do not consider the social chat problem. Therefore, we analyze datasets of task-oriented and question answering conversations in regards to the above goals. To the best of our knowledge, none of the existing datasets contain all three goals. Table 1 compares MG-ShopDial with a selection of well-established and recent datasets. From the selection of datasets, only the Movie dialogue datasets [15] contain synthetic conversations. The majority of large human-human datasets use crowdsourcing [26, 38, 48], with the Wizard-of-Oz (WoZ) setup employed for some [8, 11]. Conversely, smaller datasets such as CAsT-19 [14] and MISC [42] are created with volunteers or experts. Regarding the target domain, we notice that most datasets for *recommendation* focus on movies [19, 26, 48], while datasets for *QA* and *search* tend to target more than a single domain [12, 14, 38]. Moreover, few datasets mix conversational goals. It is worth noting that meta-communication is becoming a valuable characteristic and is included in most recent datasets. Below, we briefly discuss about existing datasets mixing goals and motivate the need for the development of new ones that target multi-goal conversations.

2.3.1 Multi-goal Conversational Datasets. Dodge et al. [15] release five datasets that aim to test the abilities of end-to-end dialogue systems. Among these, only two, QA+Recommendation Dataset and Joint Dataset, contain dialogues that mix goals. Yet, the synthetic nature of the dialogues make them unrealistic. Liu et al. [29] release a first dataset mixing conversational goals, which they call dialogue types. This dataset aims to tackle their newly introduced task: *"conversational recommendation over multi-type dialogues, where the bots can proactively and naturally lead a conversation from a non-recommendation dialogue (e.g., QA) to a recommendation dialogue, taking into account user's interests and feedback"* [29]. However, the conversations in this dataset are in Chinese, which limits the scope of research. To tackle this, Liu et al. [28] propose a second version of the dataset including data both in Chinese and English. Furthermore, this second version supports multilingual and cross-lingual conversational recommendation research.

The closest datasets to our problem are [28, 29], but unlike us, these do not support conversational search. Also, instead of asking the crowd worker about their actual preferences, they provide a user profile with information such name, gender, occupation, and preferences. An approach to create a dataset mixing goals is to combine several datasets focusing on one goal is presented in [15], however, this approach is not considered in this work for the following reasons. First, it would require to have datasets with conversations in the e-commerce domain for the different goals. Second, the conversations would need to be altered in order to integrate the new goals, which would likely lead to a loss of naturalness. This motivates our work to create a new dataset with natural multi-goal conversations.

[1]List of conversational datasets by Joko et al. [23]: https://t.co/4315joogAk?amp=1

Table 2: Checklists for participants.

Shopping assistant		Client	
(A1)	Greetings	(C1)	Greetings
(A2)	Determine client's need (e.g., What are they looking for? Do they have any constraints?)	(C2)	Inform the retail assistant about what you are looking for
(A3)	Ask clarification questions if necessary	(C3)	Inform the retail assistant about your preferences
(A4)	Elicit client's preferences	(C4)	Ask factual questions about the recommended items
(A5)	Make a first recommendation of 3-4 items that fit client's need and preferences	(C5)	Ask general questions about the recommended items
(A6)	Answer factual questions asked by the user	(C6)	Based on previous utterances ask the retail assistant to refine the list of recommended items with new constraints
(A7)	Answer general questions asked by the user	(C7)	Choose an item to buy or express your dissatisfaction with the recommended items
(A8)	Refine the recommendation if the client's need or preferences change	(C8)	Gracefully end the conversation
(A9)	Gracefully end the conversation		

3 PROTOCOL

The goal of this paper is to collect multi-goal conversations in the domain of e-commerce using a coached (human-human) data collection protocol. For the collection of conversations, we follow the example of Radlinski et al. [37] by providing general instructions rather than possible answers to choose from. As in [37], this method can help to reduce the bias towards to the system, and permit human-level natural language understanding and generation. However, we customize our approach to a product search scenario, thereby differing from and extending [37] in several ways:

(1) Instructions are provided to both parties in the conversation: the shopping assistant and the client. Indeed, due to the complexity of the task and to ensure the presence of multiple conversational goals, both clients and assistants need to be coached.
(2) The instructions contain a *checklist* of actions to complete. This checklist allows the participants to easily and quickly assess what remains to be done during the conversation. Table 2 presents the checklists for the shopping assistant and client. To guarantee that the assistant can make a recommendation, they need to uncover the client's need and preferences; this is reflected in actions A3 and A4. Accordingly, the client needs to disclose this information as indicated in actions C2, C3, and C6. The client should also be curious and ask different types of question about the recommended items to explore the search space and make an informed decision (actions C4 and C5).
(3) The conversation has a time limit in order to help participants stay focused and minimize digression.
(4) Conversations are collected for selected product categories. In this work, we select 4 product categories from the Amazon Product dataset [31]: Sports and Outdoors, Books, Office Products, and Cell Phones and Accessories. The motivation behind this choice is that the categories are diverse and include products from daily life and hobbies.
(5) For each category, different scenarios with specific levels of *constraints* and *complexity* are developed. Constraints are divided into 3 levels: low, medium, and high. The low level corresponds to restriction only on the product, e.g., *"You are looking for a book in the genre of your choice."* For the medium level, another constraint is added on top of the product, for example a specific color or budget. Finally, when the scenario includes at least two constraints in addition to the product to buy, it is classified as a conversation with high constraints, such as *"You are looking for a pair of red running shoes in size 7 [...] made from recycled material."* Moreover, the conversation is considered *simple* when the client is only looking for a unique product, while it is seen as *complex* if more products are sought, e.g., a client is looking to buy the necessary equipment to play ice hockey. We refer to the GitHub repository for the detailed specifications of scenarios.[2]
(6) For each category, we provide a curated *list of products* that can meet the requirements presented in the different scenarios. The shopping assistant has access to the list of products to help them reduce the time needed to generate a response, as too long waiting time could negatively impact the user experience and engagement in the conversation. However, they are allowed to recommend products that are not in the list if they know a better match to the client's need.
(7) The participants can send text or image utterances. All contemporary e-commerce platforms have images in addition to the textual descriptions of items, thus, we believe that allowing image utterances is more realistic than just text.

Regarding (6), we note that the size or composition of the product list is not a significant factor in this study. Our primary objective is to gain insight into the structure and evolution of conversational goals, rather than to optimize product recommendations among the sea of choices on an e-commerce platform. Consequently, the shopping assistant need not focus on providing the best recommendation from an extensive list of products, but rather on recommending items that align with the client's needs and preferences.

We further note that, different from [37], ours is technically not a Wizard-of-Oz protocol, as the client is aware that the role of the shopping assistant is played by another human. This, however, could be changed by adjusting the instructions given to clients.

[2] https://github.com/iai-group/MG-ShopDial/blob/main/CCC/ccc/app/chat/static/yml/topics.yml

4 COACHED CONVERSATION COLLECTOR

We design and implement an application, Coached Conversation Collector (CCC), to facilitate data collection in accordance with the protocol described in the previous section. The goal of this application is to match shopping assistants with customers in chat rooms where conversations take place. In addition, the tool allows participants to keep track of their progress using their checklists. CCC is a modular application that can be adapted to other use cases, by changing the instructions or modifying the chat room interface. Below, we describe the main components of CCC (Section 4.1), followed by implementation details (Section 4.2).

4.1 Components

The application is divided into three main components: lobby, chat rooms, and administrator page. To access the application, users need to register first and specify their assigned role, i.e., shopping assistant or client. Shopping assistants also need to specify which categories they are interested in, the first time they log in.

4.1.1 Lobby. After logging in, participants are redirected to the lobby. There, they can see the chat rooms available (see the "Lobby" mocks in Figure 1). On the one hand, a shopping assistant only sees their chat rooms that correspond to the categories they selected. For example, if a shopping assistant is interested in Sports and Outdoors, then a room for this category is created and is made available. On the other hand, the client sees a list of all rooms with a color indicating their availability: an occupied room appears in red, while a free room is displayed in green.

4.1.2 Chat room. The interface in a chat room is divided into several elements and differs depending on the role of the participant. The *client interface* has, on top, a timer displaying the remaining time of the conversation. Below, the main panel is divided into two vertical blocks: the ongoing conversation is displayed on the left, while the instructions related to the task are shown on the right (see "Chat room - Client" in Figure 1). In addition to the above elements, the *shopping assistant interface* also includes a product list and access to a search engine (see "Chat room - Shopping assistant" in Figure 1). The product list is comprised of curated products with descriptions and pictures. Moreover, the shopping assistant can search a large web corpus on the paragraph level to answer information needs that they cannot answer with product information or from their own knowledge. This last element can be used to collect the search logs related to a conversation for further analysis. The advantage of this interface is its modular aspect. Indeed, the study leader can decide which elements of the interface are needed and can remove the unnecessary ones. For example, the instructions for each participant are defined in a HTML file that is easily editable.

4.1.3 Administrator page. The application also has a password protected page where the administrator can see the active users and the chat rooms opened, and access the recorded conversations. In the future, we plan to extend the administration interface with aggregate statistics over the collected conversations (e.g., average number of turn per conversation, number of conversations per categories).

Figure 1: Mock of the Coached Conversation Collector application, showing the different interfaces per roles.

4.2 Implementation

CCC is implemented in Python, based on the Flask framework[3] as webserver and using Redis[4] as a message broker and database. For this work, we index the TREC CAsT 2022 corpora[5] (MS MARCO V2 dataset [5], KILT Wikipedia [35], TREC Washington Post 2020[6]) for the internal web search engine. The product lists associated to each product category are curated manually from real products available on Amazon.[7] Regarding chat messages, participants can send utterances in different modalities: text or image via its URL. For each conversation, CCC stores the following metadata, in addition to the timestamped utterances: participants' checklists, search logs, and scenario information.

5 DATA COLLECTION

This section presents the data collection procedure (Section 5.1) and the participants involved (Section 5.2), followed by the annotation of conversations in terms of intent and goal (Section 5.3). Finally, we give a brief summary of the MG-ShopDial dataset (Section 5.4).

[3]https://flask.palletsprojects.com/en/2.2.x/
[4]https://redis.io
[5]https://github.com/daltonj/treccastweb
[6]https://trec.nist.gov/data/wapost/
[7]https://www.amazon.com

5.1 Procedure

The data was collected over the course of several sessions. The sessions were conducted in English, using two formats: in-person with approximately ten participants per session, and remote with a one-on-one format. At the beginning of each session, the study leader provided an introduction to the task and presented the instructions.

The shopping assistant can choose which category they are interested in (based on their familiarity with the topic) and join the associated chat room. Once the shopping assistant is in the room, the client can join as well. After joining the chat room, the shopping assistant and the client can read and follow the instructions (i.e., description of the task and actions checklist) associated with the task. It is important to note that the shopping assistant does not know beforehand what product the client is looking for, only the product category. This is by design to encourage them to ask questions to uncover the client's need.

The duration of the conversation takes into consideration the time needed to carefully read the instructions, the latency to get a reply, and the complexity of the task. Several experiments were conducted with different duration before settling on 17 minutes as the limit. We tried 13 and 15 minutes, but found those too short, as participants barely had the time to start the information seeking process. A duration set to 20 minutes, however, was slightly too long as participants finished sooner or diverged from the initial scenario. During the conversation, all utterances were stored as well as the search logs from the shopping assistant. It is possible that some conversations contain incorrect information, especially from the shopping assistant (e.g., wrong price, bargain campaigns), however it is not an issue for this work, since our interests lie in the discovery of conversational patterns in multi-goal conversations.

5.2 Participants

Initially, our plan was to conduct the data collection on a crowdsourcing platform. However, we experienced poor engagement from crowd workers. For example, some participants left the chat room because they did not get answers quickly enough. Others replied only with very short utterances that did not satisfy the requirements. Also, some workers did not follow the instructions and were chatting on unrelated topics. Therefore, we decided to perform data collection on a smaller scale with the help of volunteers, who were trained to perform the task, in order to collect better quality conversations.

In total, 21 volunteers participated in the data collection. Their recruitment was performed by the authors in their social and professional circles through word-of-mouth promotion. The data collection happened in several sessions, therefore some volunteers played both roles (i.e., shopping assistant and client). The choice of product categories for the participants playing the shopping assistant is based on their self-assessment of their knowledge about these categories. After the data collection, participants were asked to complete an anonymous demographic survey to determine their general characteristics. Four dimensions are considered in the form: (1) gender, (2) age, (3) education, (4) and the geographical origin of the participant's mother tongue. Table 3 presents an overview by dimension of the 17 answers collected. Seven females and ten males participated, with the majority being between 25 and 35 years old.

Table 3: Overview of census responses.

Dimension	Responses	
Gender	Female	41.2%
	Male	58.8%
Age	25-35	76.5%
	Over 35	23.5%
Education	MSc	47.1%
	PhD	47.1%
	Other	5.9%
Geographical continent of mother tongue	Africa	5.9%
	Asia	41.2%
	Europe	47.1%
	North America	5.9%

The answers show that the linguistic background of the volunteers is diverse: they are from four different continents, although Europe and Asia are predominant. We hypothesize that this diversity can be reflected in the conversations, in their way of using English.

5.3 Conversation Annotation

In order to gain insights into the evolution of conversational goals along with the structure of the conversation, we annotate all utterances in terms of intents and goals. For intent classification, we develop a schema based on previous work [3, 9, 34]. For conversational goals, we use the three goals (QA, search, and recommendation), plus a fourth category for meta-communication. The schemata for intent and goal annotations are presented in Table 4. The advantage of these schemata are that they are generic and domain independent.

5.3.1 Intents. The intent schema is based on a selection of communication functions from the international standard ISO 24617-2; these are domain independent and can help understand the participant's communicative behavior [9]. Similarly to Papenmeier et al. [34], we use only a limited number of generic intents from ISO 24617-2, as we do not know beforehand which intents will be present in the conversations. However, unlike them, we have several intents for inform, answer, and explain, as well as for the different types of questions such as clarification and preference elicitation [3]. In total, we select 12 diverse intents to characterize conversational patterns in the multi-goal conversations collected. Indeed, some of these intents relate to the revealment of information by the client or the shopping assistant (e.g., inform, explain), while others such as positive and negative feedback represent the participants' sentiments.

To validate the intent schema, we compute the inter-annotator agreement between two experts annotators who perform intent annotation on three conversations. As a multi-annotator agreement measure of a multi-label task, we compute the Fleiss' kappa metric κ [16] per intent that we average to get a global inter-annotator agreement, as in [8]. The agreement between the expert annotators is considered substantial (κ=0.633) [25], which provides validation.

Intent annotation is carried out by crowd workers; for each conversation, five workers annotate every utterance and the intents selected by at least two annotators are kept. On the annotation user interface, the worker is shown an example along with the intent description table. Then, the conversation to annotate is displayed in a tabular form: the first column lists the utterances, the second

Table 4: Schemata for intent (top) and goal annotation (bottom).

Intent	Description
Greetings	Indicates the beginning or end of the conversation
Interaction structuring	Utterances that make the conversation structured and natural (e.g., thanking, stalling)
Disclose	The client discloses information about what they are looking for
Clarification question	The agent asks a question to make sure it understands correctly a previous statement
Other question	Asks a question that is not a clarification question (e.g., factoid, follow-up questions)
Elicit preferences	The agent asks a question to find the client's preferences (e.g., the color of an item, the budget)
Recommend	The agent recommends one or several items to the client
Answer	A participant gives an answer to the other participant's information request
Explain	Provides an explanation to a previous statement (e.g., justifies suggestion or rejection of an item)
Positive feedback	Expresses positive feedback (e.g., confirmation, accept a recommendation)
Negative feedback	Expresses negative feedback (e.g., disagreement, rejection of a recommendation)
Other	Does not fit other labels
Conversational goal	**Description**
Search	The client wants to find more information on a product or specific topic. The agent answers the client's request for information. This can take form of casual (why/how), unanswerable, or complex questions that require multiple interactions (e.g., follow-up, sub-questions) and their answers.
Recommendation	The agent elicits the client's preferences. The agent makes a recommendation based on the client's need and preferences. The client discloses what they are looking for or their preferences intentionally or as answer to the agent's questions.
Question answering (QA)	A participant asks a factoid (what/when/who/where), confirmation (yes/no), or listing question about a product or specific topic. The other participant replies with a fact-based and short answer.
Meta-communication	Makes the conversation fluid and natural but is not necessary to complete the goal of the conversation (i.e., chit-chat).

Table 5: Inter-annotator agreement and proportion of utterances per intent. Utterances can have multiple intent labels.

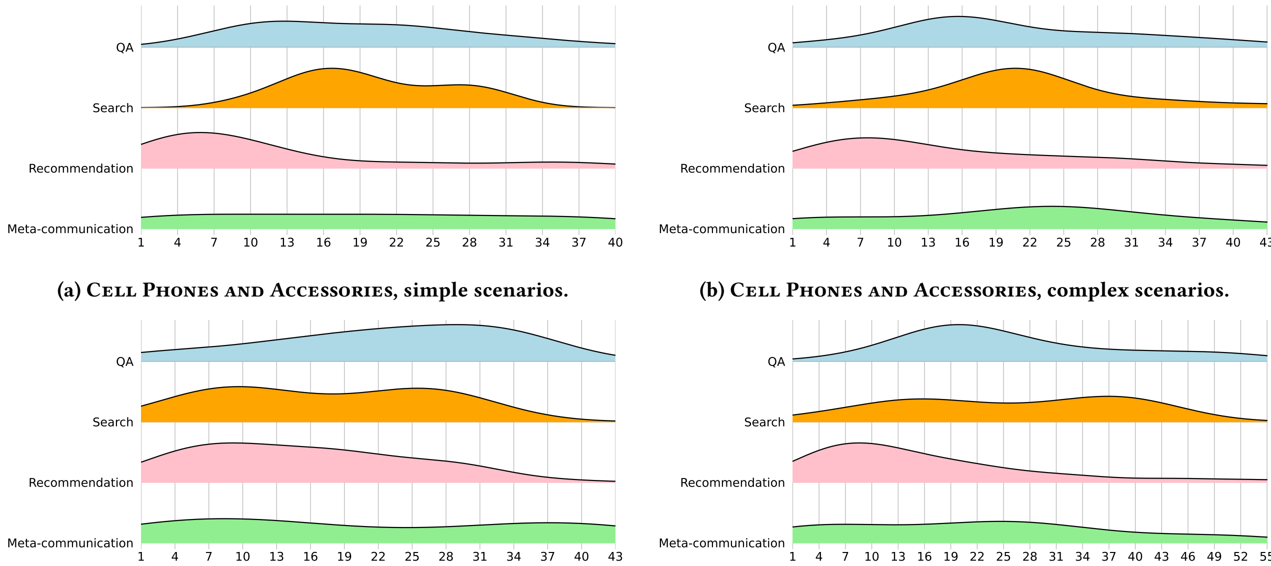
Intent	Fleiss Kappa	% utterances
Greetings	0.434	14.3
Interaction structuring	0.118	22.4
Disclose	0.132	12.8
Clarification question	0.159	23.4
Other question	0.233	20.9
Elicit preferences	0.085	11.2
Recommend	0.176	14.3
Answer	0.220	33.7
Explain	0.153	22.7
Positive feedback	0.155	16.5
Negative Feedback	0.259	2.7
Other	0.111	3.3
Weighted average	0.187	

shows the possible intents as checkboxes, and the last column has a text field where the annotator can optionally justify their choice.[8] Table 5 shows that the inter-annotator agreement is lower between crowd workers (κ=0.187) than between experts. This can be explained by several factors, including the complexity of linguistic annotation tasks, and the number of annotators and labels [2].

5.3.2 Goals. Recall that we distinguish between three conversational goals: *QA*, *search*, and *recommendation*. The delineation between the three conversational goals can be blurry, especially for *QA* and *search* [46]. In the e-commerce context, we can easily imagine a conversation mixing these goals. For example, when looking for a book, one might ask more or less complex questions to a shopping assistant about the author or literary movement of a suggested book. Furthermore, we add meta-communication to this schema. Indeed, some utterances might not contribute to the completion of one of the three previous goals but structure the conversation and make it fluid. Our motivation for annotating conversations with these goals is to observe whether some patterns emerge. For this annotation task, we provide a detailed description for each goal in Table 4. The description of *recommendation* is based on the definitions in [17, 22]; we emphasize on the preference elicitation element and the suggestion of products. The distinction between *QA* and *search* is based on the type of question a participant can ask during the conversation [45, 46]. In this work, we consider QA questions to be short and factual, hence we look for the following types of questions: (1) factoid, commonly starting with interrogative words like what, when, where, and who (e.g., "What is the price of this book?"); (2) confirmation (e.g., "Do you have it in blue?"); and (3) listing (e.g., "Can you give me the specifications of this phone?"). For *search*, we have: (1) casual questions usually starting by why or how (e.g., "How are these shoes environmentally friendly?"); (2) unanswerable questions; and (3) complex questions that require multiple interactions and can involve follow-up and subsequent questions (e.g., "What is the biggest difference between a beginner and a professional racquet? [...] Is there any difference in the string, like tension or wire thickness?").

Goal annotation is done for every conversation by the first author of this paper. A test sample with 25% of the conversations is also

[8] A screenshot of the annotation user interface is included in our GitHub repository.



(a) Cell Phones and Accessories, simple scenarios.

(b) Cell Phones and Accessories, complex scenarios.

(c) Office Products, simple scenarios.

(d) Office Products, complex scenarios.

Figure 2: Goal evolution for conversations about Cell Phones and Accessories (top) and Office Products (bottom).

Table 6: Breakdown of MG-ShopDial per category.

Category	#conversations	#utterances
Books	14	482
Office Products	14	458
Sports and Outdoors	17	697
Cell Phones and Accessories	19	559
Total	64	2,196

annotated by two crowd workers in order to ensure the clarity of the characterization of the goals. The crowdsourcing task is similar to the one for intent annotation, that is, an example is shown before starting the task, then the conversation is presented. However, instead of choosing one or multiple intents from a list, the crowd worker has to pick a single conversational goal. We compute the Fleiss' kappa metric κ [16] to assess the agreement between the annotators. Despite the complexity and subjectivity of this annotation task, the inter-annotator agreement is moderate (κ=0.415) [25], which provides validation for our schema.

5.4 Dataset Summary

The data collection procedure described above resulted in the creation of the MG-ShopDial collection, which contains a total of 64 conversations, comprising 2,196 utterances. The number of conversations within each product category are reported in Table 6. In general, we observe that the number of conversation per category is almost balanced. The complexity of the task is reflected by the number of utterances in a conversation: 75% of the conversations have at least 23 utterances. On average, conversations have 34.3±14.9 utterances that are each 7.5±6.2 words long.

We observe that some conversations did not reach the end, i.e., the selection of one or several products to buy. We hypothesize that some clients were more difficult to satisfy than others, requiring more utterances to understand their needs and elicit their preferences. However, the conversation is kept if it contains multiple goals. Furthermore, we notice that some conversations contain typos, grammatical errors, and emojis—these characteristics emphasize the naturalness of MG-ShopDial. The conversations do not include personal information; if names were mentioned, they have been anonymized.

6 ANALYSIS

We analyze the MG-ShopDial dataset with a focus on the evolution of conversational goals (Section 6.1) and the characterization of goals in terms of intents (Section 6.2).

6.1 Evolution of Conversational Goals

To identify conversational patterns, we study the evolution of conversational goals per category and scenario complexity. We make the following main observations across the different categories:

- The difference in conversation length between the simple and complex scenarios is small, even though one could reasonably expect that finding multiple items would require more turns.
- The goal distributions are similar, despite different scenario complexities.
- A common trend for all conversation is to start with *recommendation* followed by either *search* or *QA*, while having some *meta-communication* all along. More specifically, the first part of the conversation focuses on uncovering the client's needs and preferences to recommend some products. The second part mostly consists of product-related information seeking. In some cases, we observe a second peak in *recommendation* after the information seeking process, which might indicate that the client is not satisfied with the products and asks for other options.

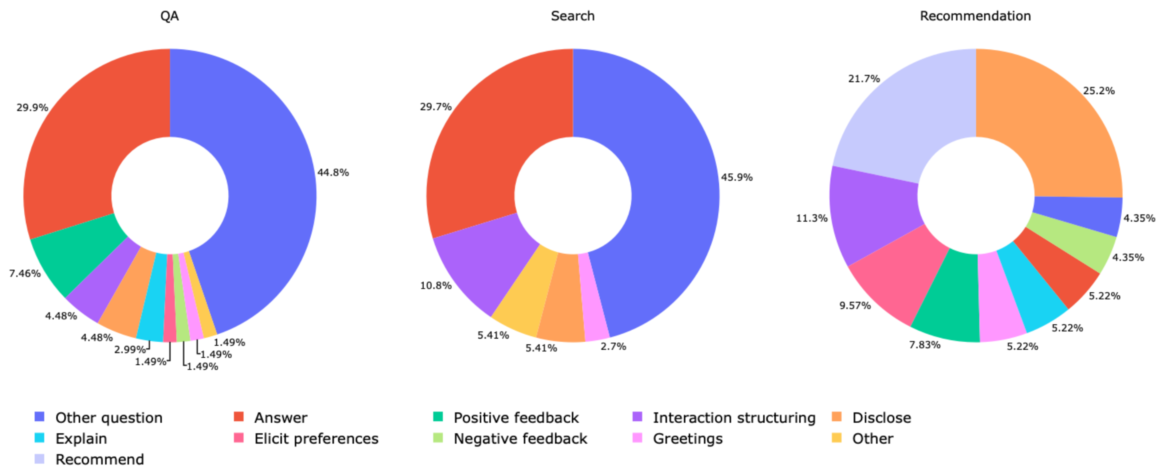


Figure 3: Intent distribution per conversational goal on a MG-ShopDial sample.

Figure 2 illustrates these observations by displaying the distribution of goals over the course of the conversation for Cell Phones and Accessories and Office Products, under simple and complex scenarios. The figure also shows the only exception to the main observations: simple scenarios for Office Products. There, the conversation length increases significantly for complex scenarios, and the goal distributions slightly deviate from the common trend. Indeed, the distributions are more balanced, yet we can still observe a decreasing probability for *recommendation* and the opposite for *QA* as the conversation progresses. Further, we note that *search* is present almost uniformly for most of the duration of the conversation. This might indicate that participants found this category more conducive to searching.

The different observations are consistent with some of the findings on product search by Papenmeier et al. [34]. Specifically, the almost uniform distribution of *meta-communication* illustrates the need of supporting interaction structuring (e.g., stalling, thanking). Further, the prevalence of *recommendation* over the first 10-15 utterances supports the idea of strategically uncovering and narrowing the client's needs and preferences.

6.2 Intent-based Characterization of Goals

Figure 3 shows the intent distribution per goal on a sample of conversations annotated with goals and intents by the main author (to ensure consistency and reduce potential noise from crowd annotations). For *QA* and *search*, *Answer* and *Other question* represent around 75% of the intents present, which is consistent with the idea of asking questions to get more information on products. Also, the *Recommend* intent is not present at all for these goals. As it can be expected for *recommendation*, the intents *Recommend* and *Disclose* represent the majority of the annotations (46.9%), followed by *Interaction structuring* (11.3%) and *Preference elicitation* (9.6%). These observations are in accord with the ambiguity around the different conversational goals, esp. with regards to *QA* and *search*. Indeed, *recommendation* can be easily distinguished from *QA* and *search* based on the intent distributions, while the distinction between the latter two is far less obvious. An idea for future work is to refine the *Other question* and *Answer* intents to see whether that would yield a better differentiation between *QA* and *search*.

7 CONCLUSION

In this work, we have introduced the MG-ShopDial dataset, along with the resources used to create it. Specifically, we have proposed a coached human-human protocol that emphasizes on guiding participants with the help of checklists instead of giving them a rigid script, and have developed the Coached Conversation Collector tool to perform the data collection following this protocol. The collected data has been annotated on the utterance level with both intents and conversational goals. Upon analyzing MG-ShopDial, we have observed a consistent conversational pattern that typically involved two or three distinct phases: initially, a recommendation is made, followed by information seeking, and in some instances, a secondary recommendation. We have also noticed that meta-communication is used throughout the conversation to keep it natural and to help transition between the different goals. Finally, the characterization of conversational goals in terms of intents has shown a clear distinction between *recommendation* and *search*/*QA*, but not so much between the latter two.

To the best of our knowledge, MG-ShopDial is the first dataset that mixes multiple conversational goals in a natural manner by situating participants in an e-commerce scenario. As such, it allows the development of conversational agents that support multiple goals. Nonetheless, the dataset is too small in size to train agents in an end-to-end manner. One solution would be to collect more data using our protocol and tool. However, creating a large collection with the same quality standards as ours is likely to be very time consuming and expensive. Another use of the dataset could be for few-shot learning with newer large language models, such as GPT-4 [33]. Alternatively, one could employ user simulation [6]; the collection is large enough to learn the parameters of models that can effectively characterize different scenarios.

ACKNOWLEDGMENTS

We thank all the volunteers and annotators who contributed to the creation of MG-ShopDial. This research was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, project number 309834).

REFERENCES

[1] Merav Allouch, Amos Azaria, and Rina Azoulay. 2021. Conversational Agents: Goals, Technologies, Vision and Challenges. *Sensors* 21, 24 (2021).

[2] Ron Artstein. 2017. *Inter-annotator Agreement.* Springer Netherlands, 297–313.

[3] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *SIGIR 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR '18)*.

[4] Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent Response Generation for Conversational Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 191–207.

[5] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[6] Krisztian Balog. 2021. Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation. In *Proceedings of the 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES '21)*. 80–90.

[7] Christina Bennett and Alexander I. Rudnicky. 2002. The Carnegie Mellon Communicator Corpus. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP '02)*. 341–344.

[8] 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 (EMNLP '18)*. 5016–5026.

[9] Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the ISO 24617-2 standard. In *Multimodal interaction with W3C standards*. 109–135.

[10] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. 33–42.

[11] Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 7302–7314.

[12] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. 2174–2184.

[13] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34––90.

[14] Jeff Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1985–1988.

[15] 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 *4th International Conference on Learning Representations (ICLR '16)*.

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

[17] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126.

[18] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. arXiv:2201.05176 [cs.IR]

[19] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*. 8142–8152.

[20] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '17)*. 1766–1776.

[21] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. 2333–2343.

[22] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (2021), 1–36.

[23] Hideaki Joko, Faegheh Hasibi, Krisztian Balog, and Arjen P. de Vries. 2021. Conversational Entity Linking: Problem Definition and Datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 2390–2397.

[24] Ivica Kostric, Krisztian Balog, and Filip Radlinski. 2021. Soliciting User Preferences in Conversational Recommender Systems via Usage-Related Questions. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21)*. 724–729. https://doi.org/10.1145/3460231.3478861

[25] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[26] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS '18)*. 9748–9758.

[27] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[28] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*. 4335–4347.

[29] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 1036–1049.

[30] Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. arXiv:1705.06476 [cs.CL]

[31] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP '19)*. 188–197.

[32] Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. 2020. Gamification Platform for Collecting Task-oriented Dialogue Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC '20)*. 7084–7093.

[33] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[34] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..." – Conversational Strategies For Product Search Assistants. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. 36–46.

[35] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL '21)*. 2523–2544.

[36] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User Intent Prediction in Information-Seeking Conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. 25–33.

[37] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL '19)*. 353–360.

[38] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL '18)*. 784–789.

[39] Tony Russell-Rose and Tyler Tate. 2013. *Designing the Search Experience.* Morgan Kaufmann.

[40] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version. *Dialogue & Discourse* 9, 1 (2018), 1–49.

[41] Idan Szpektor, Deborah Cohen, Gal Elidan, Michael Fink, Avinatan Hassidim, Orgad Keller, Sayali Kulkarni, Eran Ofek, Sagie Pudinsky, Asaf Revach, Shimi Salant, and Yossi Matias. 2020. Dynamic Composition for Conversational Domain Exploration. In *Proceedings of The Web Conference 2020 (WWW '20)*. 872–883.

[42] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proceedings of the 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR '17)*.

[43] Business Wire Voicebot.ai. 2020. Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)*. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/ Accessed: 2022-02-16.

[44] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1933–1936.

[45] Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: a survey. *Knowledge and Information Systems* 64, 12 (2022), 3151–3195.
[46] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. arXiv:2201.08808 [cs.IR]
[47] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* 63, 10 (2020), 2011–2027.
[48] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING '20)*. 4128–4139.
[49] Jinhang Zuo, Songwen Hu, Tong Yu, Shuai Li, Handong Zhao, and Carlee Joe-Wong. 2022. Hierarchical Conversational Preference Elicitation with Bandit Feedback. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. 2827–2836.