A Conversation is Worth A Thousand Recommendations: A Survey of Holistic Conversational Recommender Systems

Chuang Li[12,*], Hengchang Hu[1], Yan Zhang[1], Min-Yen Kan[1] and Haizhou Li[13]

[1]National University of Singapore, Singapore
[2]NUS Graduate School for Integrative Sciences and Engineering, Singapore
[3]The Chinese University of Hong Kong, Shenzhen, China

Abstract

Conversational recommender systems (CRS) generate recommendations through an interactive process. However, not all CRS approaches use human conversations as their source of interaction data; the majority of prior CRS work simulates interactions by exchanging entity-level information. As a result, claims of prior CRS work do not generalise to real-world settings where conversations take unexpected turns, or where conversational and intent understanding is not perfect. To tackle this challenge, the research community has started to examine *holistic CRS*, which are trained using conversational data collected from real-world scenarios. Despite their emergence, such holistic approaches are under-explored.

We present a comprehensive survey of holistic CRS methods by summarizing the literature in a structured manner. Our survey recognises holistic CRS approaches as having three components: 1) a backbone language model, the optional use of 2) external knowledge, and/or 3) external guidance. We also give a detailed analysis of CRS datasets and evaluation methods in real application scenarios. We offer our insight as to the current challenges of holistic CRS and possible future trends.

Keywords

Conversational Recommender Systems, Recommender Dialogue Systems, Interactive Dialogue Systems, Survey

1. Introduction

Conversational Recommender Systems (CRS) integrate conversational and recommendation system technologies, to facilitate users in achieving recommendation-related goals through conversational interactions [1]. In contrast to traditional recommendation systems, which act in a single (one-shot) round of interaction, CRS support multiple rounds of interaction, allowing the system to make multiple attempts in recommendation.

In much prior work on CRS, the multiple rounds of interaction are simulated by entity-level interaction, consisting of a sequence of entity-level features [2, 3]. For example in Figure 1(a), the entity-level interaction process is illustrated by how the system selects the "Feature ID" of *<Genre-Disney>* from its feature list, and the simulated human response of *<Yes>* will be directly returned to the system. Such a framing of the CRS task focuses on recommendation and decision-making strategies, which neglect the conversational element, such as possible inaccuracies in understanding the human language that makes up the conversation. Inaccurate conversation comprehension, gauging of intent and incorrect response generation [4, 5] as well as information inconsistency [6] are a regular occurrence in human conversation, yet much research on CRS have simply abstracted away from these defining characteristics. This is due to its presumption that the entity-level interaction is invariably accurate [3]. As a result, the application and evaluation of such systems in real-world situations pose significant challenges.

Thus there is a dichotomy in CRS research. Most CRS do not assume actual human conversations for interaction, only simulating the interaction with entity-level information [7, 3]. However, there are also prior work that relax this constraint and tackle conversational recommendation based on actual human conversations [8, 9]. Besides recommendation and decision strategy, these works also tackle the aforementioned conversational challenges in language understanding, generation, topic/goal planning and knowledge engagement. To distinguish these two forms of CRS research, we divide the current research works in CRS into ***standard CRS*** (the former, more prevalent form of prior CRS work), and what we term ***holistic CRS*** (which assumes a wider scoping of the CRS task) based on the input and output formats, as shown in Figure 3.

Research on holistic CRS is burgeoning, and it is timely to comprehensively survey such works to better organise and make sense of their contributions and gauge their potential future directions. This is needed to effectively utilize holistic CRS and the conversational datasets collected from real-world scenarios [10, 8] that train them, in prac-

Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, September 18–22 2023, Singapore.
*Corresponding author.
lichuang@u.nus.edu (C. Li); hengchang.hu@u.nus.edu (H. Hu); eleyanz@nus.edu.sg (Y. Zhang); kanmy@comp.nus.edu.sg (M. Kan); haizhou.li@nus.edu.sg (H. Li)
https://github.com/lichuangnus/CRS-Paper-List (C. Li)
0009-0006-8112-3505 (C. Li); 0000-0001-7847-0641 (H. Hu); 0000-0002-5336-7100 (Y. Zhang); 0000-0001-8507-3716 (M. Kan); 0000-0001-9158-9401 (H. Li)

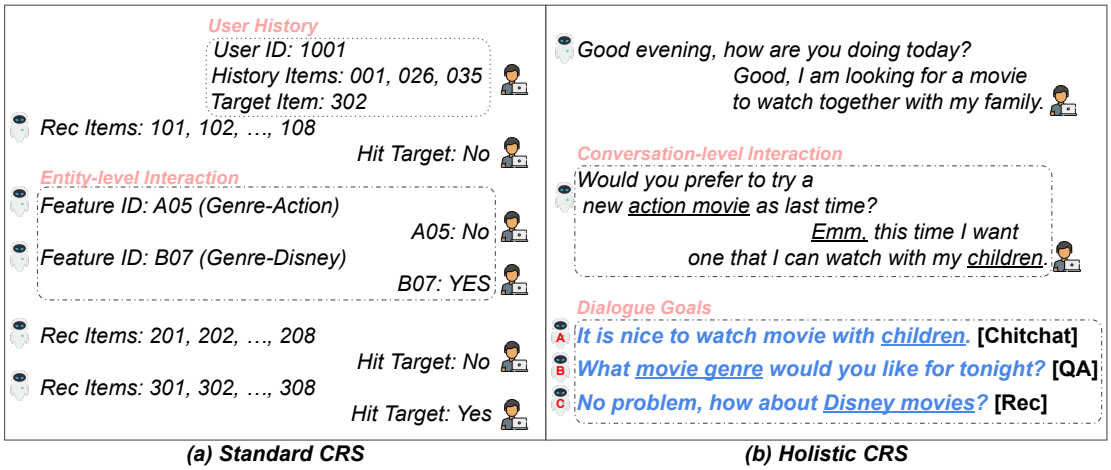

Figure 1: Examples of standard and holistic CRS. a) Standard CRS support multi-round interaction only at the entity level b) Holistic CRS support multi-round and multi-goal interaction at the conversation level.

tical contexts. Holistic CRS adopt real, conversation-level interaction and target multiple dialogue goals, as shown in Figure 1. Given the same entity pair <Genre-Disney> as the standard CRS in subfigure (a), the holistic system in subfigure (b) must generate questions like *"What movie genre would you like for tonight?"* and understand its related response correctly, before they use *<Genre-Disney>* for the recommendation. For the same question, the user may give unexpected answers like *"Show me a new movie this year!"*, inconsistent with the movie genre. Moreover, holistic CRS is required to leverage the rich contextual information inferred from the conversations [11] and from the semantic context. For example, given the input *"Emm"* in the user's second response, a holistic CRS might infer that the previous recommendation was unsatisfactory, prompting it to make a new and different recommendation.

The main challenges in the task of a holistic CRS are thus ones such as the following: *How to understand the users' intentions with limited contextual information? How should we generate reasonable responses with high recommendation quality? When faced with different inferred conversation goals, which goal should be pursued now?*

We systematically analyse the current holistic CRS work solving the above problems (§4), decomposing them into three components: 1) a backbone language model, and optional components incorporating 2) external knowledge and 3) external guidance. We follow this with an analysis of the datasets (§5) and evaluation methods (§6). We investigate the key challenges and promising research trends in this area (§7). To the best of our knowledge, this is the first survey on CRS with a special focus on conversational ("holistic") approaches. Our contributions are:

1. We provide a clear landscape of the tasks, models and hierarchical structure of holistic CRS.
2. We summarise, analyze and critique the existing methods, datasets and evaluation methods for selected works in a well-structured manner.
3. We outline key challenges, constraints and future directions for holistic CRS.

2. Definition and Background

In Figure 3, we split the field of CRS research into two distinct branches: standard and holistic CRS, further delineating them into Types 0, 1, and 2, based on their input–output dynamics.

***Type 0 standard CRS**, limited to entity-level inputs and outputs, is restricted in scope of interaction; e.g., [2, 3].*
***Type 1 holistic CRS** takes conversation as input and yields either entity-level recommendations or conversational responses, encompassing query interpretation and tailored linguistic outputs; e.g., [8, 12].*
***Type 2 holistic CRS** is more expansive, accepting and producing unrestricted inputs–outputs formats including conversations, knowledge and multimedia; e.g., [13, 14].*

Holistic CRS differ from standard CRS approaches in the following aspects: 1) The final goal for holistic CRS is to guide or convince users to accept the recommendation through multi-rounds of conversations. 2) Holistic CRS start from the conversations and ends by generating either recommendation results or responses. 3) Holistic CRS methods are evaluated on both recommendation and language quality using both automatic and human evaluation measures.

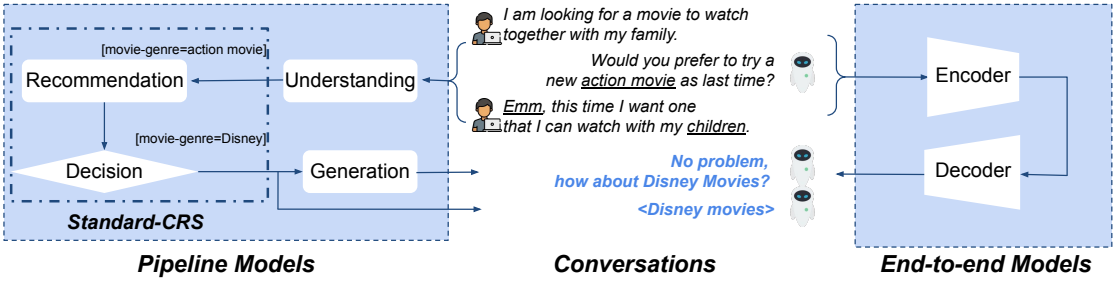

Figure 2: Pipeline models and end-to-end models in holistic CRS. Left: Pipeline models for holistic CRS include understanding, recommendation, decision and generation units while standard CRS only contain recommendation and decision units. Right: End-to-end holistic CRS with an encoder–decoder structure.

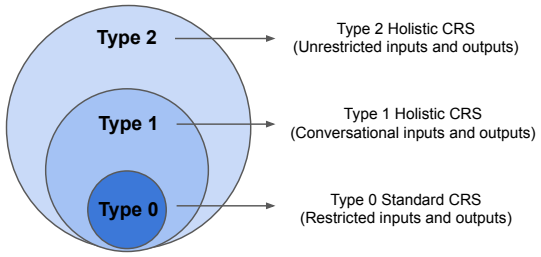

Figure 3: Hierarchical structure of CRS in terms of input and output types

2.1. Task Definition

In a task-oriented dialogue system, we restrict our consideration to the scenario where a singular system interacts with one individual user, denoted by u, and pre-determined items, represented by i. Each dialogue contains T turns of conversations, denoted as $C = \{s_j^{system}, s_j^{user}\}_{j=1}^{T}$, where each turn contains a single turn from the system and its associated response from the user. The user's entity-level interaction history of past j-th turn is denoted as $E_j^u = \{i_1^{(u)}, ..., i_j^{(u)}\}$ and dialogue history with past j-th turns is denoted $C_j^u = \{[s_1^{system}, s_1^{user}], ..., [s_j^{system}, s_j^{user}]\}$. Some methods provide knowledge or external guidance, which we denote as K. The target function for holistic CRS is expressed in two parts: to generate 1) next item prediction i_{j+1} and 2) next system response s_{j+1}^{system}. In summary, at the j-th turn, given the user's interaction history and contextual history, CRS generates either an entity-level recommendation results i_{j+1} or a conversation-level system response s_{j+1}^{system}, shown in Formula 1.

$$y^* = \prod_{j=1}^{T} P_\theta \left(i_{j+1}, s_{j+1}^{system} | E_j^u, C_j^u, K\right) \quad (1)$$

2.2. Structure of CRS

Figure 2 shows the two prevalent model structures in holistic CRS, which are pipeline and end-to-end models.

The pipeline structure of CRS contain four parts: *understanding unit, recommendation unit, decision unit* and *generation unit.* The understanding unit takes in dialogue and converts them into an entity–value pair for the recommendation unit to generate possible entity outputs. The decision unit controls the dialogue flow, while the generation unit generates the response accordingly.

With the development of Hierarchical Recurrent Encoder–Decoder (HRED) structure [8] and transformer-based encoder–decoder structure [15], components such as the understanding, decision and generation units are merged together to form an end-to-end structure.

3. Ontology and Existing Surveys

We aim to conduct an exhaustive survey on holistic CRS, focusing on Types 1 and 2 of our hierarchy. Our primary sources comprise leading NLP and Information Retrieval (IR) conferences and journals, as exemplified by premier venues such as ACL, ACM, AAAI and ScienceDirect. Furthermore, we delve into publicly accessible online resources, filtering papers by all variants of search terms in "conversational recommender systems". Matching work are then refined based on the following three criteria, regarding the features of the presented work: 1) It supports conversations as an input type. 2) It provides recommendation responses at either entity or conversation levels. 3) It facilitates multi-round interactions. For each selected work, we focus on the methodologies, datasets, and evaluation metrics.

While there exist surveys that offer an all-encompassing view of CRS, encompassing both standard and holistic CRS [1, 7, 16, 17], our survey purposefully structured and limited in scope to illuminate the evolution and development of holistic CRS only, particularly in their handling of conversational data.

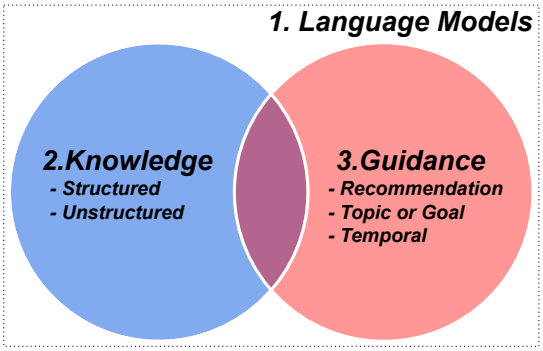

Figure 4: Components of holistic (Type 3) Conversational Recommendation System approaches: 1) requisite backbone language models, and optional components incorporating 2) external knowledge and/or 3) external guidance

Works centred on Type 0 standard CRS, given their lack of conversational aspects, are intentionally omitted.

4. Main Approaches & Discussion

Current holistic CRS approaches are primarily structured around three main components, as illustrated in Figure 4: 1) Language Models (LMs); 2) Knowledge; and 3) Guidance. A majority of holistic CRS systems hinge on LMs (§4.1), encompassing machine learning, deep learning, and pre-trained language models (PLMs), for foundational dialogue operations. However, these LMs often fall short in recommendation and commonsense reasoning. To bridge this gap, additional external knowledge (§4.2) and guidance (§4.3) are integrated, either independently or jointly. This section delineates the evolutionary path of their development, offering insights into their limitations and potential avenues for future progress.

4.1. Language Models

LMs serve as the backbone for holistic CRS in recommendation response generation with the evolution from machine learning [10], deep learning [8, 18] to PLMs [15, 12, 19]. The most popular LMs for response generation are HRED-based sequential models and transformer-based PLMs. These language models adopt a framework of end-to-end training, enabling them to be simultaneously trained in both conversation and recommendation tasks [8, 18].

Recent advancements in natural language processing (NLP) highlight the efficacy of PLMs like BERT and GPT [20, 21] in language generation and commonsense reasoning. Although those PLMs are not inherently optimized for CRS, researchers have explored their capabilities for holistic CRS tasks like recommendations and response generation. Penha and Hauff evaluated BERT's innate ability for recommendations using text-format probes for item or genre predictions without fine-tuning. In another line of work, Hayati et al. enhanced conversational tasks by adapting PLMs to produce varied recommendation responses incorporating social strategies, like encouragement or persuasion [12, 19]. Taking a multifaceted approach, Deng et al. segmented recommendation response generation into multiple tasks, including goal or topic planning, item recommendation and response generation. While having distinct tasks, they pre-trained a PLM end-to-end, underscoring the connection between holistic CRS and LMs and validating the effectiveness of the end-to-end training paradigm.

Discussion. *While PLMs can generate context-specific recommendation responses, they often fall short of meeting the dual requirements of recommendation accuracy and language quality, resulting from the phases of 1) pre-training and 2) online training.*

The inherent limitation of PLMs stems from their design for universal application. In contrast, recommendation tasks are focused and specific to certain domains [8, 23]. The implicit knowledge derived from general pre-training is insufficient to support them in making high-quality recommendations. Pre-training LMs with explicit task-specific knowledge is a solution, but comes associated with high costs and complications [24, 22]. Transferring such knowledge across diverse domains or user groups for real-world applications still poses a considerable challenge.

Holistic CRS rely heavily on online training, enabled by conversational interactions with benchmark datasets (§5). However, the restricted knowledge available in those datasets poses a formidable challenge for PLMs to generate quality recommendation responses, necessitating a model capable of integrating additional knowledge or guidance to facilitate preference tracking and response generation.

4.2. External Knowledge

Inherent limitations regarding implicit knowledge stored in PLMs are addressed in holistic CRS by integrating external knowledge. This enhances their capabilities in prediction, reasoning, and explanation. Methods augmented with knowledge often utilize graph convolutional networks (GCNs) [25] or relational graph convolutional networks (R-GCNs) [26] to extract knowledge representation from structured sources like knowledge graphs (KGs), or unstructured ones such as reviews. This representation is then incorporated into PLMs through semantic alignment or knowledge fusion techniques, enabling the production of refined recommendations [27, 28, 29]. We now delve into holistic CRS approaches that leverage both structured and unstructured knowledge sources.

4.2.1. Structured knowledge

Knowledge Graphs (KGs) are a prevalent source of structured knowledge. However, to be employed for holistic CRS tasks, they need to be transformed into an appropriate representation before the knowledge and textual features can be integrated.

KGs are typically represented by triplets comprising entities and relationships; e.g., *<Movie A-Genre-Disney>* where nodes representing item entities (*Movie A*) are connected to non-item entities (*Disney*) via edges that indicate relationships (*Genre*). In knowledge-enhanced CRS, the entities mentioned in conversations are first matched with entities in external KGs. Subsequently, graph propagation is performed to encode the KG's structural and relational information into knowledge representations [30]. Techniques like GCN and RGCN are employed in this stage to recurrently update node representations based on their neighbouring nodes. With the obtained knowledge representations, there are two main research directions in applying KGs to holistic CRS, which we denote as 1) node-level entity prediction and 2) edge-level path reasoning [31].

Node-level entity prediction in holistic CRS enhances response generation by incorporating additional item entities from the KG [30, 32]. In this usage, LMs extract knowledge representations from the KG and convert them into item-specific vocabularies, which are then integrated into recommendation responses. As a result, such responses are more fluent and informative, aligning closely to the original conversations and consistent with the user's interests [30, 32, 33].

Edge-level path reasoning provides a better approach to interpret users' preferences and dynamic shift in interests through the knowledge presentation than node-level entities [34, 35, 31, 36]. A strict, 2-hop KG reasoning is first proposed to interpret the user's preference through two steps (e.g.,*Movie A* ⇒*Actor1*⇒*Movie B*). For instance, given the user's watching history of Movies A and B, the model can infer the user's preference for Actor 1 and subsequently confirm its inference through conversation. However, due to the rule-based setting, 2-hop reasoning works well only when users have clearly-defined and straightforward preferences [35]. In situations where users demonstrate shifting interests, a multi-hop or tree-structure reasoning method is more suitable, translating implicit preference paths in KGs to explicit explanations in dialogues [34, 37, 38].

Well-constructed KGs enhance comprehensive knowledge representation in entity-level item selection and conversation-level preference reasoning or interpretation [31, 38]. However, due to the static nature of KGs, inferring the latest features of an item from structured knowledge sources poses significant challenges.

4.2.2. Unstructured knowledge

In unstructured knowledge sources (e.g., reviews or documents), a text retriever is employed to extract relevant textual segments from external documents. These segments are subsequently either transformed into nodes or edges of a new KG or merged into an existing KG [39, 29, 40, 41, 42]. The resultant KG can then be transferred into knowledge representations [41, 42, 43]. This method allows unstructured knowledge to supplement static knowledge graphs with contemporary information, allowing holistic CRS to be more versatile.

Knowledge Fusion and **Semantic Alignment** serve as the primary strategies to bridge the entity and semantic spaces in graph reasoning, leveraging both structured and unstructured knowledge resources. Knowledge Fusion integrates graph embeddings from KGs with text embeddings from LMs, enhancing both entity recommendations and conversational preference interpretations [30, 28]. Recently, Zhou et al. demonstrate a method that surpasses the performance of current fusion methods for entities and dialogues. They address the semantic gap between conversations and external knowledge with fine-grained semantic alignment techniques that align word-level semantic graphs with entity-level KGs [44, 45, 46]. Similarly, for models utilizing unstructured knowledge bases, contrastive learning strategies bridge the semantic gap across embeddings in dialogues, KGs and document reviews, potentially leveraging a spectrum of such knowledge resources [28].

Discussion. *The existing knowledge sources for holistic CRS are constrained in item space. However, as LMs become more robust, the reliance on conventional knowledge sources might decrease, while the necessity for guidance in other modalities may increase. Specifically, specialized knowledge (such as user profile representation and user–item relationship extraction) is likely to become crucial.*

The advent of powerful large language models (LLMs) serving as LMs, reduces reliance on external knowledge sources. This potentially makes the use of external sources redundant [47, 48]. The integration of external knowledge within LMs should start by evaluating a model's capabilities before knowledge incorporation, such as examining the capability of PLM in processing content-based recommendations [47, 49]. Recognizing the limitations of LMs before introducing the appropriate knowledge sources is a key issue in the advancement of holistic CRS.

4.3. External Guidance

Holistic CRS using external guidance train models for supplementary tasks — inclusive of recommendation, topic/goal planning, and temporal feature representation

Dataset	# P	# C	# T	# I	# M	Domain	Language	IR	# Pos	# Neg
REDIAL	20	10,006	182,150	6,223	8.50	Movie	Freestyle	0.96	94,150	15,377
TG-ReDial*	5	8,495	109,892	11,447	2.22	Movie	Topic-guided	0.40	89,693	2,971
DuRecDial*	4	5,678	87,301	531	15.15	Movie+Music	Multi-type	0.48	47,547	14,217
INSPIRED	4	801	16,982	1,378	10.05	Movie	Social strategy	0.35	12,589	2,395
OpenDialKG	3	13,802	91,209	4,232	82.49	Movie+Book	Knowledge path	0.38	63,856	15,798
GoRecDial	2	8,209	16,743	1,532	20.44	Movie	Game-play	0.77	88,601	19,354
MultiWOZ	2	8,420	221,588	1,737	238.58	7 Other Domains	Multi-domain	0.68	75,732	26,724

Table 1
Statistical analysis of the datasets in holistic CRS research. # P, # C, # T, # I, # M, # Pos, and # Neg stand for the number of papers, conversations, single turns, items, mentions for each item, positive and negative single turns in training data, IR: Informative turns rate. * Datasets are originally collected in Mandarin Chinese.

— in contrast to knowledge-enhanced models which fuse knowledge into PLMs. Results from these tasks serve as auxiliary guidance for LMs during recommendation response generation. Some models align both external knowledge and guidance, adopting a hybrid strategy that capitalizes on both dimensions for more robust response generation.

Recommendation guidance utilises approaches akin to template-based generation methods, decoupling conversation and recommendation result generation. LMs are conditioned to separately produce dialogues with placeholders that align with the original context and suggested items or attributes consistent with the user's history [50, 51, 52, 32, 53]. These placeholders are later substituted with corresponding recommendations.

Topic or goal guidance enhances the LM's proficiency in topic or goal planning. Although reinforcement learning techniques are predominantly employed in traditional CRS for action or goal planning, they are challenging to adapt as a representation for LMs [3, 18, 22].

Topic-guided systems initiate by building topic graphs, capturing or predicting specific target topics like "action movie" or "Disney movie". LMs subsequently use these graphs to guide recommendation response generation [54, 55, 33]. **Goal-guided systems** create hierarchical goal-type graphs derived from existing KGs and dialogues. The goal-planning module of the LMs is then trained on diverse dialogue goals, encompassing "QA", "recommendation", "greeting" or "chitchat" [9, 49, 56, 22]. These objectives also influence the dialogue policy and decision-making processes within holistic CRS.

Temporal guidance in CRS incorporates temporal features to formulate a time-aware representation, emphasizing the explicit and dynamic shift in users' preferences [57, 58]. Unlike traditional sequential recommendation systems that have access to users' historical profiles, holistic CRS often lack this depth of historical data. To address this gap, temporal features discern between historical dialogue sessions and the ongoing dialogue session, thereby capturing the multifaceted nature of users' preferences [59]. This differentiation allows the modelling of historical user preferences and continues to gather fresh preferences from active interactions. Additionally, such features aid in the construction of user profiles based on past behaviours, facilitating the retrieval of similar user profiles based on their relevance, enhancing preference modelling in a time-aware collaborative manner [60, 58]. In a distinct approach, Xu et al. put forth the idea of a user temporal KG, which contains both offline user knowledge in historical conversations and online knowledge in current or future conversation sessions. Representing a leap beyond traditional static knowledge graphs, temporal KGs have garnered significant interest [60, 37]. In the context of holistic CRS, dynamic reasoning utilizing temporal KGs represents an innovative and burgeoning research domain [37, 61, 38, 46].

Discussion. *Present methodologies for integrating external knowledge or guidance largely involve training LMs to interpret fed knowledge or representation, rather than guiding them to independently explore and extract the required information from external resources. This method, akin to "spoon-feeding" LMs with knowledge or guidance, contrasts with the envisioned future for holistic CRS. In our view, LMs should be provided with a knowledge "buffet", empowering autonomous gathering of necessary information and prioritising reasoning over interpretation [62].*

5. Datasets

In the realm of holistic CRS, the interaction between users and systems has led to the collection of several benchmark datasets. While some surveys have primarily summarized data from an item space perspective [1], our focus is to dive deeper into the publicly-available holistic CRS datasets. Our intention is understand datasets beyond traditional boundaries, expounding specifically on two dimensions: entity information and language quality [8, 12, 54, 18, 31, 63, 54].

Recommendation Accuracy		Language Quality			
Metrics	**# Papers**	**Metrics**	**# Papers**	**Human Evaluation**	**# Papers**
Recall@K	19	***Distinct-n***	18	***Fluency***	19
Hit@K	7	***BLEU***	15	***Informativeness***	17
NDCG@k	6	***Perplexity***	9	***Coherence***	8
MRR@K	6	Knowledge Precision	2	Relevance	4
F1	7	Entity Accuracy	1	Proactivity	2
Precision	2	Average Entity Number	1	Knowledge	2
Turn@K	1	***Topic Consistency***	1	Appropriateness	2
RMSE	1	***Success Rate***	1	***Consistency***	1

Table 2
Evaluation methods in holistic CRS. # Papers indicate the volume of work using the associated evaluation method.

5.1. Statistical analysis

Table 1 presents a statistical analysis of various datasets, detailing each dataset in terms of both entity and linguistic characteristics. In terms of entity space, the scale of a dataset is measured by the number of conversations and items it contains; while the informativeness is measured by the number of conversation turns and the number of mentions of specific items within them. Interestingly, our analysis reveals that a longer conversation does not necessarily correspond to mentions of more items. Rather we believe that ensuring a consistent frequency of item mentions is paramount for the recommendation system's learning efficacy [64].

From the perspective of language, most datasets are compiled from predominately English data and focus on the movie domain. Recent datasets indicate a decline in the ratio of informative turns. This trend aligns with real-world conversational patterns, where interactions are transforming into conversations that contain a growing amount of general or chit-chat content [12, 19]. This observation reinforces our belief that an optimal dataset should capture authentic human behaviour and not merely translate entity-centric data into dialogues. The data also suggests that positive turns — ones that provide constructive or affirmative feedback –– are more valuable for recommendations compared to negative ones [65, 66]. In sum, it is not merely about the volume of training data, but about the quality, authenticity, and informativeness of the conversations therein.

5.2. Limitations

The objective of Holistic CRS datasets is to accurately emulate real-world scenarios and offer labelled information for efficient learning. However, our evaluation reveals three primary limitations in the existing datasets: First, some datasets diverge from real-world conversations, which impedes the quality of learned interactions [18]. A notable example is the game setting where the dialogue's objective is to guess a target item, disrupting the natural flow of conversation as seekers are already privy to the target item's identity. Second, a significant proportion of datasets predominantly focus on the movie domain [8, 30], potentially damaging the generalizability of conclusions drawn on CRS research. Third, current datasets do not offer sufficient labels outside the confines of the item space [8, 12]. Addressing these shortcomings will be pivotal for productive future research in holistic CRS.

6. Evaluation Methods

CRS generate both recommendation results and responses. Their evaluation require appropriate mechanisms to assess the quality of both the recommended items and the resulting dialogue as a whole. Existing evaluation methods examine both recommendation accuracy (as in traditional recommendation systems) and language quality (as in NLP language modelling) separately, using both metrics and human evaluation. We compile the frequency of these methods from the works in §4 as Table 2.

6.1. Recommendation Evaluation

Recommendation evaluation metrics categorise along three lines: point-wise accuracy methods (RMSE), decision support methods (F1) and ranking-based methods (Recall@K). The evaluation metrics for holistic CRS are similar to those in standard CRS, where they mostly evaluate the recommendation from the item level. However, for holistic CRS, it is equally important to evaluate the recommendation performance separately at the conversation level in order to ensure information consistency in response generation [32].

6.2. Language Evaluation

While most of the recommendation results can be evaluated with metrics, it still requires human beings to evaluate the language generation quality as the golden stan-

dard. Metric-based approaches, as auxiliary solutions, provide a fast and simple evaluation of holistic CRS. Language evaluation metrics such as *Distinct n-gram, BLEU* and *Perplexity* evaluate language quality regarding diversity and fluency.

Human evaluation provides a fair evaluation of different models from the viewpoints of users and in a double-blind way [51, 10]. It is relatively fast and convenient for human annotators to provide a high-quality evaluation in terms of *fluency* and *informativeness*. However, as the human evaluation may only be limited to one or few turns over the whole conversation, it is challenging for the annotators to fully examine the *coherence* and *consistency*, which generally requires the full understanding of dialogue [6].

Unlike recommendation systems which merely compare item rankings with respect to the target item, in holistic CRS, implicit features like personality, persuasion, and encouragement also contribute to the success of a recommendation [12]. Evaluating a system based on user experience remains challenging. It is imperative to introduce automatic assessment methods for both system-generated quality and user-centric experiences. [17, 67, 68, 69].

7. Challenges & Future Trends

As we have detailed the development of holistic CRS, we now highlight current challenges and suggest future directions to round out our overview.

Language generation quality and style. Current holistic CRS methods do not meet the requirements for practical application due to their inferior language quality scores in human evaluation, even when compared to retrieval-based methods [70, 51, 71]. Successful recommendation responses need to supplement explicit prediction results by accounting for implicit features like social strategy and language styles (e.g., encouragement and informativeness [12, 65, 66]). As recommendation outcomes often draw from an external or enriched knowledge structure, future research should focus on 1) elevating language quality to garner positive user feedback [72], and 2) emphasizing preferred language styles to enhance user acceptance [73].

User-centric holistic CRS. Holistic CRS has made strides towards user-centricity by facilitating conversational feedback between the user and the system. Nonetheless, its feedback and recommendation spectrum is still restricted. To enhance its efficacy, future versions of holistic CRS should prioritize personalised experiences for individual users by harnessing multi-modal data from item categories and user profiles. Moreover, attending to users' personal feedback and latent preferences is key for building a superior user modelling framework, resulting in more pertinent recommendations [74]. Additionally, incorporating other LMs or AI-generated content (AIGC) into recommendation feedback could also be a promising avenue [75, 76].

Unified model for holistic CRS. Large Language Models (LLMs) have significantly advanced task-oriented dialogue systems, allowing for integrated handling of various tasks in a conversational manner [77, 78]. In the realm of recommendation systems, some research has adopted a two-phase training approach (pre-training and fine-tuning), leveraging text for recommendations, reasoning and explanation [61, 79, 80]. Yet, while there's a push to integrate PLMs into CRS tasks using a text-to-text paradigm, the broader holistic CRS research domain has not achieved a standardized problem framework, which would enable seamless integration with task-specific models and swift adaptation to similar tasks across different domains [32, 44, 24]. LLMs, on their own, cannot address every CRS challenge. Current holistic CRS models lean heavily on complex ensemble architectures that merge LMs with external knowledge or guidance. As such, crafting a unified model framework with consistent problem definitions remains a pivotal research avenue [32, 44].

8. Conclusion

Despite the rising interest in standard conversational recommendation systems which are restricted to entity-level input and output, our study reveals the necessity and current negligence of holistic CRS, which encompasses all forms of input and output, catering for real-world situations. In this paper, we systematically describe the important components of holistic CRS, including 1) language models, 2) knowledge resources, and 3) external guidance. To the best of our knowledge, our survey is the first systematic review specifically dedicated to holistic CRS with conversational approaches, which further summarized common datasets, evaluation methods and challenges. Existing ascendant works enlighten a number of promising future directions from the above perspectives. Through clear landscapes in holistic CRS, we hope to attract more attention to explore a more natural and realistic setting in this challenging but promising area.

References

[1] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Comput. Surv. 54 (2021). doi:10.1145/3453154.

[2] Y. Zhang, X. Chen, Q. Ai, L. Yang, W. B. Croft, Towards conversational search and recommendation: System ask, user respond, in: Proceedings of the

27th acm international conference on information and knowledge management, 2018, pp. 177–186.

[3] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, T.-S. Chua, Estimation-action-reflection: Towards deep interaction between conversational and recommender systems, in: Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 304–312. doi:10.1145/3336191.3371769.

[4] J. Ni, T. Young, V. Pandelea, F. Xue, E. Cambria, Recent advances in deep learning based dialogue systems: A systematic survey, Artificial intelligence review (2022) 1–101.

[5] Y. Dai, H. Li, Y. Li, J. Sun, F. Huang, L. Si, X. Zhu, Preview, attend and review: Schema-aware curriculum learning for multi-domain dialogue state tracking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 879–885. URL: https://aclanthology.org/2021.acl-short.111. doi:10.18653/v1/2021.acl-short.111.

[6] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438.

[7] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, AI Open 2 (2021) 100–126. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000164. doi:https://doi.org/10.1016/j.aiopen.2021.06.002.

[8] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 9748–9758.

[9] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, T. Liu, Towards conversational recommendation over multi-type dialogs, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1036–1049. URL: https://aclanthology.org/2020.acl-main.98. doi:10.18653/v1/2020.acl-main.98.

[10] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 815–824.

[11] H. Weld, X. Huang, S. Long, J. Poon, S. C. Han, A survey of joint intent detection and slot filling models in natural language understanding, ACM Computing Surveys 55 (2022) 1–38.

[12] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, Inspired: Toward sociable recommendation dialog systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8142–8152. URL: https://www.aclweb.org/anthology/2020.emnlp-main.654.

[13] T. Yu, Y. Shen, H. Jin, Towards hands-free visual dialog interactive recommendation, in: AAAI, 2020.

[14] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, Information Fusion 77 (2022) 149–171. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001512. doi:https://doi.org/10.1016/j.inffus.2021.07.009.

[15] G. Penha, C. Hauff, What does bert know about books, movies and music? probing bert for conversational recommendation, in: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 388–397. doi:10.1145/3383313.3412249.

[16] F. Radlinski, C. Boutilier, D. Ramachandran, I. Vendrov, Subjective attributes in conversational recommendation systems: challenges and opportunities, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12287–12293.

[17] D. Jannach, Evaluating conversational recommender systems: A landscape of research, Artificial Intelligence Review 56 (2023) 2365–2400.

[18] D. Kang, A. Balakrishnan, P. Shah, P. A. Crook, Y. Boureau, J. Weston, Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue, CoRR abs/1909.03922 (2019). URL: http://arxiv.org/abs/1909.03922. arXiv:1909.03922.

[19] A. Manzoor, D. Jannach, Inspired2: An improved dataset for sociable conversational recommendation, arXiv preprint arXiv:2208.04104 (2022).

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[21] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, Dialogpt: Large-

scale generative pre-training for conversational response generation, arXiv preprint arXiv:1911.00536 (2019).

[22] Y. Deng, W. Zhang, W. Xu, W. Lei, T.-S. Chua, W. Lam, A unified multi-task learning framework for multi-goal conversational recommender systems, ACM Trans. Inf. Syst. 41 (2023). doi:10.1145/3570640.

[23] C. Yang, Y. Hou, Y. Song, T. Zhang, J.-R. Wen, W. X. Zhao, Modeling two-way selection preference for person-job fit, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 102–112. doi:10.1145/3523227.3546752.

[24] S. Geng, S. Liu, Z. Fu, Y. Ge, Y. Zhang, Recommendation as language processing (rlp): A unified pretrain, personalized prompt predict paradigm (p5), in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 299–315. doi:10.1145/3523227.3546767.

[25] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, CoRR abs/1609.02907 (2016). URL: http://arxiv.org/abs/1609.02907. arXiv:1609.02907.

[26] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: European semantic web conference, Springer, 2018, pp. 593–607.

[27] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, 2018. URL: http://arxiv.org/abs/1808.09781. arXiv:1808.09781 [cs].

[28] T. Zhang, Y. Liu, P. Zhong, C. Zhang, H. Wang, C. Miao, KECRS: Towards knowledge-enriched conversational recommendation system, 2021. URL: http://arxiv.org/abs/2105.08261. arXiv:2105.08261 [cs].

[29] Y. Zhou, K. Zhou, W. X. Zhao, C. Wang, P. Jiang, H. Hu, C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 1488–1496.

[30] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, J. Tang, Towards knowledge-based recommender dialog system, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1803–1813. URL: https://aclanthology.org/D19-1189. doi:10.18653/v1/D19-1189.

[31] S. Moon, P. Shah, A. Kumar, R. Subba, OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 845–854. URL: https://aclanthology.org/P19-1081. doi:10.18653/v1/P19-1081.

[32] L. Wang, H. Hu, L. Sha, C. Xu, K. Wong, D. Jiang, Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph, CoRR abs/2110.07477 (2021). URL: https://arxiv.org/abs/2110.07477. arXiv:2110.07477.

[33] J. Zhang, Y. Yang, C. Chen, L. He, Z. Yu, KERS: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1092–1101. URL: https://aclanthology.org/2021.findings-emnlp.94. doi:10.18653/v1/2021.findings-emnlp.94.

[34] J. Zhou, B. Wang, R. He, Y. Hou, CRFR: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 4324–4334. URL: https://aclanthology.org/2021.emnlp-main.355. doi:10.18653/v1/2021.emnlp-main.355.

[35] W. Ma, R. Takanobu, M. Tu, M. Huang, Bridging the gap between conversational reasoning and interactive recommendation, CoRR abs/2010.10333 (2020). URL: https://arxiv.org/abs/2010.10333. arXiv:2010.10333.

[36] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, P. S. Yu, User memory reasoning for conversational recommendation, arXiv preprint arXiv:2006.00184 (2020).

[37] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, B. Liu, P. Yu, User memory reasoning for conversational recommendation, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5288–5308. URL: https://aclanthology.org/2020.coling-main.463. doi:10.18653/v1/2020.coling-main.463.

[38] W. Li, W. Wei, X. Qu, X.-L. Mao, Y. Yuan, W. Xie, D. Chen, TREA: Tree-structure reasoning schema for conversational recommendation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2970–2982. URL:

https://aclanthology.org/2023.acl-long.167.

[39] L. Liao, R. Takanobu, Y. Ma, X. Yang, M. Huang, T. Chua, Deep conversational recommender in travel, CoRR abs/1907.00710 (2019). URL: http://arxiv.org/abs/1907.00710. arXiv:1907.00710.

[40] Y. Lu, J. Bao, Y. Song, Z. Ma, S. Cui, Y. Wu, X. He, RevCore: Review-augmented conversational recommendation, 2021. URL: http://arxiv.org/abs/2106.00957. arXiv:2106.00957 [cs].

[41] Y. Li, B. Peng, Y. Shen, Y. Mao, L. Liden, Z. Yu, J. Gao, Knowledge-grounded dialogue generation with a unified knowledge representation, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 206–218. URL: https://aclanthology.org/2022.naacl-main.15. doi:10.18653/v1/2022.naacl-main.15.

[42] B. Yang, C. Han, Y. Li, L. Zuo, Z. Yu, Improving conversational recommendation systems' quality with context-aware item meta-information, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 38–48. URL: https://aclanthology.org/2022.findings-naacl.4. doi:10.18653/v1/2022.findings-naacl.4.

[43] X. Zhang, X. Xin, D. Li, W. Liu, P. Ren, Z. Chen, J. Ma, Z. Ren, Variational reasoning over incomplete knowledge graphs for conversational recommendation, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 231–239.

[44] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1006–1014.

[45] J. Wu, B. Yang, D. Li, L. Deng, A semantic relation-aware deep neural network model for end-to-end conversational recommendation, Applied Soft Computing 132 (2023) 109873.

[46] J. Zhou, B. Wang, M. Huang, D. Zhao, K. Huang, R. He, Y. Hou, Aligning recommendation and conversation via dual imitation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 549–561. URL: https://aclanthology.org/2022.emnlp-main.36.

[47] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, arXiv preprint arXiv:1909.01066 (2019).

[48] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, J. McAuley, Large language models as zero-shot conversational recommenders, arXiv preprint arXiv:2308.10053 (2023).

[49] Z. Liu, D. Zhou, H. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, T. Liu, H. Xiong, Graph-grounded goal planning for conversational recommendation (2022) 1–1. doi:10.1109/TKDE.2022.3147210, conference Name: IEEE Transactions on Knowledge and Data Engineering.

[50] A. Manzoor, D. Jannach, Generation-based vs. retrieval-based conversational recommendation: A user-centric comparison, in: Fifteenth ACM Conference on Recommender Systems, ACM, ????, pp. 515–520. URL: https://dl.acm.org/doi/10.1145/3460231.3475942. doi:10.1145/3460231.3475942.

[51] A. Manzoor, D. Jannach, Towards retrieval-based conversational recommendation, Information Systems (2022) 102083.

[52] X. Wang, K. Zhou, J.-R. Wen, W. X. Zhao, Towards unified conversational recommender systems via knowledge-enhanced prompt learning, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1929–1937. doi:10.1145/3534678.3539382.

[53] Z. Liang, H. Hu, C. Xu, J. Miao, Y. He, Y. Chen, X. Geng, F. Liang, D. Jiang, Learning neural templates for recommender dialogue system, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7821–7833. URL: https://aclanthology.org/2021.emnlp-main.617. doi:10.18653/v1/2021.emnlp-main.617.

[54] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, 2020.

[55] L. Liao, R. Takanobu, Y. Ma, X. Yang, M. Huang, T.-S. Chua, Topic-guided conversational recommender in multiple domains, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 2485–2496. doi:10.1109/TKDE.2020.3008563.

[56] D. Lin, J. Wang, W. Li, Target-guided knowledge-aware recommendation dialogue system: An empirical investigation, in: Proceedings of the Joint KaRS & ComplexRec Workshop. CEUR-WS, 2021.

[57] X. Zeng, J. Li, L. Wang, Z. Mao, K.-F. Wong, Dynamic online conversation recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 3331–3341. URL: https://aclanthology.org/2020.

acl-main.305. doi:10.18653/v1/2020.acl-main.305.

[58] J. Zou, E. Kanoulas, P. Ren, Z. Ren, A. Sun, C. Long, Improving conversational recommender systems via transformer-based sequential modelling, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2319–2324. doi:10.1145/3477495.3531852.

[59] S. Li, R. Xie, Y. Zhu, X. Ao, F. Zhuang, Q. He, User-centric conversational recommendation with multi-aspect user modeling, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 223–233. URL: http://arxiv.org/abs/2204.09263. doi:10.1145/3477495.3532074. arXiv:2204.09263 [cs].

[60] L. Wang, S. Joty, W. Gao, X. Zeng, K.-F. Wong, Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge, 2022. URL: http://arxiv.org/abs/2209.11386. arXiv:2209.11386 [cs].

[61] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, CoRR abs/1904.06690 (2019). URL: http://arxiv.org/abs/1904.06690. arXiv:1904.06690.

[62] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, et al., Check your facts and try again: Improving large language models with external knowledge and automated feedback, arXiv preprint arXiv:2302.12813 (2023).

[63] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4335–4347. URL: https://aclanthology.org/2021.emnlp-main.356. doi:10.18653/v1/2021.emnlp-main.356.

[64] H. Hu, X. He, J. Gao, Z.-L. Zhang, Modeling personalized item frequency information for next-basket recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1071–1080.

[65] V. Bursztyn, J. Healey, N. Lipka, E. Koh, D. Downey, L. Birnbaum, "it doesn't look good for a date": Transforming critiques into preferences for conversational recommendation systems, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1913–1918. URL: https://aclanthology.org/2021.emnlp-main.145. doi:10.18653/v1/2021.emnlp-main.145.

[66] Y. Wu, C. Macdonald, I. Ounis, Multimodal conversational fashion recommendation with positive and negative natural-language feedback, in: Proceedings of the 4th Conference on Conversational User Interfaces, CUI '22, Association for Computing Machinery, New York, NY, USA, 2022. doi:10.1145/3543829.3543837.

[67] R. Lowe, M. D. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, J. Pineau, Towards an automatic turing test: Learning to evaluate dialogue responses, CoRR abs/1708.07149 (2017). URL: http://arxiv.org/abs/1708.07149. arXiv:1708.07149.

[68] C. Zhang, L. F. D'Haro, R. E. Banchs, T. Friedrichs, H. Li, Deep am-fm: Toolkit for automatic dialogue evaluation, in: Conversational Dialogue Systems for the Next Decade, Springer, 2021, pp. 53–69.

[69] S. Zhang, K. Balog, Evaluating conversational recommender systems via user simulation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1512–1520. doi:10.1145/3394486.3403202.

[70] A. Manzoor, D. Jannach, Generation-based vs retrieval-based conversational recommendation: A user-centric comparison, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 515–520. doi:10.1145/3460231.3475942.

[71] S. Zhang, K. Balog, Evaluating conversational recommender systems via user simulation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1512–1520. URL: http://arxiv.org/abs/2006.08732. doi:10.1145/3394486.3403202. arXiv:2006.08732 [cs].

[72] D. Jannach, A. Manzoor, End-to-end learning for conversational recommendation: A long way to go?, in: IntRS@ RecSys, 2020, pp. 72–76.

[73] P. P. Rau, Y. Li, D. Li, Effects of communication style and culture on ability to accept recommendations from robots, Computers in Human Behavior 25 (2009) 587–595. URL: https://www.sciencedirect.com/science/article/pii/S0747563208002367. doi:https://doi.org/10.1016/j.chb.2008.12.025, including the Special Issue: State of the Art Research into Cognitive Load Theory.

[74] D. Pramod, P. Bafna, Conversational recommender systems techniques, tools, acceptance, and adop-

tion: A state of the art review, Expert Systems with Applications 203 (2022) 117539.
[75] H.-C. Kuo, Y.-N. Chen, Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 249–258. URL: https://aclanthology.org/2023.findings-acl.17.
[76] W. Wang, X. Lin, F. Feng, X. He, T.-S. Chua, Generative recommendation: Towards next-generation recommender paradigm, arXiv preprint arXiv:2304.03516 (2023).
[77] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020.
[78] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. S. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. H. Chi, Q. Le, Lamda: Language models for dialog applications, CoRR abs/2201.08239 (2022). URL: https://arxiv.org/abs/2201.08239. arXiv:2201.08239.
[79] S. Chen, X. Liu, J. Gao, J. Jiao, R. Zhang, Y. Ji, Hitter: Hierarchical transformers for knowledge graph embeddings, CoRR abs/2008.12813 (2020). URL: https://arxiv.org/abs/2008.12813. arXiv:2008.12813.
[80] J. Li, M. Wang, J. Li, J. Fu, X. Shen, J. Shang, J. McAuley, Text is all you need: Learning language representations for sequential recommendation, arXiv preprint arXiv:2305.13731 (2023).
[81] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, M. Gašić, 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 5016–5026. URL: https://aclanthology.org/D18-1547. doi:10.18653/v1/D18-1547.
[82] A. Iovine, F. Narducci, M. de Gemmis, A dataset of real dialogues for conversational recommender systems., in: CLiC-it, 2019.
[83] Y. Lu, J. Bao, Z. Ma, X. Han, Y. Wu, S. Cui, X. He, AUGUST: an automatic generation understudy for synthesizing conversational recommendation datasets, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10538–10549. URL: https://aclanthology.org/2023.findings-acl.670.

9. Appendix

9.1. Datasets for CRS

We provide a detailed description of each dataset in Table 3. From the perspective of language, each dataset has a different focus. GoRecDial [18] uses a game setting to guide the dialogues while TG-Redial [54] uses topic to guide the crowd workers. That guidance are utilized to facilitate the CRS towards the target goals. OpenDialKG [31] pairs each conversation with a corresponding KG path while DuRecDial [9] further includes user profile and different goals (QA, chitchat, recommendation). These two datasets provide additional knowledge to item space and they are important for knowledge-enhanced models. INSPIRED [12] emphasizes more on the social strategies in making a successful recommendation with more than half utterances involving a social strategy. MultiWoz [81] collects mainly human-to-human dialogues in multiple domains. Instead of focusing on target item prediction, these two datasets demonstrate real scenarios that aim for successful acceptance in real life. Other datasets that are not publicly available are not included in this survey [36, 82, 83]

Dataset	Description
REDIAL [8]	First CRS dataset collected from crowd workers using a paired mechanism, where one person acts as a recommender and the other person acts as a movie seeker. Crowd workers are free to generate dialogues that meet the basic quality instructions.
TG-ReDial* [54]	A Chinese CRS datasets with topic-guided dialogues. Using real watching records of real online users to create different topic threads that further generate conversations.
DuRecDial* [9, 63]	A bi-lingual CRS datasets with additional annotation of users' profile, dialogue goals(QA, chitchat, recommendation) and knowledge. It is collected in Chinese with paired mechanisms and translated into the English version.
GoRecDial [18]	A goal-driven CRS dataset where the recommender aims to look for the target items by chatting with the seeker. A pair mechanism is adopted and candidate items are provided for each conversation.
OpenDialKG [31]	A dialogue dataset on movie and book domain with annotated knowledge graphs and relation paths related to each conversation.
INSPIRED [12, 19]	First CRS dataset proposed to create dialogues with different social strategies and preference elicitation strategies using the paired mechanism. Crowd workers are asked to finish 3 pre-task personality tests and a post-task survey with demographic questions.
MultiWoz [81]	A large transcript of human-to-human dialogues among 7 domains, eg: hotels, restaurants, attractions, taxis, trains, hospitals, police. It contains a large corpus of multi-domain dialogues with labelled dialogue states.

Table 3
Description of datasets. *Datasets are first collected in Mandarin Chinese.