# Soliciting User Preferences in Conversational Recommender Systems via Usage-related Questions

Ivica Kostric
University of Stavanger
Stavanger, Norway
ivica.kostric@uis.no

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

Filip Radlinski
Google
London, UK
filiprad@google.com

## ABSTRACT

A key distinguishing feature of conversational recommender systems over traditional recommender systems is their ability to elicit user preferences using natural language. Currently, the predominant approach to preference elicitation is to ask questions directly about items or item attributes. These strategies do not perform well in cases where the user does not have sufficient knowledge of the target domain to answer such questions. Conversely, in a shopping setting, talking about the planned use of items does not present any difficulties, even for those that are new to a domain. In this paper, we propose a novel approach to preference elicitation by asking implicit questions based on item usage. Our approach consists of two main steps. First, we identify the sentences from a large review corpus that contain information about item usage. Then, we generate implicit preference elicitation questions from those sentences using a neural text-to-text model. The main contributions of this work also include a multi-stage data annotation protocol using crowdsourcing for collecting high-quality labeled training data for the neural model. We show that out approach is effective in selecting review sentences and transforming them to elicitation questions, even with limited training data.

## CCS CONCEPTS

• **Information systems → Recommender systems**; *Users and interactive retrieval.*

## KEYWORDS

Conversational recommender systems, preference elicitation, question generation

**ACM Reference Format:**
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), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3460231.3478861

## 1 INTRODUCTION

Traditionally, recommender systems predict users' preference towards an item by performing offline analysis of past interaction data (e.g., click history, past visits, item ratings) [7]. These systems often do not take into account that users might have made mistakes in the past (e.g., regarding purchases) [28] or that their preferences change over time [9]. Additionally, for some users, there is little historical data which makes modeling their preferences difficult [11]. A *conversational recommender system* (CRS), on the other hand, is a multi-turn, interactive recommender system that can elicit user preferences in real-time using natural language [10]. Given its interactive nature, it is capable of modeling dynamic user preferences and take actions based on users current needs [7].

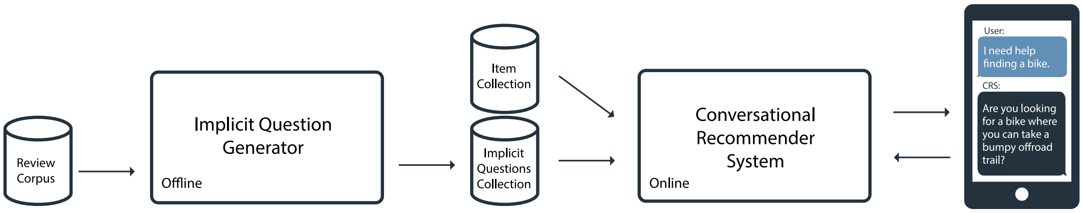
One of the main tasks of a conversational recommender system is to elicit preferences from users. This is traditionally done by asking questions either about items directly or item attributes [4–7, 12, 25, 26, 30, 32]. Asking directly about specific items is inefficient due to a vast number of items in the collection; therefore, the majority of the research is focused on the estimation and utilization of users preferences towards attributes [7]. Common to these approaches is that the user is explicitly asked about the desired values for a specific product attribute, much in the spirit of slot-filling dialogue systems [8]. For example, in the context of looking for a bicycle recommendation, we might have wheel dimensions or the number of gears as attributes in our item collection. In this case, a system might want to ask a question like *How thick should the tires be?* or *How many gears should the bike have?* However, ordinary users often do not possess this kind of attribute understanding, which might require extensive domain-specific knowledge. Instead, they only know where or how they intend to use the item. For example, a user might only be interested in using this bike for commuting but does not know what attributes might be good for that purpose. The novel research objective of this work is to generate *implicit* questions for eliciting user preferences, related to the intended use of items. This stands in contrast with explicit questions that ask about specific item attributes.

Our approach hinges on the idea that usage-related experiences are captured in item reviews. By identifying review sentences that discuss particular item features or aspects (e.g., *fat tires*) that matter in the context of various activities or usage scenarios (e.g., *for conquering tough terrain*), those sentences can then be turned into preference elicitation questions. In our envisaged scenario, a large collection of implicit preference elicitation questions is generated offline, and then utilized later in real-time interactions by a CRS; see Fig. 1 for an illustration.

In this paper, our focus is on the offline question generation part. Specifically, we start with *candidate sentence selection*, which can effectively be implemented based on part-of-speech tagging and simple linguistic patterns. Given a candidate sentence as input, *question generation* produces an implicit question or the label

*RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands*

ACM ISBN 978-1-4503-8458-2/21/09.
https://doi.org/10.1145/3460231.3478861


Review Corpus
Implicit Question Generator
Offline
Item Collection
Implicit Questions Collection
Conversational Recommender System
Online
User:
CRS:

I need help finding a bike.

Are you looking for a bike where you can take a bumpy offroad trail?

**Figure 1: Conceptual system overview. Our focus in this paper is on the implicit question generator component.**

N/A (not applicable). This is done by fine-tuning a pre-trained, sequence-to-sequence model for text generation [24]. The main challenge associated with this task is the collection of high-quality training data. We develop a multi-stage data annotation protocol via crowdsourcing to generate a sentence-to-question dataset. The process consists of generating questions, validating them, as well as expanding the variation of questions. Evaluating our proposed approach against held-back test data shows its effectiveness and its capability of generating questions that are suitable for preference elicitation, can simply be answered, and are grammatically correct.

In summary, our main contributions in this paper are as follows: (1) Introduce the novel task of eliciting preferences in CRSs via implicit (usage-oriented) questions. (2) Devise an approach for generating usage-related questions based on a corpus of item reviews, consisting of two main steps: candidate sentence selection (based on linguistic patterns) and question generation (using a neural sequence-to-sequence model). (3) Develop a multi-stage data annotation protocol using crowdsourcing for collecting high-quality ground truth data. (4) Perform an experimental evaluation of the proposed approach, followed by an analysis of results. The resources developed in this paper (crowdsourced dataset and question generation model) are made publicly available at https://github.com/iai-group/recsys2021-crs-questions.

## 2 RELATED WORK

In this paper, we focus on question-based user preference elicitation and natural language generation, both of which are identified as major challenges in [7]. That is, we provide novel answers to questions *what to ask* and *how to ask*.

### 2.1 Preference Elicitation

Commonly, preference elicitation questions target either items or their attributes. Typical of early studies on CRS, *item-based elicitation* approaches ask for users' opinions on an item itself, using a combination of methods from traditional recommender systems, such as collaborative filtering, with user interaction in real time [27, 35]. The selection of items may be approached as an optimization problem using a static preference questionnaire method [25] or multi-armed bandit algorithms that capture the exploration-exploitation tradeoff [6, 27]. Asking about items directly has been found inefficient, as large item sets would require several conversational turns and in turn increase the likelihood of users getting bored [7]. Alternatively, *attribute-based elicitation* aims to predict the next attribute to ask about. It is often cast as a sequence-to-sequence prediction problem, lending itself naturally to sequential neural networks. However, obtaining large conversational datasets to train conversational recommender systems is challenging [10], therefore, non-conversational data is often leveraged. Christakopoulou et al. [5] propose a question & recommendation (Q&R) method, utilize data from a non-conversational recommendation system, and develop surrogate tasks to answer questions: *What to ask?* and *How to respond?* A similar approach of training a sequential neural network on non-conversational data is taken by Zhang et al. [32], who convert Amazon reviews into artificial conversations. The assumption is that the earlier aspect-value pairs appear in the review, the more important they are to the user and should be prioritized as questions. Additionally, they develop a heuristic trigger to decide whether the model should ask about another attribute or recommend an item. Another way to elicit preferences is in the form of *critiques*, i.e., feedback on attribute values of recommended items [4]. For example if the recommendation is for a *phone*, a critique might be *not so big* or *something cheaper*. Such methods often employ heuristics as elicitation tactics [18, 19]. In recent work, Balog et al. [1] study the problem of robustly interpreting an unconstrained natural language feedback on attributes.

### 2.2 Question Generation

While there is research on end-to-end frameworks to enable CRSs to both understand user intentions as well as generate fluent and meaningful natural language responses [13], the predominant approach is still to use templates or construct the utterances using predefined language patterns [7]. Looking at the broader field of dialogue systems, there are two additional strands of research that could be applied to CRSs as well: retrieval-based and generation-based methods. Instead of relying on a handful of templates, *retrieval-based methods* utilize a large collection of possible responses. The basic approach to retrieving the appropriate response is based on some notion of similarity between the user query and candidate responses, with the simplest being inner product [31]. *Generation-based methods* in dialogue systems are typically based on sequence-to-sequence modeling. These models are usually trained on a hand-labeled corpus of task-oriented dialogue [3]. Our proposed approach shares elements of both of these methods: it generates questions using a sequence-to-sequence model and stores them in a collection that can be queried using retrieval-based methods.

## 3 APPROACH

We present an approach for generating a collection of implicit elicitation questions from a review corpus. Item review datasets tend to be very large, with both the number of items and reviews in the thousands or even millions, making labeling the entire dataset extremely expensive [14]. To overcome this, we extract candidate sentences from a corpus that have a high probability of mentioning item-related activity or usage (Section 3.1). We wish to train a model that can take a candidate sentence as input and generate a preference elicitation question out of that or the label N/A if it is not possible (Section 3.2). We opt for a pre-trained, transformer-based, state-of-the-art, sequence-to-sequence model (T5).

## 3.1 Candidate Sentence Selection

We identify sentences that describe some item feature or aspect (§3.1.1) and mention some activity or usage (§3.1.2).

$$\texttt{The}\ \overbrace{\texttt{fat}}^{value}\ \overbrace{\texttt{tires}}^{aspect}\ \texttt{are perfect for}\ \overbrace{\texttt{conquering tough terrain}}^{usage/activity}\texttt{.}$$

*3.1.1 Aspect-Value Pair Extraction.* An aspect in this context is a term that characterizes a particular feature of an item [17] (e.g., *wheel*, *seat* or *gear* are aspects of a bicycle). Value words are terms that describe an aspect (e.g., a *wheel* might be *large* or *small*, a *seat* can be *hard* or *comfortable*). Here, we extract all sentences that mention some aspect-value pair for a given category of items, using phrase-level sentiment analysis proposed by Zhang et al. [33, 34]. The motivation for this step stems from the assumption that an activity or usage can be mapped to a particular aspect of an item.[1]

*3.1.2 Activity Identification.* In this step, the goal is to classify sentences that mention some item-related activity or usage. Inspired by Benetka et al. [2], our approach revolves around using *part-of-speech* (POS) analysis and rules of the English language. We filter for the preposition *for* followed by a verb in progressive tense heuristically, by looking for *-ing* endings (e.g., *for commuting*, *for hiking*). Note that there might be other formulations that describe activity or usage. Our goal is not to extract all possible sentences containing mentions of activity or usage; a high recall approach would likely come at the cost of a larger fraction of false positives. Instead, we focus on achieving high precision.

## 3.2 Question Generation

The main motivation for this step is generating natural-sounding questions that are easy for users to understand and answer, without needing any additional context. Consider the sentence *The fat tires are perfect for conquering tough terrain.* An example of converting it to a yes or no usage-related question might be *Would you like a bike that is great for conquering tough terrain?* Note that not all candidate sentences that pass our selection heuristic are viable for conversion to a question, e.g., *Thank you so much for coming up with such a great product.* This sentence is too vague and does not mention any action or usage for the item, and thus should be labeled as not applicable (N/A).

Learning to generate questions is done by fine-tuning a large, pre-trained, sequence-to-sequence language model. There are two main benefits of using transfer learning from a pre-trained model. First, it increases the learning speed; as both syntax and semantics of the English language are already learned, there are fewer things the model needs to learn. Second, it reduces the amount of labeled data needed to train models to high performance. Specifically, we use T5 [24] as it is a state-of-the-art approach that can be used for both N/A-classification and generation in one go, where N/A-classification is posed as a text-to-text problem. Obtaining high-quality labeled data for fine-tuning the model is a challenge on its own; we develop a multi-step data collection protocol using crowdsourcing, which we discuss in Section 4.2. In our experiments (in Section 5), we evaluate our question generation models using different number of parameters for pre-training and varying amount of training data for fine-tuning.

# 4 DATA COLLECTION

This section describes the process of creating our dataset, which consists of a set of review sentences and either (i) a corresponding set of five preference elicitation questions or (ii) the label not applicable (N/A) for each.

## 4.1 Candidate Sentence Selection

The starting point for getting the candidate sentences are the Amazon review and metadata datasets [22],[2] where item reviews from Amazon are extracted along with product metadata information such as *title*, *description*, *price*, and *categories.* There are, in total, 233.1M reviews about 15.5M products. Due to the sheer dataset size, we focus our research on three main categories: *Home and Kitchen*, *Patio, Lawn and Garden*, and *Sports and Outdoors.* From these, we further sub-select 12 diverse subcategories (simply referred to as *categories* henceforth): *Backpacking Packs*, *Tents*, *Bikes*, *Jackets*, *Vacuums*, *Blenders*, *Espresso Machines*, *Grills*, *Walk-Behind Lawn Mowers*, *Birdhouses*, *Feeders*, and *Snow Shovels.*

Sentence splitting and *aspect-value* pair extraction is performed using the Sentires toolkit [33, 34].[3] This step discards many non-viable sentences. The remaining ones are POS-tagged using the Stanford NLP toolkit [20]. Finally, we filter for sentences that match our activity detection heuristic (*"for + [verb in progressive tense]"*). Our sentence selection process is designed to favor precision over recall, and was validated by manual inspection of the results. Upon completion of the crowdsourcing tasks (described in Section 4.2), we find that over 75% of the selected sentences can be turned into questions. This shows that our simple method can indeed identify candidate sentences with high precision.

Our final *candidate sentence set* contains 100 sentences for each of the categories, except *Birdhouses*, where only 15 candidate sentences are found due to the size of that category, that is, a total of 1,115 sentences over 12 categories.

## 4.2 Question Generation using Crowdsourcing

Crowdsourcing was done on the Amazon Mechanical Turk (AMT) platform in three steps. The task was available to workers with 95% approval rate and with at least 1,000 approved human intelligence tasks (HITs).

*4.2.1 Step 1: Question Collection.* Crowd workers are given a review sentence (describing some aspect or use for a product) and a

[1] This concerns future utilization of responses given to these elicitation questions, where the CRS might want to map activities to specific attribute values.

[2] https://nijianmo.github.io/amazon/index.html

[3] https://github.com/evison/Sentires

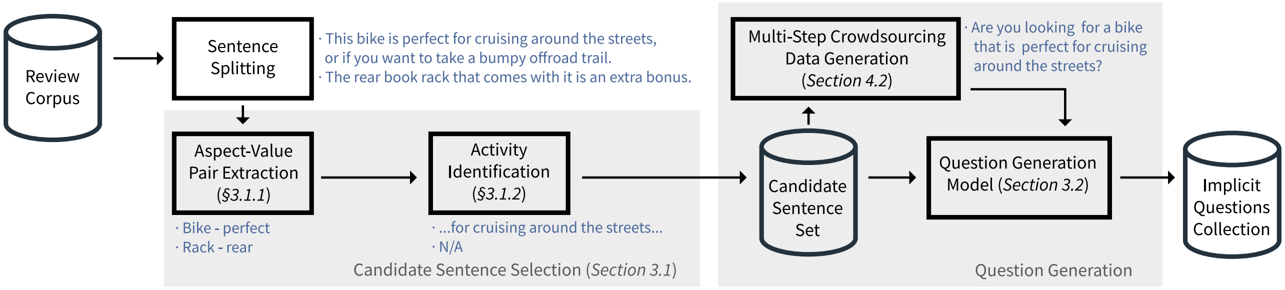

Review Corpus
Sentence Splitting
· This bike is perfect for cruising around the streets, or if you want to take a bumpy offroad trail.
· The rear book rack that comes with it is an extra bonus.
Aspect-Value Pair Extraction (§3.1.1)
· Bike - perfect
· Rack - rear
Activity Identification (§3.1.2)
· ...for cruising around the streets...
· N/A
Candidate Sentence Selection (Section 3.1)
Multi-Step Crowdsourcing Data Generation (Section 4.2)
· Are you looking for a bike that is perfect for cruising around the streets?
Candidate Sentence Set
Question Generation Model (Section 3.2)
Question Generation
Implicit Questions Collection

**Figure 2: Components of our question generation system.**

product category as input, and tasked with rewriting it into a question or marking it as not applicable. They are specifically instructed to formulate a question that a salesperson or a recommender agent might ask a customer, such that it is a standalone question that can simply be answered with yes/no. For every input sentence, we collected responses from three different workers. Sentences found non-applicable by at least two workers are set as N/A. The task was re-run if a single worker responded with N/A. This process resulted in approx. 2,600 sentence-question pairs.

*4.2.2 Step 2: Validation and Filtering.* Next, we validate all responses (i.e., generated questions) for applicable sentences collected in Step 1 using crowdsourcing. We employ three different workers in Step 2, who are requested to answer four multiple-choice questions: (1) *Is the question grammatically correct?* [Yes/No] (2) *Can the question be answered by yes or no?* [Yes/No] (3) *Does the question mention any trait or use for a product?* [Yes/No] (4) *Who is most likely to ask this question in a sales setting?* [Buyer/Salesperson/Neither]. Generated questions that are found invalid by all three workers on a single aspect or at least two workers on at least two aspects are automatically rejected. Those that are marked invalid on multiple aspects but do not fall into the former category are manually checked by an expert annotator (one of the authors). All other questions are approved. Steps 1 and 2 were run multiple times until all questions were approved.

*4.2.3 Step 3: Expanding Question Variety.* Our main motivation for expanding the question variety is to add new ways of asking implicit questions. To this end, we task a new set of workers to paraphrase the questions we obtained and validated in Steps 1 and 2. Each worker receives all three versions of the questions from Step 1 as input and is asked to produce a new (paraphrased) question the expresses the same meaning. Note that this set of workers do not get to see the original sentences, only the questions generated from them by other workers. For each set of three questions, two additional paraphrases were collected. Considering that generating paraphrases proved to be a much simpler task than generating questions from review sentences, no additional quality assurance steps were necessary.

### 4.3 Final Dataset

Out of the 1,115 candidate sentences, 277 were labeled as non-applicable (not containing relevant usage-related information), which is below 25%. This shows that our high-precision approach to selecting candidate sentences is effective. We note that our sentence selection method works better for some categories than for others. The fraction of viable sentences ranges between 52% (*Espresso machine* category) to 84% (*Backpacking pack* category). For the remaining 838 sentences, a total of five questions are generated, three based on the candidate sentence and two via paraphrasing. Table 1 shows two example sentences from our dataset. The total cost of generating the dataset was $1,200.

## 5 RESULTS AND ANALYSIS

With our experiments, we aim to answer the following research question: How effective is our method for generating implicit questions for preference elicitation? Specifically, given a candidate sentence as input, our approach should either generate a question or label it as non-applicable (N/A), if a usage-related question cannot be generated.

### 5.1 Experimental Setup

We train *small*, *base*, and *large* T5 models, which vary in the number of layers, self-attention heads, and the dimension of the final feedforward layer. The difference is shown in the number of parameters in Table 2. We use 80% of the data for training, while the rest is test data. In our training, we employ teacher forcing [29], regularization by early stopping [21], and adaptive gradient method AdamW [16] with linear learning rate decay. For each sentence, we have either N/A or a set of reference questions. We evaluate question generation both as a classification task, in terms of Accuracy (detecting N/A), and as a machine translation task, where the set of human-generated questions serve as reference translations. Specifically, we report on BLEU-4, which uses modified n-gram precision up to 4-grams [23], and ROUGE-L, a recall-based metric based on the longest common subsequence [15].

### 5.2 Results

Table 2 shows the results in terms of non-applicability classification (Accuracy) and question generation (BLEU and ROUGE). On both tasks, larger pre-trained models tend to perform better, which is expected. The difference, however, is more pronounced for non-applicability detection than for question generation.

We further investigate how the amount of training data affects model performance, by considering different ways of data reduction. We use the best performing model for this experiment, i.e., *T5 large*.

**Table 1: Example sentence-question pairs from our dataset.**

| **Category** | Blender | Snow shovel |
|---|---|---|
| **Sentence** | Great for making smoothies with frozen fruit. | This product is excellent for doing the job. |
| **Generated questions** | - Are you looking for a blender that's great for making smoothies with frozen fruit?<br>- Would you be interested in a blender that is great for making smoothies with frozen fruit?<br>- Are you interested in a blender for making smoothies with frozen fruit? | N/A<br>(The input sentence passes our candidate selection heuristic, but the activity is too broad and can apply to any item.) |
| **Paraphrases** | - Do you want a blender that's great for making smoothies with frozen fruit?<br>- Would you like a blender that is great for making smoothies with frozen fruit? | |

**Table 2: Question generation performance using different pre-trained models that are fine-tuned on all available training data (i.e., five questions or N/A per sentence). Best scores for each measure are in boldface.**

| Model | #Parameters | Accuracy | BLEU-4 | ROUGE-L |
|---|---|---|---|---|
| T5-small | 60.5 M | 0.76 | 0.56 | 0.69 |
| T5-base | 222 M | 0.84 | 0.53 | 0.66 |
| T5-large | 737 M | **0.88** | **0.58** | **0.71** |

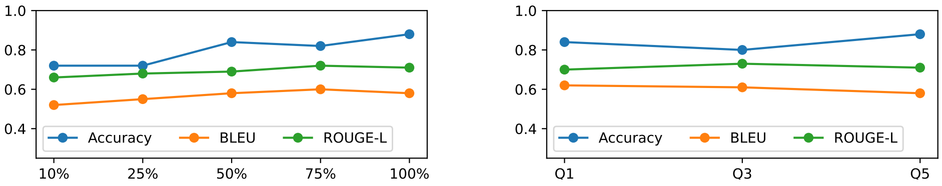


**Figure 3: Model performance (T5-large) with sentence-based (Left) or question-based (Right) training data reduction.**

In *sentence-based* data reduction, shown in Fig. 3 (Left), only a subset of the available sentences is used for training (using all available questions corresponding to those sentences). We observe a drop in Accuracy when we reduce the amount of training data to 25% or lower, while question generation performance is less severely affected. In *question-based* data reduction, shown in Fig. 3 (Right), we split the dataset based on the number of questions available for each sentence. We consider using a single question (Q1), the three initially generated questions (Q3), and the three initial questions plus the two paraphrases (Q5). We find that reducing the number of questions has surprisingly little effect. This suggests that it is more beneficial to collect a small number of questions for a larger set of sentences than vice versa.

## 6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have studied the question of how a conversational recommender system can solicit user's needs through natural language by using *indirect* questions about how the product wanted will be used. This contrasts with most prior work that considers how to directly ask about desired product attributes. Our method starts with a corpus of reviews, then identifies statements that characterize how products are used, and how this ties to product attributes. These statements are then transformed into preference elicitation questions. We show that our approach effectively selects such statements (with high precision), and transforms them into effective questions.

We emphasize that this work focuses on this first stage of recommendation, understanding the user's needs, and doing this in an engaging way. The most important future direction is determining how answers to these questions should best be applied to the recommendation task once the user's need is understood. Here, we believe that sentence embedding techniques are likely to be effective. Second, as this work builds on top of large language models, language safety is a key consideration warranting further study before our approach could be used in practice. Nevertheless, during experimentation we did not observe concerning language nor hallucinations. We also note that the offline question generation process lends itself well to even manual control over the language model output.

## REFERENCES

[1] Krisztian Balog, Filip Radlinski, and Alexandros Karatzoglou. 2021. On Interpretation and Measurement of Soft Attributes for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 890–899.

[2] Jan R. Benetka, John Krumm, and Paul N. Bennett. 2019. Understanding Context for Tasks and Activities. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. 133–142.

[3] Pawel Budzianowski, Tsung–Hsien Wen, Bo–Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ – A Large–Scale Multi–Domain Wizard–of–Oz Dataset for Task–Oriented Dialogue Modelling. arXiv:1810.00278

[4] Li Chen and Pearl Pu. 2012. Critiquing–based recommenders: Survey and emerging trends. *User Modeling and User–Adapted Interaction* 22 (2012).

[5] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. 2018. Q&R: A Two–Stage Approach toward Interactive Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 139–148.

[6] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 815–824.

[7] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat–Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. arXiv:2101.09459

[8] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *The 41st International ACM SIGIR Conference on Research &*

*Development in Information Retrieval (SIGIR '18)*. 1371–1374.

[9] Rolf Jagerman, Ilya Markov, and Maarten de Rijke. 2019. When People Change Their Mind: Off–Policy Evaluation in Non–Stationary Recommendation Environments. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. 447–455.

[10] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5, Article 105 (2021), 36 pages.

[11] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta–Learned User Preference Estimator for Cold–Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. 1073–1082.

[12] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive Path Reasoning on Graph for Conversational Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. 2073–2083.

[13] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. arXiv:1812.07617

[14] Yuan–Hong Liao, Amlan Kar, and Sanja Fidler. 2021. Towards Good Practices for Efficiently Annotating Large–Scale Image Classification Datasets. arXiv:2104.12690

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.

[16] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. arXiv:1711.05101

[17] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. 2011. Automatic Construction of a Context–Aware Sentiment Lexicon: An Optimization Approach. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. 347–356.

[18] Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent Linear Critiquing for Conversational Recommender Systems. In *Proceedings of The Web Conference 2020 (WWW '20)*. 2535–2541.

[19] Kai Luo, Hojin Yang, Ga Wu, and Scott Sanner. 2020. Deep Critiquing for VAE–Based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1269–1278.

[20] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 55–60.

[21] N. Morgan and H. Bourlard. 1989. Generalization and Parameter Estimation in Feedforward Nets: Some Experiments. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems (NIPS'89)*. 630–637.

[22] 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)*. Association for Computational Linguistics, 188–197.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text–to–Text Transformer. arXiv:1910.10683

[25] Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference Elicitation as an Optimization Problem. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. 172–180.

[26] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 235–244.

[27] Qing Wang, Chunqiu Zeng, Wubai Zhou, Tao Li, Larisa Shwartz, and Genady Ya. Grabarnik. 2017. Online Interactive Collaborative Filtering Using Multi–Armed Bandit with Dependent Arms. arXiv:1708.03058

[28] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2021. Denoising Implicit Feedback for Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. 373–381.

[29] Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Comput.* 1 (1989), 270–280.

[30] Ga Wu, Kai Luo, Scott Sanner, and Harold Soh. 2019. Deep Language–Based Critiquing for Recommender Systems. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. 137–145.

[31] Wei Wu and Rui Yan. 2019. Deep Chit–Chat: Deep Learning for Chatbots. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. 1329.

[32] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 177–186.

[33] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase–Level Sentiment Analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. 83–92.

[34] Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014. Do Users Rate or Review? Boost Phrase–Level Sentiment Labeling with Review–Level Sentiment Classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. 1027–1030.

[35] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. 2013. Interactive Collaborative Filtering. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. 1411–1420.