Performance Predictors for Conversational Fashion Recommendation

Maria Vlachou[1], Craig Macdonald[1]

[1]University of Glasgow, UK

Abstract

In Conversational Recommendation Systems (CRS), a user can provide natural language feedback on suggested items, which the recommender uses to produce improved suggestions. Therefore, the success of a user's conversation with the CRS is determined by how well the system is able to interpret the user's feedback and the quality of the recommendations. Knowing whether a conversation is likely to be successful may allow the CRS to adjust accordingly - for instance, changing its retrieval strategy, or asking a clarifying question. Existing work on Query Performance Prediction (QPP) has examined a number of predictors that indicate the effectiveness of a search engine's ranking in response to a query. Inspired by existing work in QPP, we propose a framework for Conversational Performance Prediction (CPP) that aims to predict conversation failures by considering the recommendation ranking at different turns of a conversation, either one turn at a time, or by considering multiple consecutive turns. In this regard, we adapt post-retrieval predictors to address the multi-turn nature of the CRS task. We conduct our analysis on Shoes and FashionIQ Shirts & Dresses datasets. In particular, as a ground truth, we measure conversation difficulty by the effectiveness of the ranking at a given turn of the conversation. Overall, we find some promise in score-based retrieval predictors for CPP, obtaining medium strength correlations with conversation difficulty - for instance, observing a Spearman's ρ of 0.423 on the Shoes dataset, which is comparable to correlations observed for standard QPP predictors on adhoc search tasks.

1. Introduction

Traditionally, Recommender Systems (RS) help users to find items of interest on the basis of user feedback in terms of ratings, clicks or reviews. In contrast, *Conversational Recommendation Systems (CRS)*, such as personal digital assistants [1], have facilitated more complex recommendation settings by suggesting items in response to voice or (natural language) chat interactions. In particular, a CRS allows a multi-turn dialogue with users and aim to assist them with achieving a number of task-oriented goals [2]. Indeed, at each turn users can provide their feedback or *critique* [3], which helps the system to improve recommendations [4].

One important aspect of natural language-based CRS is that they allow users to explore the range of available options and elicit their preferences. For example, Bursztyn et al. [7] created a multi-modal system, where users navigate in a setting of limited options, such as finding a restaurant near their location. In this setting, users start exploring an initial set of restaurants and have the opportunity to see their details by clicking through the options, while they are asked about the reasons for any negative feedback they provide. Another example of user exploration is the MusicBot [8, 9], a music chatbot that first collects users' preferences and then makes suggestions based on different techniques of critiquing the song recommendations. In our work, we are focused upon conversational fashion image recommendation [5, 6, 10], an example of which is shown in Figure 1. In this task, the user has a target item in mind, and provides textual feedback (critiques) to direct the system towards retrieving images of fashion products that are more similar to to their perceived target item.

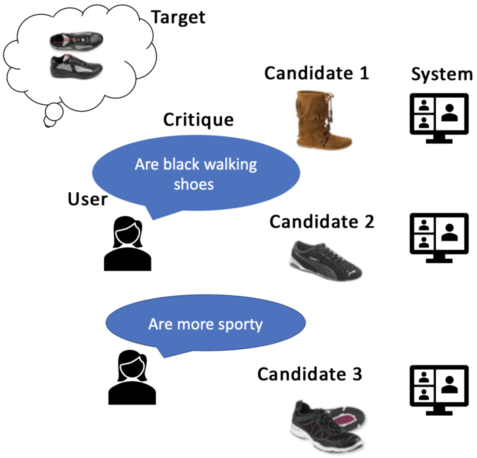


Figure 1: Example of Dialog-based recommendation in CRS. Pictures and dialogues from the Shoes dataset [5, 6].

However, not all conversations may lead to a satisfying outcome for the user. This can be easily quantified in offline evaluation scenarios, where the CRS is evaluated across a pre-defined number of turns. For example, Yu et al. [11] found that, although users had the option to

4th Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2022, September 18–23 2023, Seattle, WA, USA.
m.vlachou.1@research.gla.ac.uk (M. Vlachou); craig.macdonald@glasgow.ac.uk (C. Macdonald)

explore a list of options for a number of turns, the system was unable to find a relevant recommendation by turn 7, which might mean that the algorithm was still exploring the space. Also, in Wu et al. [10], the target item was found by the system at rank 1 in only 42% of conversations after (a maximum of) 10 turns. Therefore, exploration might result in an increased number of turns, which on one hand might mean more engaged users [8], but at the same time suggests that often the conversations might fail (i.e., target item not found). In this regard, we are interested in identifying indicators that can detect how this happens – for example, a conversation could fail because the system is unable to find the target item or because the target item is not available.

In what follows, inspired by existing work on Query Performance Prediction (QPP) (e.g., [12, 13, 14, 15]), we aim to predict conversational failures by identifying specific indicators that are correlated with failure. In particular, we aim to determine the quality of multi-turn critiquing-based CRS recommendation by proposing predictors that consider the multi-turn aspect of conversational recommendation. The proposed predictors address characteristics of the retrieved scores of the top-recommended items and can predict poor performance across a shorter or longer number of turns in the conversation, which we call *prediction horizons*. In summary, this work makes the following contributions: (i) We propose a framework for Conversational Performance Prediction (CPP), which extends the existing work on QPP to a conversational recommendation setting; (ii) We show how to adapt QPP evaluation methodology to a multi-turn conversational setting which allows to evaluate CPP predictors for both short- and long-term prediction horizons; (iii) We evaluate some of our proposed predictors on the Shoes [5, 6] dataset and the Fashion IQ Dresses and Shirts categories [16], using a state-of-the-art user simulator [6]. The rest of the paper is structured as follows: Section 2 presents the existing research on QPP, including pre- and post-retrieval predictors, as well as their probabilistic interpretation; Section 4 outlines our new proposed framework and predictors; Section 5 describes our experimental setup; Sections 6 & 7 present our results and provide concluding remarks.

2. Related Work

In order to predict why a conversation with a CRS might fail, we need to identify indicators that show when the user is unable to find the target item during the interaction. In this regard, we are inspired by existing work from *Query Performance Prediction (QPP)*, which we discuss in Section 2.1; Later, in Section 2.2, we discuss applications of QPP in conversational contexts.

2.1. Query Performance Prediction

Traditionally, QPP is used to predict the effectiveness of a search results page performed in response to a query in the absence of human relevance judgments [15]. It has applications to selective retrieval approaches [17, 18] and query features for learning-to-rank [19], to name but a few. *Query performance predictors* are generally grouped into pre-retrieval, and post-retrieval, which we discuss further below.

2.1.1. Pre-retrieval Query Performance Predictors

Pre-retrieval predictors are used to estimate the performance of queries before the retrieval stage, and therefore, are independent of the search performed and the ranked list of results [14]. This means that pre-retrieval predictors base their predictions on properties of query-terms or corpus-based statistics [12, 13, 14, 20, 21, 22]. Examples of pre-retrieval predictors that describe the statistical properties of the query terms or the corpus include the *query length* (number of non-stop words in the query), the *standard deviation of the inverse document frequency* of the query terms, the simplified query clarity score *(SCS)*, which measures the occurrence of a query term in the query relative to its occurrence in the collection, and AvICTF, which considers the overall informativeness of the query terms using the collection model [23]. Another class of pre-retrieval predictors refers to linguistic features of the queries, such as *syntactic complexity* (distance between syntactically linked words) and *word polysemy* (number of semantic classes a word belongs to) [22]. Overall, with limited information available before retrieval commences, pre-retrieval predictors are widely considered less accurate for performance prediction than post-retrieval predictors [14].

2.1.2. Post-retrieval Query Performance Predictors

On the other hand, *post-retrieval* predictors are applied on the list of the top-ranked retrieved documents, and therefore use the relevance scores or the (textual) contents of the returned items. A first group of post-retrieval predictors examines the difference of the result list from the corpus, or the *focus of the result list*. For example, *Clarity* [12] measures the focus of the resulting ranking with respect to the corpus using the KL divergence between their respective language models, while the *Weighted Information Gain (WIG)* corresponds to the difference between the average retrieval score of the result list and of that of the corpus [24]. A second group includes the distribution of the retrieval scores of the top-ranked items. Predictors in this group include *Normalized Query Commitment (NQC)* [25] (the standard deviation of the retrieval scores in the result list). The standard deviation is considered to be negatively correlated with the amount of query drift (the non-related information in the result list) [26]. Also, this group includes the modeling of retrieval scores; the top-ranked items could be modeled as a certain mixture of distributions corresponding to relevant and non-relevant items [27]. Another related predictor is *autocorrelation* [28], which assumes that documents whose vector space embeddings are closely related receive similar scores, and therefore, closely related scores would indicate similar performance.

A third group of post-retrieval predictors refers to the relation of the top-ranked retrieval scores with a particular reference list. Recently, a more generalised approach for estimating the effectiveness of a ranking was proposed, based on the assumption that high association with *pseudo-effective* reference lists and low association with *pseudo-ineffective* lists improves effectiveness [29]. One example refers to the *utility estimation framework (UEF)* [30], which estimates the utility of a given ranking with respect to how much it represents an underlying information need [31]. The utility is estimated by the expected similarity between a given document ranking and those induced by estimates of relevance language models (these rankings are assumed to be representative of the information need). [32]. A similar predictor to the UEF approach is *query feedback (QF)* [24], which measures the overlap of top items between the result list and a reference list retrieved from the corpus using a language model induced from the result list. Autocorrelation [28] could also fall under this category, if we compare the result list of the original retrieval scores with a reference list that contains either a perturbed version of the scores diffused in space or a list with the averaged values from multiple retrievals for the same query. Lastly, an inter-list similarity predictor is a measure of *rank-biased overlap (RBO)*, which measures the expected average overlap between two rankings [33] and can be applied to the QPP task.

Finally, we note some recent QPP work (e.g. [29, 34]) has focused upon probabilistic frameworks for QPP, which can integrate both pre-retrieval and post-retrieval predictors. However, many of the underlying intuitions encapsulated by these frameworks are already addressed in the previously described predictors.

2.2. Query Performance Prediction in Conversational Search

Natural language-based conversational systems allow users to express complex feedback through a dialogue, thus resulting in more natural interactions [35]. To be able to predict the likelihood of success of a conversation, we need to consider the salient aspect of the conversational setting, such as the users' feedback and the iterative *turn*-based nature of the interaction process.

However, while QPP has been widely explored for (single turn) queries in search settings, the area of conversational search or recommendation has seen much less work. For example, one recent work examines the predicted effectiveness of the top-retrieved documents for deciding to generate clarifying questions, and specifically some extracted features, such as noun phrases or named entities [36]. Indeed, clarifications are useful for both the user and the system [37, 38, 39]. Also, Roitman et al. [40] examined a constrained retrieval setting, namely the interaction with a conversational assistant, where the assistant needs to decide whether the provided answer could be accepted. The authors built a classifier that determines the answer quality by adapting some existing QPPs to the answer level (using the score of the top item, which is provided as the answer).

However, QPP for conversational recommendation has not been addressed. In particular, we are interested in creating a prediction framework for identifying poorly performing or failed conversations in a recommendation setting. We postulate that these predictors can be useful in several use cases, for instance knowing when to ask for clarifications, or when the users target item cannot be found. Towards achieving this goal, we explore score-based predictors, adapting to the multiple turn nature of the task. In the next section, we define the CRS task; Later in Section 4, we define our CPP framework.

3. Conversational Image Recommendation

Figure 1 describes the context of dialog-based image recommendation in a CRS. At each interaction turn, the user provides a critique of the current recommendation (candidate item) back to the system, aimed at directing it towards the desired target item. More formally, at a given interaction turn k, the user provides textual feedback f_k on the current top-ranked candidate item $i_{k,1}$. Based on this feedback, the conversational recommendation system $\mathcal{C}()$ provides a new ranking , i.e.: $\mathcal{C}(i_{k,1}, f_k) \rightarrow S_k$, where S_k is a ranking of n items with corresponding descending retrieval scores $s_1 \ldots s_n$, i.e.: $S_k = [\langle i_{k+1,1} s_1 \rangle, \ldots \langle i_{k+1,n}, s_n \rangle]$.

However, it is challenging to train and evaluate a natural language-based CRS. For training, *reinforcement learning (RL)* is widely used, as it allows optimising the recommendation model based on the long-term rewards [41], i.e. based not just on retrieving the correct item in any current iteration, but also retrieving it in later iterations. However, such a model needs to be trained while interacting with an environment, and obtaining many samples is hard by relying on real users [41, 42]. For evaluation, ideally human users are needed to judge the system's efficiency and user satisfaction [43]. Instead, user simulators are deployed as surrogates for human users, trained on *relative caption* data - a form of human-annotated dialogues on pairs of images. Recommendation models trained and evaluated using user simulators have been found to be correlated with human satisfaction [6].

Specifically, for the purposes of training a user simulator with human-annotated dialogues, Guo et al. [6] proposed the *relative captioning* task. In this task, human annotators recruited through crowdsourcing are placed in a context of online shopping, where the CRS acts as the shopping assistant and they play the role of the customer. During the process, annotators are presented with *candidate* recommended images of items and they are asked to provide single instance critiques. In each interaction round, they are shown a given *candidate* item and based on a given *target* item, they provide a critique on the current candidate item. These differences between the candidate and the target image are described with natural language phrases and form the *relative captions*. Hence, a relative captioning dataset contains tuples of the following form: $\langle i_t, i_c, tq_{c,t}\rangle$ where i_t is a representation of the target item (for instance an image), i_c is the current candidate item being presented to the user and $tq_{c,t}$ is the critique by the user on the candidate, intended to direct the system more towards the target. Relative captioning data can be used to train a user simulator, which is then deployed for training or evaluating a CRS [6, 10, 11, 16, 44, 45].

Using a user simulator for evaluation, the overall success of a CRS system can be reliably measured, in a offline Cranfield-like setting, by using ranking evaluation measures, such as NDCG, upon the ranked list of recommendations produced at each turn. From such an evaluation, it can be seen that even after 10 turns, some CRS models may not be able to identify the target item for some conversations. For this reason, making a prediction as to the likelihood of a user being satisfied with a conversation may have utility to improving the user experience. In the next section we introduce our proposal for conversation performance prediction for CRS.

4. Performance Prediction in Conversational Recommendation

Our aim for conversational performance prediction differs from existing approaches on QPP in a number of ways. While QPP focuses on estimating the relevance of a ranking to a given single query (single-turn), to predict the user's satisfaction of a conversation, we need to take into account the nature of the task, which is to consider the ranking quality across multiple turns. Another important difference is that many QPP techniques are based on textual queries and textual documents. In contrast, in our fashion-based CRS, the "units of retrieval" are images, with embedded representations - this precludes the use of textual content-based predictors. Furthermore, our "query units" are critiques, which are based on the retrieval of the previous turn. Therefore, it can be seen that there is no clear distinction between pre-retrieval and post-retrieval predictors, since what is considered post-retrieval of one turn could be seen as a pre-retrieval predictor of the following turn. For this reason, we propose a new framework for performance prediction in a conversational setting, in particular conversational fashion retrieval, which we describe in Section 4.1 below. Later in Section 4.2, we describe the initial score-based predictors we can adapted to this framework.

4.1. CPP Framework

We present a framework for Conversational Performance Prediction (CPP) applied to the domain of fashion recommendation for image retrieval [6, 16]. In this regard, we define recommendation success as the identification of the target image item by the system before a maximum number of turns is reached, which corresponds to a user being satisfied with the conversation. More formally, the CPP task can be described as a function of the form

$$CPP(F, S) \to \mathbb{R}$$

where F is a sequence each containing f feedback critiques over 1 or more turns, and S is a sequence of results lists consisting of retrieval scores, over 1 or more turns.

This framework can be instantiated for single-turns, or multiple turns. For instance, in a single-turn setting, we can instance CPP task at a given turn k, i.e.:

$$CPP_{\text{single}}([f_k], [s_k]).$$

On the other hand, for two consecutive turns, k and $k+1$, prediction takes the following form:

$$CPP_{\text{consecutive}}([f_k, f_{k+1}], [s_k, s_{k+1}]).$$

Overall, from the above different formulations, it is clear that CPP is a distinct task from QPP that can be addressed by different families of predictors. In this initial work, we adapt one category of score-based QPP predictors into the CPP framework, which we discuss further below.

4.2. Score-based Predictors for CPP

In this work, we are inspired by post-retrieval predictors that study the distributions of retrieval scores and the use of reference lists, as introduced in Section 2.1.2. In particular, we have the following initial intuitions concerning successful interactions in the CRS task:

Table 1
Proposed CPP predictors according to number of turns involved.

Single-turn	Consecutive Turns
Top-1 item score (maximum score)	Difference in maximum score
Mean score of top-n items	Overlap of top-ranked items
Standard deviation (sd) of top-n items	

- For a single turn, if the score of the top-ranked item(s) is high, then the system has a clear representation of the user's desired item, and it can find item(s) that closely matches that representation.
- In a successful conversation, the scores of the top-ranked item(s) will increase across multiple turns, as the system becomes more confident in its predictions.
- In a successful conversation, the retrieved items become more similar across turns as the system becomes more confident in its predictions and focuses on the correct part of the item catalogue.

Adapting the notation of Section 4.1 to disregard the feedback sequences, we define a number of score-based CPPs, for single turns – in the form of $CPP([s_k])$ and for consecutive turns – $CPP([s_k, s_{k+1}])$. All predictors are described in Table 1. For instance, top-1 denotes the maximum score of any retrieved item, while mean denotes the average of the scores of the retrieved items. When applying these predictors, we also denote the turn k that the predictor is calculated, i.e. top-1@k is the maximum score of any item retrieved in the ranking produced for turn k. In the remainder of this paper, we evaluate these predictors on several conversational fashion recommendation datasets.

5. Experimental Setup

We now experiment to address salient aspects upon both the nature of the predictors (single-turn and consecutive turn), as well as upon the accuracy of the predictors on different prediction *horizons*, i.e., at what point can a prediction be made, and how does it correspond to the effectiveness of the CRS, as measured at a later turn. In particular, we measure *short-term* horizons (i.e., can we predict the effectiveness of the next turn?); and *long-term* horizons (i.e., can we predict the effectiveness of the last turn); as well as measuring the *longevity* of the prediction (i.e., how useful is an early prediction?). Focusing initially on single-turn predictors, our first research question is:

RQ1 Can we predict conversation performance with predictors based on retrieval scores of a single turn, in terms of (a) long-term and (b) short-term prediction, as well as (c) longevity?
Secondly, we consider the consecutive-turn predictors:

RQ2 Can we predict conversation performance with predictors based on (a) differences in retrieval scores between consecutive turns and (b) overlap in retrieved items of two consecutive turns?

To evaluate our CPP approaches, we use the Shoes dataset [5, 6], which contains one relative critique (describing relative differences between recommended and target image pairs) for pairs of shoe images, and the Dresses & Shirts categories of the Fashion IQ dataset [16], which contains two relative captions per candidate-target pair.

For a CRS, we apply a supervised GRU sequential recommendation model [6, 46], which is trained using triplet loss and uses the natural language feedback and the previous recommended images as input, thus maximizing short-term rewards. To train our recommendation model, we use a recently developed user simulator for dialog-based interactive image retrieval based and the relative captioning task [6]. The GRU model is configured to retrieve 100 items at each turn.

In QPP, the accuracy of predictors is evaluated at the query level (a given query is easy or difficult compared to other queries in a set). Specifically, a ranking of queries by the effectiveness of a system, i.e., in terms of Mean Average Precision (ground truth) is correlated with a ranking induced by a predictor. In contrast, we evaluate CPP predictors at the conversation level (across multiple dialog turns). Consequently, for the ground truth, we evaluate the effectiveness of each *conversation* at identifying the user's target item – more specifically, by considering the rank of the target item at a specific turn of the conversation. Following existing CRS work [6, 10, 11, 16, 44], we set the maximum number of turns to be 10.

In this regard, for our proposed single-turn predictors in Table 1, we use three different ground truth settings: the rank of the target item at the end of the conversation (turn 10); the rank of the target item during the conversation, i.e. at a given turn k; and the rank of the target item directly after the prediction is made (i.e. $k + 1$ for a prediction at turn k). Through these different ground truth settings, we can measure CPP accuracy at both short-term and long-term horizons, as well as their longevity.

Finally, for quantifying the correlations, we report Spearman's ρ. Significance testing is achieved by examining the p-value associated with ρ, which indicates the probability of an uncorrelated ranking producing a Spearman correlation as high as that observed.[1]

6. Results

In this section we report experiments for score-based CPP predictors, for single-turn (Section 6.1) and consecutive-turn (Section 6.2) scenarios.

[1] See also https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Table 2

Results of single-turn predictors for short and long-term prediction of rank of target items at various turns. * denotes significant correlations; for Shoes, all correlations are significant, so * is omitted ($p < 0.05$). In the first group of columns, bold values denote the maximum correlation over all turns for the same predictor and the same ground truth value. For the other two sets of columns, bold values denote the highest performing predictor of the three examined single-turn predictors in the given evaluation setting for each turn – this is because comparison of correlation values across turns (rows) is not possible, since the ground truth changes for each row.

Prediction at turn k with rank@turn10				Prediction at turn 2 with rank@turn k				Prediction at turn k with rank@turn $k+1$			
k	top-1@k	mean@k	sd@k	rank@k	top-1@k	mean@k	sd@k	k, rank@k	top-1@k	mean@k	sd@k
Shoes											
2	-0.144	-0.141	-0.081	2	**-0.405**	-0.385	-0.059	2,3	**-0.423**	-0.413	-0.201
3	-0.145	-0.145	-0.097	3	**-0.423**	-0.413	-0.201	3,4	**-0.356**	-0.355	-0.254
4	-0.148	-0.148	**-0.105**	4	**-0.357**	-0.349	-0.183	4,5	**-0.318**	-0.317	-0.211
5	-0.155	-0.153	-0.089	5	**-0.314**	-0.309	-0.177	5,6	**-0.293**	-0.292	-0.180
6	-0.165	-0.165	-0.093	6	**-0.270**	-0.267	-0.163	6,7	**-0.254**	-0.254	-0.135
7	-0.173	-0.173	-0.100	7	**-0.230**	-0.226	-0.140	7,8	**-0.235**	-0.234	-0.126
8	-0.178	-0.177	-0.073	8	**-0.213**	-0.210	-0.136	8,9	**-0.208**	-0.207	-0.067
9	**-0.184**	**-0.183**	-0.064	9	**-0.175**	-0.173	-0.1149	9,10	**-0.183**	-0.183	-0.064
10	-0.183	-0.181	-0.026	10	**-0.144**	-0.141	-0.081				
Dresses											
2	0.012	0.003	-0.036	2	**-0.281***	-0.279*	-0.161*	2,3	-0.248*	**-0.256***	-0.197*
3	-0.017	-0.015	-0.004	3	-0.248*	**-0.256***	-0.197*	3,4	**-0.262***	-0.257*	-0.075*
4	-0.045*	-0.047*	-0.014	4	-0.187*	**-0.198***	-0.173*	4,5	**-0.246***	-0.239*	-0.038
5	-0.055*	-0.051*	-0.007	5	-0.128*	**-0.140***	-0.137*	5,6	**-0.206***	-0.198*	-0.008
6	-0.063*	-0.063*	**-0.041***	6	-0.079*	-0.092*	**-0.102***	6,7	**-0.172***	-0.168*	-0.034
7	-0.069*	-0.072*	-0.033	7	-0.052*	-0.067*	**-0.091***	7,8	-0.139*	**-0.142***	-0.044*
8	-0.075*	-0.076*	-0.021	8	-0.039	-0.051*	**-0.072***	8,9	**-0.103***	-0.101*	-0.000
9	-0.073*	-0.071*	-0.018	9	-0.005	-0.018	**-0.053***	9,10	**-0.073***	-0.071*	-0.018
10	**-0.080***	**-0.078***	0.003	10	0.0127	0.003	**-0.036**				
Shirts											
2	-0.092*	-0.089*	**-0.074***	2	**-0.305***	-0.298*	-0.141*	2,3	-0.297*	**-0.305***	-0.201*
3	-0.124*	-0.119*	-0.033	3	-0.297*	**-0.305***	-0.201*	3,4	**-0.336***	-0.326*	-0.03*
4	-0.145*	-0.137*	0.011	4	-0.264*	**-0.273***	-0.192*	4,5	**-0.323***	-0.308*	0.019
5	-0.148*	-0.142*	-0.016	5	-0.228*	**-0.231***	-0.157*	5,6	**-0.305***	-0.293*	0.018
6	-0.139*	-0.134*	-0.003	6	-0.198*	**-0.206***	-0.155*	6,7	**-0.248***	-0.238*	0.026
7	-0.152*	-0.150*	-0.003	7	-0.166*	**-0.168***	-0.122*	7,8	**-0.203***	-0.196*	0.017
8	**-0.160***	**-0.153***	0.031	8	-0.1346*	**-0.135***	-0.096*	8,9	**-0.192***	-0.184*	0.049*
9	-0.149*	-0.142*	0.003	9	**-0.120***	-0.118*	-0.089*	9,10	**-0.149***	-0.142*	0.003
10	-0.147*	-0.138*	0.053*	10	**-0.092***	-0.089*	-0.074*				

6.1. RQ1 - Single-Turn predictors

Table 2 shows the results for the three single-turn predictors, namely: the score of the top-ranked item at a given turn k (denoted top-1@k); the mean value of all top-ranked items in the recommendation list at a given turn (mean@k); and the standard deviation values of the scores of all top-ranked items (sd@k).

The table is grouped into three sets of columns defining the prediction turn and the ground truth turn. Specifically, Prediction at turn k with rank@turn10 addresses long-term prediction; the middle group, Prediction at turn 2 with rank@turn k, addresses whether prediction at an early turn can help identify success at early or late turns; finally, the third group, Prediction at turn k with rank@turn $k+1$, addresses short-term prediction.

We first examine the first group of columns, which aims to determine the extent that the overall conversation can be successfully predicted (i.e. the ground truth is the rank of the target item at turn 10). Overall, the correlations[2] are weak (-0.184 is the strongest observed for Shoes, and -0.160 for Shirts; Dresses is lower still at -0.080), yet significant ($p < 0.05$). This suggest the difficulty of the long-term prediction task. We do observe that correlations are relatively higher as the prediction turn increases - thus indicating that it is easier to predict performance at turn 10 using evidence of the ranking at turn 10. Finally, among the predictors, the maximum score at each turn, along with the mean score, exhibit higher correlations the standard deviation. To answer RQ1 (a), we cannot sufficiently predict long-term conversation

[2] In our analysis, we ignore the sign of the correlation - indeed, the observed correlations are negative, as our CRS system uses representation *distances* rather than similarities.

performance using single-turn score-based predictors.

Turning next to the second group of columns, we observe stronger correlations. Indeed, the overall higher correlations suggests that predicting at turn 2 gives more accurate predictions, particularly when aiming to predict conversation performance at turn 2 or shortly thereafter. In particular, for the Shoes datasets, medium strength correlations of -0.423 are observed - these are in line with the best accuracy of some QPP predictors for adhoc search tasks [12, 25, 30, 24]. Correlations of -0.305 and -0.281 are observed for Shirts and Dresses, respectively. Among the predictors, top-1@k is again most successful on Shoes, but on Dresses and Shirts, where correlations are lower, the overall picture is less clear across different prediction horizons (i.e. as the ground truth k is varied). For these datasets, mean is the most accurate for most values of $k \geq 2$. In general, when predicting conversation performance using single-turn retrieval scores, prediction becomes less accurate as the longevity of the prediction increases, thus answering RQ1(c).

Finally, the last set of columns of the table shows the correlation of the scores of each turn k (as a predictor) when the effectiveness of the following turn $k+1$ is used as the ground truth (i.e. applying a short-term horizon). The scores of both the top-ranked item and the average score of the top-ranked items at turn k sufficiently predict the rank of turn $k+1$, especially for early turns. This trend weakens as the number of turns increases, but the observed correlations remain quite high for some cases. For example, for Shoes, we start with a correlation of -0.423 (maximum score) and -0.413 (average score) for turns $2, 3$ and at turns $8-9$ the correlation is still -0.20. For Shirts, the maximum and average score of top items sufficiently predict the ranking of turn 3 at -0.30 and the score of turn 8 still at -0.20. Finally, although weaker than the other two datasets, the two predictors work reasonably well for Dresses, achieving a maximum value of -0.26 for predicting the rank of turn 3. These values suggest some evidence for short-term prediction when using single-turn score-based predictors, to answer RQ1(b).

Overall, we observe that there is some evidence for short (score of one turn predicting the rank of the following turn) and early prediction (a score of initial turn predicting the rank of some turns ahead). The score of the top-ranked item and the mean scores of the recommendation list are shown to be the most promising single-turn predictors. However, contrary to previous QPP research [25], the results for the standard deviation are not as encouraging. The results for long-term prediction are weaker, but still, the score of the initial turn is predictive of later stages. In general, prediction of the system performance (whether it finds the target in the context of a conversation) is possible by using single-turn score-based predictors, particularly for the success of the conversation at early turns and prediction of the next turn. In RQ2 below, we focus on short-term (next turn prediction), as the most promising CPP setting.

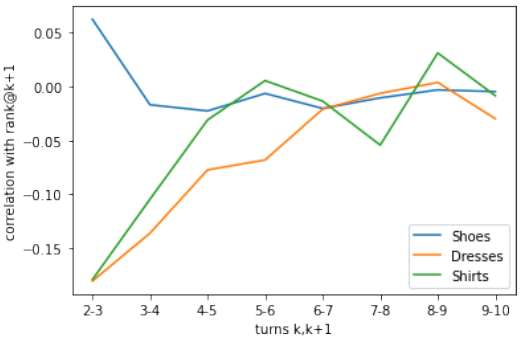


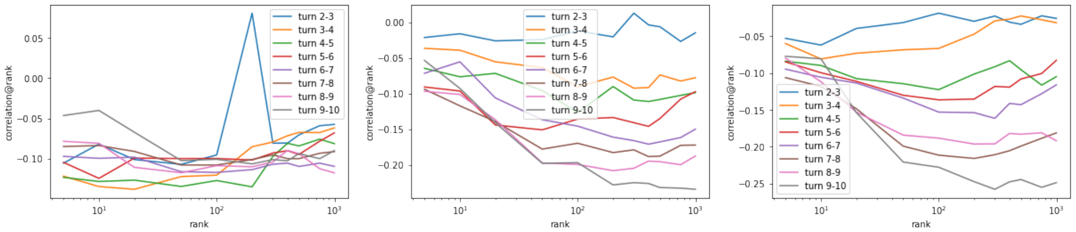
Figure 2: Results of the difference in the top-1 ranked item (maximum score) between pairs of consecutive turns as a consecutive turn CPP predictor for each of the datasets.

6.2. RQ2 - Consecutive-Turn predictors

Figure 2 presents the results of our first consecutive-turn predictor, namely the difference in maximum score (top-1 item) for each pair of turns $k, k+1$ when predicting the rank of the target item at turn $k+1$. Within the figure, each dataset is represented as a separate curve. Considering the different datasets, for Shirts and Dresses, we observe a similar trend across turns, starting from a correlation of -0.18 (the maximum value obtained for this predictor) at turns 2-3, which gradually decreases as the number of turns increases. In contrast, Shoes does not achieve any correlation stronger than -0.016 at turns 3-4. Therefore, we observe only weak correlations for this predictor at short-term prediction, although some correlations are significant. To answer RQ2(a), using the scores of two consecutive turns, does not sufficiently predict conversation performance, and is indeed generally less effective than the predictors examined in RQ1.

Next, we test our final predictor, which considers the overlap of top-ranked items (i.e., the size of intersection) between consecutive turns. We considered various rank cutoff values for calculating the overlap, ranging from rank 5 to rank 1000, and all pairs of turns. Figure 3 reports the observed correlations (y-axis), where each pair of turns is a curve, and the x-axis is the rank cutoff at which overlap is calculated. Recall that we expect that when the retrieved items are generally similar, this may be indicative that the CRS is reaching a stable conclusion of the likely relative items. If this occurs at a later turn, we may be further confident in the likely positive performance of the system.

On analysing Figure 3, we note that Dresses & Shirts (Figure 3(b) & (c), respectively) – which are both Fashion IQ datasets – we observe a strengthening trend in the correlations as we increase the rank cutoff value (more



(a) Shoes (b) Dresses (c) Shirts

Figure 3: For each dataset, results for overlap of top-ranked items as a consecutive predictor for all pairs of turns $k, k+1$ for a number of rank cutoff values.

items are considered). This happens for all pairs of turns except the initial turn. In addition, the correlations are stronger for later turns than earlier turns, indicating that this predictor is more useful for later turns (as expected). Indeed, improved prediction at later turns is particularly notable, as this contrasts with our results in RQ1, where earlier prediction was more accurate.

On the other hand, for the Shoes dataset, the highest correlations are observed for turns 3-4 and 4-5, and for cutoff values at 50 and 100. The correlations for item overlap in Shoes are weaker than the other two datasets, contrasting with the observations in RQ1 (where Shoes exhibited higher correlations for the single-turn predictors than Dresses or Shirts). We note that, as a CRS dataset, Shoes is "easier" than Dresses (e.g. the GRU model can attain Mean Reciprocal Rank 0.2 at turn 10 on Shoes, compared to Mean Reciprocal Rank 0.075 at turn 10 on Dresses [10]). We postulate that early single-turn prediction works well on Shoes, as more conversations are answered at earlier turns; in contrast, on Dresses, more critiques are required for successful conversations, and the overlap-based evidence later in the conversation is therefore more useful for prediction.

Overall, these results suggest some weak-medium correlations (upto -0.25 ρ) on the overlap-based consecutive turn predictor, thereby answering RQ2(b).

7. Conclusions

We have presented a novel framework for conversational performance prediction (CPP) that aims to detect the factors that indicate effective performance by taking into account the multi-turn aspect of the task of conversational interactive image retrieval. In this regard, we proposed a number of predictors that can be used for both short-term and long-term prediction, and explored the retrieval scores and retrieved items, of both a single turn and consecutive turns. We conducted our analyses on three widely-used relative captioning datasets for conversational recommendation systems (CRS) and examined the extent to which our proposed predictors are indicative of the ranking of the users' target items in the recommendation list.

In our analysis of the proposed single-turn predictors, we found that examining the score of the top-ranked items had a medium correlation with the effectiveness of the conversation, particularly the effectiveness at early turns. Indeed, we observed a Spearman's ρ of 0.423 on the Shoes dataset, which is comparable to correlations observed for standard QPP predictors on adhoc search tasks [12, 24, 25, 30]. However, these single-turn predictors became less useful at predicting the success of later turns. On the other hand, among our consecutive turn predictors, simply examining the overlap of the retrieved lists had a weak-medium correlation with late turn effectiveness on two out of our three datasets.

Overall, the weak-medium correlations observed for our simple unsupervised predictors of different families suggests that there is significant scope to extend this work, for instance by introducing supervised predictors. Moreover, our proposed framework for CPP is generalisable - for instance, we can also envisage predictors that examine aspects of the critiques (for instance, repeated critiques), or characteristics of the retrieved images (are item colours or styles varied). We leave these for future work. Furthermore, we also aim to extend our analyses to a classification task that aims to predict whether a conversation would fail, as well as testing the efficacy of interventions for failing conversations.

Finally, this study takes place in the context of user simulators for evaluation of CRS - such user simulators are common in the training and evaluation of conversational systems. Logging the interactions of a deployed CRS would allow to verify the results depicted here.

Acknowledgments

Maria Vlachou's work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant number EP/S02266X/1.

References

[1] T. M. Brill, L. Munoz, R. J. Miller, Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications, Journal of Marketing Management 35 (2019) 1401–1436.

[2] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2021) 1–36.

[3] F. N. Tou, M. D. Williams, R. Fikes, D. A. Henderson Jr, T. W. Malone, Rabbit: An intelligent database assistant., in: AAAI, 1982, pp. 314–318.

[4] L. Chen, P. Pu, Critiquing-based recommenders: survey and emerging trends, User Modeling and User-Adapted Interaction 22 (2012) 125–150.

[5] T. L. Berg, A. C. Berg, J. Shih, Automatic attribute discovery and characterization from noisy web data, in: Proc. ECCV, 2010, pp. 663–676.

[6] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, R. Feris, Dialog-based interactive image retrieval, in: Proc. NeurIPS, 2018, pp. 678–688.

[7] V. S. Bursztyn, J. Healey, E. Koh, N. Lipka, L. Birnbaum, Developing a conversational recommendation systemfor navigating limited options, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–6.

[8] Y. Jin, W. Cai, L. Chen, N. N. Htun, K. Verbert, Musicbot: Evaluating critiquing-based music recommenders with conversational interaction, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 951–960.

[9] W. Cai, Y. Jin, L. Chen, Critiquing for music exploration in conversational recommender systems, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 480–490.

[10] Y. Wu, C. Macdonald, I. Ounis, Partially observable reinforcement learning for dialog-based interactive recommendation, in: Proc. RecSys, 2021, pp. 241–251.

[11] T. Yu, Y. Shen, H. Jin, A visual dialog augmented interactive recommender system, in: Proc. KDD, 2019, pp. 157–165.

[12] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 299–306.

[13] B. He, I. Ounis, Inferring query performance using pre-retrieval predictors, in: International symposium on string processing and information retrieval, Springer, 2004, pp. 43–54.

[14] C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors, in: Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 1419–1420.

[15] D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, Synthesis Lectures on Information Concepts, Retrieval, and Services 2 (2010) 1–89.

[16] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris, Fashion iq: A new dataset towards retrieving images by natural language feedback, 2020. arXiv:1905.12794.

[17] J. Peng, C. Macdonald, B. He, I. Ounis, A study of selective collection enrichment for enterprise search, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, 2009, p. 1999–2002.

[18] S. Cronen-Townsend, Y. Zhou, W. B. Croft, A framework for selective query expansion, in: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04, 2004, p. 236–237.

[19] C. Macdonald, R. L. Santos, I. Ounis, On the usefulness of query features for learning to rank, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, 2012, p. 2559–2562.

[20] Y. Zhao, F. Scholer, Y. Tsegay, Effective pre-retrieval query performance prediction using similarity and variability evidence, in: European conference on information retrieval, Springer, 2008, pp. 52–64.

[21] F. Scholer, S. Garcia, A case for improved evaluation of query difficulty prediction, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 640–641.

[22] J. Mothe, L. Tanguy, Linguistic features to predict query difficulty, in: ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop, 2005, pp. 7–10.

[23] B. He, I. Ounis, Query performance prediction, Information Systems 31 (2006) 585–594.

[24] Y. Zhou, W. B. Croft, Query performance prediction in web search environments, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 543–550.

[25] A. Shtok, O. Kurland, D. Carmel, Predicting query performance by query-drift estimation, in: Conference on the Theory of Information Retrieval, Springer, 2009, pp. 305–312.

[26] M. Mitra, A. Singhal, C. Buckley, Improving automatic query expansion, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 206–214.

[27] R. Cummins, Document score distribution models for query performance inference and prediction, ACM Transactions on Information Systems (TOIS) 32 (2014) 1–28.

[28] F. Diaz, Performance prediction using spatial autocorrelation, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 583–590.

[29] A. Shtok, O. Kurland, D. Carmel, Query performance prediction using reference lists, ACM Transactions on Information Systems (TOIS) 34 (2016) 1–34.

[30] A. Shtok, O. Kurland, D. Carmel, Using statistical decision theory and relevance models for query-performance prediction, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 259–266.

[31] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, pp. 111–119.

[32] V. Lavrenko, W. B. Croft, Relevance-based language models, in: ACM SIGIR Forum, volume 51, ACM New York, NY, USA, 2017, pp. 260–267.

[33] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems (TOIS) 28 (2010) 1–38.

[34] O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, O. Rom, Back to the roots: A probabilistic framework for query-performance prediction, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 823–832.

[35] J. Kang, K. Condiff, S. Chang, J. A. Konstan, L. Terveen, F. M. Harper, Understanding how people use natural language to ask for recommendations, in: Proc. RecSys, 2017, pp. 229–237.

[36] I. Sekulić, M. Aliannejadi, F. Crestani, Exploiting document-based features for clarification in conversational search, in: European Conference on Information Retrieval, Springer, 2022, pp. 413–427.

[37] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, in: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, 2019, pp. 475–484.

[38] J. Kiesel, A. Bahrami, B. Stein, A. Anand, M. Hagen, Toward voice query clarification, in: The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 1257–1260.

[39] H. Zamani, S. Dumais, N. Craswell, P. Bennett, G. Lueck, Generating clarifying questions for information retrieval, in: Proceedings of the Web conference 2020, 2020, pp. 418–428.

[40] H. Roitman, S. Erera, G. Feigenblat, A study of query performance prediction for answer quality determination, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, 2019, pp. 43–46.

[41] W. Shi, K. Qian, X. Wang, Z. Yu, How to build user simulators to train RL-based dialog systems, arXiv preprint arXiv:1909.01388 (2019).

[42] X. Li, Z. C. Lipton, B. Dhingra, L. Li, J. Gao, Y.-N. Chen, A user simulator for task-completion dialogues, arXiv preprint arXiv:1612.05688 (2016).

[43] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: Proc. IEEE data engineering workshop, IEEE, 2007, pp. 801–810.

[44] Y. Wu, C. Macdonald, I. Ounis, Partially observable reinforcement learning for dialog-based interactive recommendation, in: Proceedings of ACM RecSys, 2021.

[45] Y. Wu, C. Macdonald, I. Ounis, Multimodal conversational fashion recommendation with positive and negative natural-language feedback, in: Proceedings of ACM Conversational User Interfaces, 2022.

[46] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, arXiv preprint arXiv:1511.06939 (2015).