Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval

Aliannejadi, M.; Chakraborty, M.; Ríssola, E.A.; Crestani, F.

Link to publication

Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval

Mohammad Aliannejadi*†
University of Amsterdam
m.aliannejadi@uva.nl

Esteban Andrés Ríssola*
Università della Svizzera italiana (USI)
esteban.andres.rissola@usi.ch

Manajit Chakraborty*
Università della Svizzera italiana (USI)
manajit.chakraborty@usi.ch

Fabio Crestani
Università della Svizzera italiana (USI)
fabio.crestani@usi.ch

ABSTRACT

With the improvements in speech recognition and voice generation technologies over the last years, a lot of companies have sought to develop conversation understanding systems that run on mobile phones or smart home devices through natural language interfaces. Conversational assistants, such as Google Assistant™ and Microsoft Cortana™, can help users to complete various types of tasks. This requires an accurate understanding of the user's information need as the conversation evolves into multiple turns. Finding relevant context in a conversation's history is challenging because of the complexity of natural language and the evolution of a user's information need. In this work, we present an extensive analysis of language, relevance, dependency of user utterances in a multi-turn information-seeking conversation. To this aim, we have annotated relevant utterances in the conversations released by the TREC CaST 2019 track. The annotation labels determine which of the previous utterances in a conversation can be used to improve the current one. Furthermore, we propose a neural utterance relevance model based on BERT fine-tuning, outperforming competitive baselines. We study and compare the performance of multiple retrieval models, utilizing different strategies to incorporate the user's context. The experimental results on both classification and retrieval tasks show that our proposed approach can effectively identify and incorporate the conversation context. We show that processing the current utterance using the predicted relevant utterance leads to a 38% relative improvement in terms of nDCG@20. Finally, to foster research in this area, we have released the dataset of the annotations.

*The three authors contributed equally to this work.

†Work done while Mohammad Aliannejadi was affiliated with Università della Svizzera italiana (USI).

1 INTRODUCTION

Recent emergence of intelligent assistants, such as Google Assistant™ and Microsoft Cortana™ have led to an increasing interest in research on conversational systems. Conversational assistants can help users complete various types of tasks. The tasks can range from as simple as setting an alarm to more complex cases like health advice. Conversational assistants have been employed for information seeking [35] and recommendation [34] among other information systems applications. Moreover, such systems can be used as home assistants, *e.g.,* Alexa and Google Home; or be integrated in a smartphone [1, 2] or wearable devices, *e.g.,* Apple Siri. Several Information Retrieval (IR) tasks have been investigated in a conversational setting. Some examples include response ranking [40], item recommendation [11], evaluation [19, 21], and asking clarifying questions for users intent disambiguation [3].

Although much work has been devoted towards studying single-turn conversations, several new challenges, that a multi-turn dialogue based conversational system poses, remain to be explored. Multiple turns in a conversation can be used to understand the user information need more effectively [30]. For instance, while searching for a new smartphone, user and machine could discuss various features and options in multiple turns. In order to facilitate research on multi-turn information seeking conversations, TREC introduced the Conversational Assistance Track (CAsT)[1] in 2019. The track contains 80 conversations on different topics, each of which consists of 8-12 turns (or utterances). Much work has been done on multi-turn conversational question answering [10, 16]. However, not much is known about how users interact with a machine in a multi-turn conversation and how their utterances can help the system to understand their information needs. Moreover, various aspects of multi-turn conversations are yet to be investigated, such as the way a user's information need and intent evolves as the conversation develops.

In this work, we aim to provide an understanding of users' behavior and how their information needs evolve during the course of a multi-turn conversation. Furthermore, we investigate the modeling of conversation context by fine-tuning BERT. To this end, we provide an in-depth analysis on the structure and evolution of

[1]http://www.treccast.ai/

Number:	37
Title:	prison psychology studies
Description:	The Stanford Experiment and its ramifications in human testing, particularly on prisoners and its ethics.
1	What was the Stanford Experiment?
2	What did it show?
3	Tell me about the author of the experiment.
4	Was it ethical?
5	What are other similar experiments?
6	What happened in the Milgram experiment?
7	Why was it important?
8	What were the similarities and differences between the studies?
...	...

Figure 1: Identifying Relevant Utterances.

users' utterances on the TREC CAsT 2019 collection. To understand the dependency of conversation turns, we have annotated each turn of the conversation with related turns in the conversation's context. This turn-relevance annotation enables us to understand and visualize the conversation evolution and dependencies. An utterance is *relevant* to the *current* one when: (i) it is needed to clarify the *current* utterance or (ii) it augments the information need for the *current* utterance or (iii) it entails the *current* utterance. As an example, consider the following conversation extracted from the TREC CAsT 2019 dataset in Figure 1. As we can observe, utterance u_8 cannot be answered unless additional information from previous utterances, *i.e.*, the context, is included in the question.[2] In this case, utterances u_1 and u_6 provide the necessary context and, thus, are considered *relevant* to u_8.

Moreover, we propose a neural model for predicting relevant conversation turns. Our model, called BERT-CUR, learns a joint high-dimensional representation of the user's utterance and its position in the conversation. We evaluate the effectiveness under two settings, namely, classification and retrieval performance. For the former, we study the effectiveness of our model in predicting the relatedness labels. For the latter, we follow a simple technique for incorporating the related turn into the user's current utterance. Moreover, we report the performance of classical IR models using different techniques to improve the current utterance.

The contributions of this paper can be summarized as follows:

- We create CAsTUR, a dataset of conversation turn relevance annotations for the TREC CAsT 2019 collection. CAsTUR is publicly available for research purposes.[3]
- We conduct an extensive analysis on the language, relevance, and dependency of multi-turn information-seeking conversations.
- We propose a neural model for predicting related conversation turns.

---

[2]In this paper, we use the terms "utterance" and "question" interchangeably.
[3]https://github.com/aliannejadi/castur

- We evaluate and analyze the performance of BERT-CUR with two sets of experiments both as a classifier and as part of an IR system.

Experimental results show the effectiveness of our model in improving the user's utterance by adding the relevant conversation context. In particular, we see that updating utterances with their related turns results in 38% relative improvement in terms nDCG@20. Moreover, we see in the analysis that the majority of turns are related to the first turn of the conversation. This suggests that a simple heuristic model – assuming the first utterance of the conversation is related – could also improve the retrieval performance.

2 RELATED WORK

Conversational search has been a long standing research problem in the IR community. However, with the recent advances in automatic voice recognition and the proliferation of Intelligent Personal Assistants such as Siri®, Alexa™, Google Assitant™, and Cortana™on personal devices, the area of Conversational search has received renewed attention in the past few years. One of the earliest attempts towards conversational mode in IR can be traced back to the work of Oddy [27] who proposed the introduction of *dialogue* for searching documents. Levinson and Shipley [23] later proposed a system for retrieving airline information and ticket reservation using speech. Another IR system directed towards medical health professionals within a conversational setting was put forward by Bolc et al. [9] in 1985 for searching documents related to Gastroenterology. However, it was the work by Croft and Thompson [12] on I³R, an expert interface communicating with the user in a search session, which laid the foundation for conversational IR. A few years later Belkin et al. [7] characterized information-seeking strategies for conversational IR, offering users choices in a search session based on case-based reasoning. Since then, the problem of conversational systems has been studied by researchers from both the fields of IR and Natural Language Processing (NLP) with varied interests. Conversational Agents have forayed their applications in various domains ranging from conversational recommender systems [11, 34], human memory augmentation [6], e-Health systems [25], personality recognition [31] to museum tour guidance [22]. Gao et al. [17] provides a systematic review on neural approaches to conversational AI developed in the last few years. Recently, rule-based conversational IR system [24, 38, 39] have given way to learning based approaches [20] and even more recent methods based on deep learning [41]. Among the several facets of conversational IR systems, one research direction is focused on analyzing user-behavior and interaction with voice-only systems [33]. Along the same line, Radlinski and Craswell [30] proposed a theoretical framework for conversational search highlighting the need for multi-turn interactions with users for narrowing down their specific information needs. Trippas et al. [35] studied conversations of real users to determine the frequently-used interactions and inform a conversational search system design. A close line of research deals with identifying user-intent while searching for information. Much work has been done in this direction, some of which include query suggestion to clarify users' intent in a traditional IR setting [30], asking clarifying questions from users to

understand users' intent and redirect the search [3], clarifying user-intent by eliminating non-relevant items through negative user feedback in a conversational search [8]. On the other end of the spectrum, Azzopardi et al. [5] posited that while understanding user intent and actions is important, little work has been addressed towards understanding the action taken by a conversational agent in the same context. They thus provided a framework for under-standing the human-computer interaction from an agent's point of view. Vtyurina et al. [37] listed the important factors to consider while designing a conversational assistant. One of their key find-ings was that it is essential to maintain the conversational context which is one of the focal points of our paper.

3 DATA ANNOTATION AND ANALYSIS

3.1 TREC CAsT

TREC CAsT was organized for the first time in 2019 [13]. Its main purpose is to advance research on conversational search systems. In particular, the goal of the track is to create a reusable benchmark for open-domain information centric conversational dialogues. It aspires to establish a concrete and standard collection of data with information needs to make systems directly comparable. The task was organized to focus on candidate information ranking in context involving two stages:

- *Read the dialogue context*: Track the evolution of the informa-tion need in the conversation, identifying salient information needed for the current turn in the conversation.
- *Retrieve Candidate Response Information*: Perform retrieval over a large collection of paragraphs (or knowledge base content) to identify relevant information.

The track is built on top of three open-source collections, namely, MS MARCO (MAchine Reading COmprehension) Ranking pas-sages [26], TREC CAR 2018 [15] paragraph collection and News article from Washington Post (WAPO)[4]. MS MARCO collection was released by Microsoft in 2018 and comprises of 8,841,823 pas-sages – extracted from 3,563,535 web documents retrieved by Bing. TREC-CAR (Complex Answer Retrieval) is a corpus of 20 million paragraphs harvested from a snapshot of Wikipedia.

We performed experiments on MS MARCO and TREC-CAR 2018. We did not use the WAPO collection because of inconsistencies as reported by TREC and discrepancies associated with them in the relevance judgment file. The TREC CAsT 2019 track came with a set of 30 training conversations and 50 evaluation conversations[5]. Each conversation consisted of several turns.

3.2 Data Annotation

One way to determine the relevant context in a conversation is to track the evolution of the information need and identify salient information in the conversation such that the current utterance is enhanced to express user's information need more accurately. For instance, consider the conversation in Figure 1 and assume that the current turn is at utterance 8. Up to this point, previous utterances, u_1 and u_6 are *relevant* to u_8. In fact, without taking into account such information the utterance u_8 could hardly be addressed.

Table 1: Descriptive Statistics for TREC-CAst and CAsTUR.

Parameter	Value
# conversations	80
# turns	748
Avg. # turns per conversation	9.35 ± 1.32
Median # turns per conversation	9
Avg. # terms per turn	7.21 ± 2.02
# relevance annotations	752
% utterances with **0** rel. turns	18%
% utterances with **1** rel. turns	71%
% utterances with **2** rel. turns	10%
% utterances with **3** rel. turns	< 1%
% utterances with **4** rel. turns	< 1%

Given a conversation, a turn and all the previous utterances until that turn, three human annotators were asked to select one or more utterances which could help to better express the information need presented in the current turn of the conversation. After independent labeling, we computed the percentage of agreement among the annotators. If the agreement score was greater than or equal to 66.67% (*i.e.,* at least two of the three annotators agreed on the same set of relevant utterances), we recorded the relatedness label. For turns that had an agreement score below 66.67%, the three annotators deliberated and made an agreement on the relevance of the corresponding utterances. From the 748 utterances, 625 were annotated without any major disagreements. The remaining 123 annotations were resolved by a rigorous discussion between the annotators. This procedure was conducted for both the training and evaluation sets.

We name this set of annotated utterances as *CAsTUR* and we released it for public use for research purposes.

3.3 Analysis

In this section, we report our analysis based on the TREC CAsT data, as well as our CAsTUR annotations.

Basic statistics. In the first half of Table 1 we present a summary statistics of TREC CAsT 2019 conversations dataset. We note that on average conversations span over nine turns with an average length of seven terms per turn. In particular, we notice that there are only 19 conversations out of 80, which have eleven turns or more. In the second half of the table, we outline statistics associated with CAsTUR (Section 3.2). We observe that 18% of the utterances do not have any related utterances. This indicates that those utterances are either self-contained or no additional information in previous conversation turns existed. Majority of the utterances have only one relevant turn. Very few of the utterances (< 1%) have three or more relevant turns.

Question language and categories. Depending on the type of question, we manually segregate them into one of the following nine categories: *What, When, Where, Which, Who, Why, How, Yes/No* and *Compare*. The category *What* includes all questions that either starts[6] with a "What" or ask for a description like "Tell me about
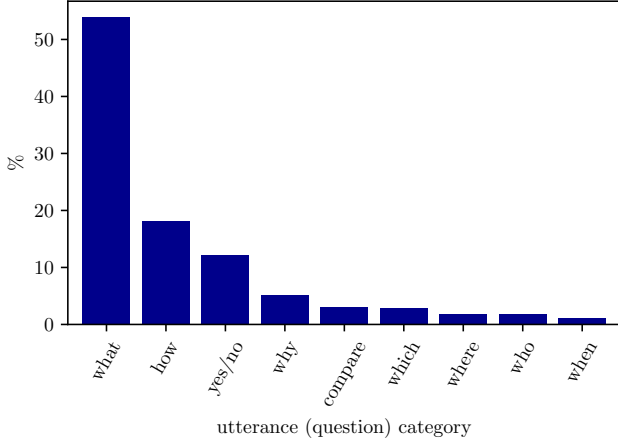
Figure 2: Question category distribution.

Boer goats." Questions that either collate or contrast two or more entities are categorized into *Compare*, *e.g.,* "What were the similarities and differences between the studies?" The rest of the categories are self-explanatory. The distribution of questions across categories is presented in Figure 2. Additionally, we computed for each category the number of conversations in which it is the leading category. We found that *What* questions dominate 67 conversations, while *How* dominates 12 conversations and the remaining one is dominated by *Why* questions.
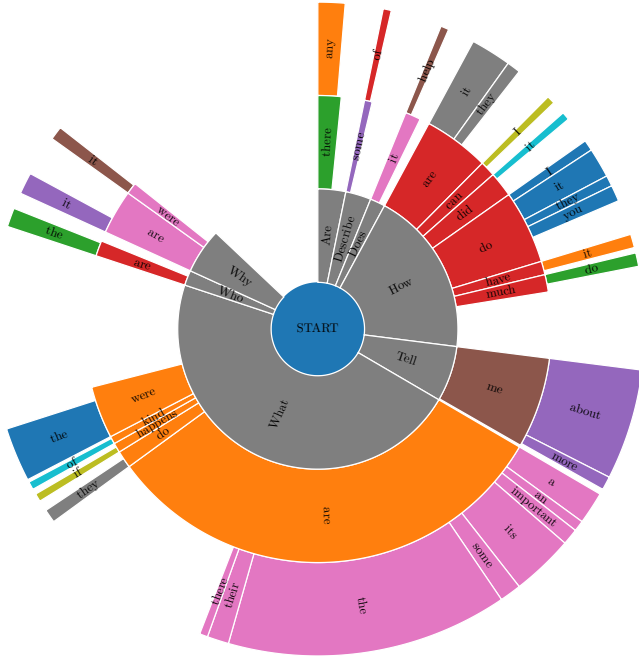

Figure 3: Distribution of top 40 trigrams of the questions (best viewed in color).

In Figure 3, we present the distribution of top 40 trigrams of the questions across all conversations. This diagram helps us identify the share and sequence of terms in a question across all conversations. To provide a more meaningful representation some of the auxiliary verbs have been modified (*e.g.,* "is" → "are"). For example,
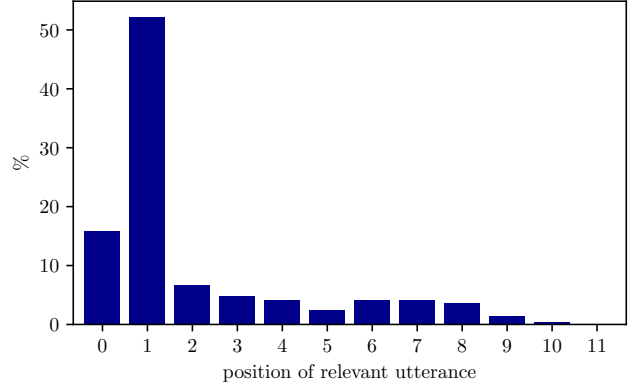

Figure 4: Distribution of *relevant* turn number $\mathcal{P}_{RT}$.

we see that many of the utterances start with "What" and we see that in many cases the users are asking for more details or additional information about specific topics (*e.g.,* "What are its" and "What happens if"). On closer observation we find that the third word in the trigrams are usually pronouns, which indicates that they are referring to the context of the conversation. In fact, we found that 50% of CAsT utterances contained at least one kind of English pronouns, 42% of which was the third word in the sentence. This highlights the significance of determining *relevant* context in a conversation.

Distribution of relevant turns. We also analyze the position of relevant turn, $\mathcal{P}_{RT}$, for each of the turns and how they are distributed across all conversations (Figure 4). We find that for more than 50% of the total turns, a relevant turn is found at the first position in the conversation.

The position 0 in the figure represents the number of self-contained utterances, *i.e.,* the ones which do not have any prior relevant utterance. This position does not account for the first utterances in each conversation, since they are assumed to have no *relevant* priors. Hence, they constitute 16% of the total number of relevant utterances.

Evolution of conversation. Since the topic of the conversation generally evolves or changes with each turn, we want to observe the difference in position of the relevant turn (RT) with the current turn (CT), $\Delta\mathcal{P}_{(CT,RT)}$, where $\mathcal{P}$ denotes position of an utterance. We find that for a large share of utterances, a relevant turn is found in close proximity (Figure 5). In fact, as the figure suggests, more than 30% of the utterances have a relevant turn at a position immediately prior to them.

Next, we study the relation between the position of the current utterance and its relevant turn. For instance, as observed in Figure 6, for turn number 10, the average relevant question position is 2; while for turn 11, the relevant question is usually found at position 7 in the conversation. This suggests that in most conversations of the TREC CAsT, the first seven turns are mostly dependent on the first turn, suggesting that users explore a single topic in several turns. However, we can also see that as the conversation progresses, the later turns are less dependent on the first turn (*e.g.,* turns 10, 11, and 12). This indicates the possibility that longer conversations may include multiple topics (*i.e.,* either elaborating a subtopic or starting a new topic).
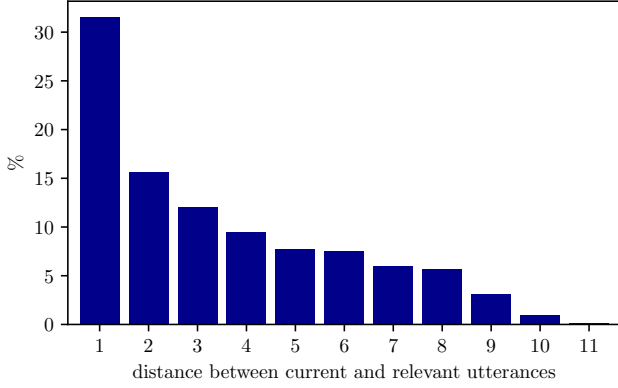
Figure 5: Distribution of difference in position of current and relevant question $\Delta \mathcal{P}_{(CT, RT)}$.
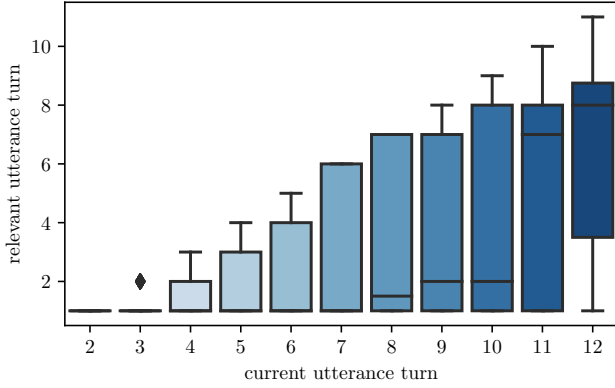

Figure 6: Distribution of relevant turn ($\mathcal{P}_{RT}$) per current turn ($\mathcal{P}_{CT}$).

Dependency among question categories. We use a Markov Model to analyze the flow patterns in the conversations as shown in Figure 7. The aim of this analysis is to understand the natural progression of a conversation in terms of question categories. Each of the nine question categories forms a node in a graph with two additional placeholder nodes, START and END which denote the beginning and end of the conversation. Therefore, for the conversation example in Figure 1, the labeled sequence would be "START→What→What→What→Yes/No→What→What→Why→Compare→ ... →END". The edges between the nodes are weighted by the transition rate. We denote the highest probable transition from each node by marking its respective edge in blue (*e.g.,* "which → what"). This figure represents the question category transition pattern in an information seeking process. We can make some observations from the graph: (1) For most cases, we note that the conversation starts with a *What* question while there is no clear candidate for end of conversation. (2) *What* questions are more likely to be followed by another *What* question, indicated by a blue self-loop. (3) *Which* questions are followed by more diversified questions *i.e.,* they can be followed by different category questions.

4 RETRIEVAL FRAMEWORK

In this section, we describe the framework that we design for utterance dependency prediction and document retrieval. First, we explain our model for predicting the related utterance to the current
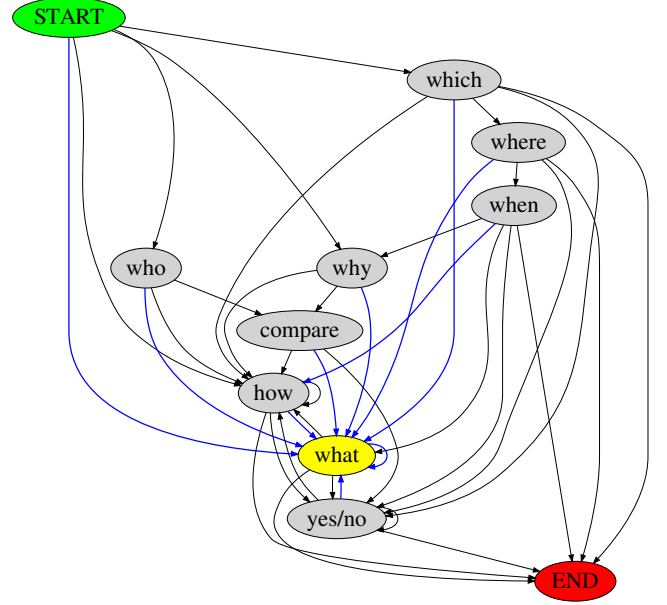
Figure 7: Flow pattern of conversations with a Markov model. Nodes are question categories. Edges are directed and weighted by transition probability (best viewed in color).

one. Further, we explain how we use the prediction to improve the current utterance and retrieve documents.

4.1 Utterance Relevance Prediction Model

We now describe our BERT[7] Utterance Relevance Prediction model, called BERT-CUR. Here, our goal is to maximize the performance of our retrieval model by selecting the most relevant utterances as the context of the current utterance. The aim is to select related utterances such that the current utterance is enhanced to express the user information need more accurately. As we saw in Section 3.3, predicting the relevant context is critical as many of the utterances contain a pronoun, referring to their context in the conversation. Based on our analysis, the relative distance of two utterances is negatively correlated with their relevance. Therefore, finding the related utterance among multiple ones in the conversation history requires joint modeling of the utterance language and position in the conversation. We learn a high-dimensional language and position representation for predicting the most relevant utterances to the current utterance. Formally, BERT-CUR estimates the probability $p(R = 1|u_i, u_j)$, where R is a binary random variable indicating whether the utterance u should be marked as relevant ($R = 1$) or not ($R = 0$). Let u_i and u_j denote the current and the candidate utterance, respectively. The utterance relevance probability in the BERT-CUR model is estimated as follows:

$$p(R = 1|u_i, u_j; i, j) = \psi\big(\phi_U(u_i), \phi_C(u_j), \eta_U(i), \eta_C(j)\big) , \quad (1)$$

where ϕ_U and ϕ_C respectively denote the current and candidate utterance representations. Whereas η_U and η_C denote the position representation of the current and candidate utterances, respectively. ψ is the matching component that takes the aforementioned representations and produces an utterance relevance score. There are various ways to implement any of these components.

---

[7]BERT: Bidirectional Encoder Representations from Transformers [14]

We implement ϕ_U and ϕ_C similarly using a function that maps a sequence of words to a d-dimensional representation ($V^s \rightarrow \mathbb{R}^d$). We use the BERT [14] model to learn these representation functions. BERT is a deep neural network with 12 layers that uses an attention-based network called Transformers [36]. We initialize the BERT parameters with the model that is pre-trained for the language modeling task on Wikipedia and fine-tune the parameters on CAsTUR. BERT has recently outperformed state-of-the-art models in a number of language understanding and retrieval tasks [14]. In addition, BERT shows promising results in modeling short texts and small collections.

We model the utterance position using η_U and η_C, mapping the conversation turn position to a high-dimensional representation ($\mathbb{N}^t \rightarrow \mathbb{R}^{dt}$). The component ψ is modeled using a fully-connected feed-forward network with the output dimensionality of 2. Rectified linear unit (ReLU) is employed as the activation function in the hidden layers, and a softmax function is applied on the output layer to compute the probability of each label (*i.e.,* relevant or non-relevant). To train BERT-CUR, we use a cross-entropy loss function and follow the point-wise learning approach. We train BERT-CUR using two BERT models. BERT-CUR $_{\text{BASE}}$ fine-tunes the uncased base BERT model, whereas BERT-CUR $_{\text{LARGE}}$ fine-tunes the uncased large BERT model.

4.2 Utterance Reformulation and Document Retrieval

After finding the relevant utterances as conversation context, the next step is to augment and reformulate the current question for improved document retrieval. This step involves processing of the original utterance, as well as adding terms from its context to help the document retrieval model rank relevant documents higher. The query processing and reformation is done in the following stages:

(1) In a conversation, it is imperative that there will be some form of relationship between the turns. We observed from the data that a large share of utterances were dependent on previous turns in the conversation. For example, in Figure 1, u_2 is clearly dependent on u_1; u_3, u_4 also depends on u_1 to make a meaningful question. Hence, it is essential to perform anaphora (coreference) resolution on such utterances to make them a complete question in itself. This step is performed only on FirstUtterance (FU) technique. To this end, we use AllenNLP [18] coreference resolution tool . For self-contained and first turns in the conversations, the output of the tool is the same as the original turn.

(2) On this modified query, we perform stopword removal, removal of special characters, and tokenization.

(3) We perform an additional step on utterances in BERT-CUR to ensure that the utterances retain the same order in the final query as in the conversation. This procedure ensures that BERT learns from the relative ordering of questions as well. It is to be noted that for our proposed BERT-CUR model, we did not perform Step 1 on the utterances since our objective is to learn the correct relevant utterance from the context.

After updating the utterances, we pass the resulting query set to various document retrieval models. For document retrieval we

Table 2: Names and descriptions of utterance reformulation techniques.

Name	Description
Orig.	The original unmodified utterance.
FU	The current utterance reformulated using the first utterance of the conversation.
PU	The current utterance reformulated using the immediate previous utterance of the conversation.
FPU	The current utterance reformulated using the first and immediate previous utterance of the conversation.
AU	The current utterance reformulated using all the previous utterances of the conversation until that point.
PrU	The current utterance reformulated using the relevant utterance predicted by BERT-CUR.
Oracle	The current utterance reformulated using the true relevance labels.

employ Galago[8] and index both the MS MARCO and TREC-CAR collections together. The total number of documents indexed was 38,429,852. We employ four different techniques to reformulate the current utterance and compare the performance of three term-matching IR models, namely, Okapi-BM25 [32], Query Likelihood (QL) [29] and Divergence From Randomness (DFR) [4]. We investigate four utterance selection strategies for each of the retrieval models. More details of the utterance reformulation strategies can be found in Table 2. As we see in Table 4, the performance of the proposed BERT-CUR is compared with the original utterance (*i.e., Orig.*), as well as the reformulated utterance using the first utterance of the conversation (*i.e., FU*). We use the first utterance as a heuristic model, motivated by our analysis which showed that the majority of relevant utterances are found in the first turn of the conversation. Comparison with oracle models allows us to validate the effectiveness of utterance relevance annotations. Also, it gives us an idea of the upper bound in improving the retrieval performance, following the mentioned utterance reformulation steps.

Moreover, it is worth mentioning that we do not investigate the performance of neural ranking or re-ranking approaches such as BERT fine-tuning models. Our decision is motivated by the relevance annotation used to create the TREC CAsT collection. Since the pooling and labeling is done using only term-matching models, the superiority of more sophisticated approaches such BERT is not necessarily reflected in the evaluation, simply because the documents that they retrieve were not present in the evaluation pool. Therefore, the validity of the experiments can be guaranteed only on term-matching approaches.

5 EXPERIMENTS

In this section, we describe the data, metrics, and compared methods. Furthermore, we evaluate the performance of the proposed models

---

[8]https://www.lemurproject.org/galago.php

in comparison with the baselines. We also study the performance of the models at different turns of the conversation.

5.1 Experimental Setup

Dataset. Since we conduct the evaluation on the TREC CAsT data, we did not have the relevance judgment for the 50 evaluation topics provided by TREC, at the point of writing this paper. In the absence of gold standard, we used the 30 training topics (which had 269 utterances in total) from the TREC challenge as our *test* set while using the TREC evaluation topics as our *training* set for BERT-CUR. Out of 50 training topics, which had 479 utterances, we randomly selected 10% as *dev* set. We used these sets to train, tune, and test BERT-CUR, as well as the compared methods.

Utterance relevance prediction evaluation metrics. Since we model the utterance relevance problem as classification, we consider three standard metrics to evaluate the effectiveness of the models, namely, Precision (Prec.), Recall and F1-Measure. For computing these scores, we treat each of the 748 turns (*i.e.,* as a single instance) individually.

Ad-hoc retrieval evaluation metrics. Here, effectiveness is measured considering the standard evaluation metrics of IR, as well as the ones used in the TREC CAsT track. Therefore, we evaluate the retrieval models using eight standard evaluation metrics: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), normalized discounted cumulative gain at the top 5,10 and 20 retrieved documents (nDCG@5, nDCG@10, nDCG@20) and, precision at the top 5,10 and 20 retrieved documents (P@5, P@10 and P@20). We use the relevance assessments as released by TREC.

Statistical test. We determine statistically significant differences using the two-tailed paired t-test at a 95% confidence interval ($p <$ 0.05).

Compared methods. We compare the performance of BERT-CUR with competitive classification baseline methods, details of which are as follows:

- *FirstUtterance/FirstPrevUtterance/AllPrevUtterance*: we always select the first/first and previous/all previous utterances in the conversation as relevant to the current one.
- *Random*: we select any random turn from the conversation until the current utterance to be *relevant*.
- *k-NN, k-NN-AWE*: To find the nearest neighbors in *k*-nearest neighbors (k-NN), we consider the cosine similarity between the TF-IDF vectors and turn numbers of utterances. Then, we take the labels of the nearest questions and produce the relevant question ranking. As for k-NN-AWE, we compute the cosine similarity between the average word embedding (AWE) of the questions obtained from GloVe [28] with 300 dimensions. The value of *k* is determined using 5-fold cross validation on the *dev* set.
- *MLP, MLP-AWE*: For MLP we used 2 hidden layers along with Adam optimization algorithm with default parameters. The MLP-AWE method for using word embeddings was similar to the one used in k-NN-AWE.
- *SVM, SVM-AWE*: We use linear SVM with $C = 0.1$ to predict the most relevant question. Similar to MLP-AWE, SVM-AWE also

Table 3: Question Relevance Prediction Performance. Best performances are marked in bold.

	Prec.	Recall	F1
FirstUtterance	0.60	0.65	0.62
PrevUtterance	0.29	0.31	0.30
FirstPrevUtterance	0.42	0.85	0.56
AllPrevUtterances	0.20	**1.00**	0.33
Random	0.20	0.50	0.29
k-NN	0.64	0.49	0.55
MLP	0.63	0.48	0.54
SVM	0.64	0.43	0.52
k-NN-AWE	0.64	0.49	0.55
MLP-AWE	0.64	0.53	0.58
SVM-AWE	**0.66**	0.49	0.57
BERT-CUR $_{BASE}$	**0.66**	0.69	**0.67**
BERT-CUR $_{LARGE}$	0.65	0.67	0.66

follows the same approach as in k-NN-AWE for obtaining the average word embeddings.

5.2 Results and Discussion

We evaluate the performance of BERT-CUR and the utterance reformulation techniques with respect to classification and document retrieval performance. Furthermore, we conduct various analyses on the results to understand the influence of conversation context in the performance.

Relevant utterance prediction performance comparison. We start by presenting the comparison of the relevant utterance prediction models in Table 3. We can clearly observe that our BERT-CUR outperforms all the other prediction models. Although, the *SVM-AWE* model achieves the same Precision as BERT-CUR $_{BASE}$, our proposed model supersedes the *SVM-AWE* in terms of Recall and hence F1-Measure as well. In terms of F1-Measure, the closest performing model to BERT-CUR is *FirstUtterance*, which corroborates our findings from Figure 4 that first utterance is relevant for more than 50% of the total utterances in the dataset. It states that in the majority of cases, the first turn in the conversation identifies the main topic of the conversation and the rest of the turns explore more details of the same topic, rather than shifting to a different topic. Notice that *FirstUtterance* is also the best performing heuristic for utterance prediction. Also, we observe that BERT-CUR $_{LARGE}$ also outperforms the baseline models. However, unlike other tasks, we observe better performance achieved by fine-tuning the BERT base model. This is because of smaller data size compared to the tasks where the BERT large model performs better.

Document retrieval performance comparison. Here, we also study the performance of an oracle model, *i.e.,* assuming that an oracle model is aware of the correct relevant question to all the questions. The goal is to show to what extent adding a relevant question to the current turn can improve the performance of a retrieval system. The results are presented in Table 4. Seven utterance reformulation techniques (see Table 2) are compared for each of the three retrieval models. The best retrieval performance is achieved when we use the QL model for retrieval. We see that

all utterance reformulation strategies improve the performance compared to the original utterance, indicating the high dependence of utterances on previous turns of the conversation. Also, we see that even employing simple heuristic utterance selection, such as selecting the first utterance (*i.e.,* FU) or previous utterance (*i.e.,* PU) of the conversation, leads to retrieval improvement. However, it is interesting to note that no heuristic method achieves significant improvement over the original utterance model (*i.e.,* Orgi.).

As we see in the sixth row of each section of Table 4, our proposed model *PrU*, outperforms all the baselines for all the three retrieval models. We observe that while the improvements on the DFR and QL models are consistently statistically significant (except for MRR on QL), they are less consistent on BM25. This could be due to the length of the reformulated utterances which are usually augmented with the relevant utterance(s). Notice in the table that the improvements over *FU* is statistically significant when we use the QL retrieval model, in terms of P@k. This shows the potential gain of adding good relevant utterances to the current utterance on the performance of a conversational system. It is worth mentioning that although Oracle model is an idealistic model, there were three instances in which *PrU* performed better than *Oracle* model *i.e.,* for DFR on P@10, as well as for QL on nDCG@5 and P@5. This suggests the strength of our model in capturing the semantic dependency of utterances, achieving comparable performance to the human annotators. It is worth noting that we report the performance of our proposed BERT-CUR model for precision-oriented relevant utterance selection (explained in details later in this section). Interestingly, we see that the difference between *Oracle* and *PrU* are not significant (no [6] in Table 4). This again suggests that BERT-CUR can effectively model and capture the semantic dependencies of utterances in different turns to the point where it can help the retrieval model to achieve comparable results to the perfect human annotation.

Impact of number of turns in the conversation on the retrieval performance. In this exercise, we study the effect of the utterance position (*i.e.,* turn) on the retrieval performance. Our aim is to understand the impact that the evolution of the conversation has during the course of a dialogue. Figure 8 depicts the performance of QL model using different utterance reformulation strategies in terms of nDCG@20. From Figure 8, we see that the overall performance of the system diminishes as the conversation progresses (with a few exceptions). Interestingly enough, turns 3 and 8 are positions where our proposed BERT-CUR-based *PrU* model performs better than *Oracle* model. For turn 2, we observe that adding a candidate relevant question to the existing one does not really affect performance. The same is self-explanatory for the first turn. On the other hand, as the conversation reaches its end, from turn 10 onwards we observe that the potential relationship between the relevant question and the current question decreases significantly. This is in contrast with the findings of Aliannejadi et al. [3] where they found that as the conversation evolves the retrieval performance also increases. This is because of the different nature of the conversations in this work as the conversations do not have clarifying questions. Asking multiple clarifying questions should indeed improve the performance because it helps the system to understand the user's information need more accurately. The
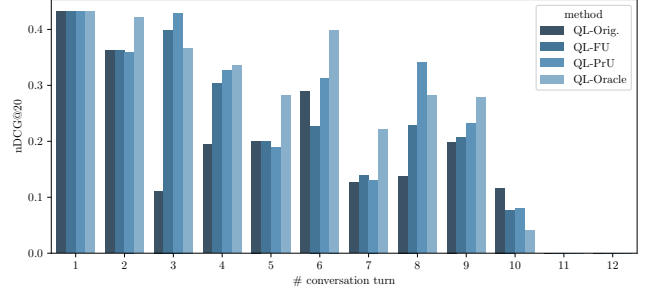

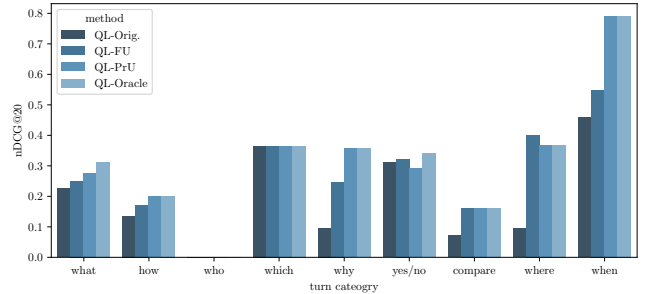**Figure 8: Impact of conversation turn number on the retrieval performance.**

Figure 9: Impact of utterance (question) categories on retrieval performance.

same does not necessarily hold in multi-turn conversations without clarifying questions because users' utterances do not necessarily elaborate more on the same information need and could diverge from the main topic of the conversation.

Impact of utterance (question) categories on retrieval performance. Next, we study the effect each question category has on the overall performance of the four models. Figure 9 shows us that for *Which* category, there is no effect of adding relevant questions to the model. This might be because there are very few utterances of *Which* category and some of those utterances are conversation starters, so dependence of such questions on other questions is low. We observe for *When* category remarkable improvement, at least over 60% compared to *Orig.* baseline, when either our proposed model or the oracle model is used. This can be explained from Figure 7, where we see that *When* category questions are always preceded by other category questions, indicating that it almost always depends on previous utterances of the conversation.

Impact of metric-oriented selection of relevant utterances on the retrieval. Our next objective is to study the effect of preference of metric (Precision vs. Recall vs. F1-measure) for relevant utterance selection on the overall performance of passage retrieval. The results of this exercise is presented in Table 5. The three runs, PrU_P, PrU_R and PrU_{F1} represent the best performing run (in terms of the metric itself) for the relevance selection model for Precision, Recall and F1-measure, respectively. From the three runs, we can see that selecting relevant questions with high precision, PrU_P, gives us the best retrieval performance[9] for all three retrieval evaluation measures. This demonstrates that for any current utterance selecting the correct relevant utterance(s) with a higher precision

---

[9]We present only QL scorer performance since it is the best performing scorer among the three.

Table 4: Passage Retrieval Performance Comparison. The superscripts 1/2/3/4/5/6 denote that the improvements over Orig./FU/PU/FPU/AU/PrU are statistically significant ($p < 0.05$). Best performing runs except Oracle model are marked in bold.

	Model	MAP	MRR	nDCG@5	nDCG@10	nDCG@20	P@5	P@10	P@20
DFR	Orig.	0.1079	0.1497	0.0988	0.1086	0.1319	0.0950	0.0800	0.0621
	FU	0.1682^5	0.2480	0.1687^5	0.1840^5	0.2189^5	0.1450^{45}	0.1108^{45}	0.0858^5
	PU	0.1438	0.2072	0.1399	0.1505	0.1793	0.1233	0.0983	0.0762
	FPU	0.1487	0.2312	0.1387	0.1654	0.1963	0.1050	0.0942	0.0763
	AU	0.1226	0.1923	0.1201	0.1299	0.1573	0.1017	0.0842	0.0667
	PrU	**0.1919**^{1345}	**0.2754**^{135}	**0.1906**^{1345}	**0.2058**^{135}	**0.2418**^{135}	**0.1650**^{1345}	**0.1317**^{1345}	**0.1004**^{1345}
	Oracle	0.2009^{12345}	0.3004^{12345}	0.2004^{1345}	0.2154^{1345}	0.2551^{12345}	0.1717^{1345}	0.1308^{12345}	0.1038^{12345}
BM25	Orig.	0.0887	0.1420	0.0841	0.0948	0.1131	0.0817	0.0733	0.0554
	FU	0.1359^5	0.2287^5	0.1432^5	0.1668^5	0.1907^5	0.1250^{45}	0.1092^{45}	0.0788^5
	PU	0.1263	0.1962	0.1268	0.1364	0.1679	0.1083	0.0908	0.0725
	FPU	0.1272	0.2108	0.1301	0.1472	0.1754	0.0967	0.0883	0.0692
	AU	0.1014	0.1654	0.1013	0.1135	0.1374	0.0883	0.0783	0.0629
	PrU	**0.1546**^{5}	**0.2465**^{5}	**0.1548**^{5}	**0.1805**^{135}	**0.2097**^{5}	**0.1400**^{45}	**0.1258**^{1345}	**0.0921**^{1345}
	Oracle	0.1633^{1345}	0.2658^{345}	0.1671^{1345}	0.1925^{1345}	0.2225^{1345}	0.1450^{1345}	0.1267^{1345}	0.0933^{12345}
QL	Orig.	0.1290	0.1603	0.1115	0.1302	0.1681	0.1100	0.0933	0.0833
	FU	0.2026^5	0.2788	0.1726^5	0.2111^5	0.2602^5	0.1500^5	0.1408^5	0.1113^5
	PU	0.1745	0.2486	0.1522	0.1799	0.2164	0.1433	0.1300	0.0963
	FPU	0.1884^5	0.2735^5	0.1602^5	0.1968^5	0.2368^5	0.1250	0.1275^5	0.0979^5
	AU	0.1429	0.2220	0.1211	0.1519	0.1848	0.1083	0.1083	0.0854
	PrU	**0.2362**^{1345}	**0.3009**^{5}	**0.2082**^{1345}	**0.2470**^{1345}	**0.2984**^{1345}	**0.1883**^{12345}	**0.1658**^{12345}	**0.1321**^{12345}
	Oracle	0.2435^{12345}	0.3243^{12345}	0.2059^{1345}	0.2493^{12345}	0.3103^{12345}	0.1817^{12345}	0.1692^{12345}	0.1383^{12345}

Table 5: Performance comparison of passage retrieval for metric oriented relevant question selection. Best performances are marked in bold.

	Turn Relevance			QL based IR performance		
	Prec.	Recall	F1	MAP	nDCG@20	P@20
PrU_P	**0.79**	0.54	0.64	**0.2362**	**0.2984**	**0.1321**
PrU_R	0.59	**0.72**	0.65	0.2113	0.1857	0.1108
PrU_{F1}	0.66	0.69	**0.67**	0.2272	0.2840	0.1229

is preferred over selecting all the relevant utterances. We have thus recorded the performance of the precision-oriented relevant question selection based retrieval model in Table 4. It is also interesting to see that PrU_P achieves the best retrieval performance although it has the least recall among the three models. Also, we see that PrU_P has the lowest F1-Measure, while achieving the highest retrieval performance.

6 CONCLUSIONS AND FUTURE WORK

In this work, we presented an extensive analysis of utterance dependency and language in multi-turn information-seeking conversations. To understand utterance dependencies we have annotated the conversations released by the TREC CAsT 2019 track. To do so, we hired three expert annotators and instructed them to identify the relevant utterance from a conversation's context. The analysis provided useful insights into the problem. For instance, we saw that there is a very strong negative correlation between relevance and distance of two utterances. This suggests that people tend to ask about information relevant to the most recent utterances.

Furthermore, we proposed a BERT fine-tuning model to predict the relevance of the utterances. Our proposed BERT-CUR model outperformed competitive classification baselines. Also, we demonstrated its effectiveness in improving the performance of document retrieval models. We conducted an extensive analysis on the performance of retrieval models. We found that incorporating the conversation context has positive impact on the retrieval performance. In particular, we discovered that for our proposed model the highest context dependency is found at the third turn of the conversation. Also, our experiments suggest that the task of relevant utterance prediction is highly precision-oriented, rather than being recall-oriented. As future work, we plan to perform a similar analysis on information-seeking conversational systems with the ability of asking clarifying questions [3]. The proposed task in the TREC CAsT track is aiming to evaluate systems that retrieve responses to the user's utterances. However, an ideal conversational system should be able to ask questions when required. This would have a great impact on the user's behavior and definition of relevance in the conversation context. As such, we plan to extend our analysis to such conversations and investigate effective modeling of relevant utterances.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2018. In Situ and Context-Aware Target Apps Selection for Unified Mobile Search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy*. ACM, 1383–1392.

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2018. Target Apps Selection: Towards a Unified Search Framework for Mobile Devices. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*. ACM, 215–224.

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. 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, SIGIR 2019, Paris, France*. 475–484.

[4] Gianni Amati and C. J. van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (2002), 357–389.

[5] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*.

[6] Seyed Ali Bahrainian and Fabio Crestani. 2018. Augmentation of Human Memory: Anticipating Topics that Continue in the Next Meeting. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA*. 150–159.

[7] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications* 9, 3 (1995), 379 – 395.

[8] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. 2019. Conversational Product Search Based on Negative Feedback. *CoRR* abs/1909.02071 (2019).

[9] Leonard Bolc, Adam Kowalski, Malgorzata Kozlowska, and Tomasz Strzalkowski. 1985. A natural language information retrieval system with extentions towards fuzzy reasoning. *International Journal of Man-Machine Studies* 23, 4 (1985), 335 – 367.

[10] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. *CoRR* abs/1808.07036 (2018).

[11] 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)*. ACM, 815–824.

[12] W. B. Croft and R. H. Thompson. 1987. I^{3}R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science* 38, 6 (1987), 389–404.

[13] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA*. 4171–4186.

[15] Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. 2018. TREC Complex Answer Retrieval Overview. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA*.

[16] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[17] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[18] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:arXiv:1803.07640

[19] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy*. ACM, 1183–1192.

[20] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 851–861.

[21] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy*. ACM, 506–516.

[22] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 329–343.

[23] S. E. Levinson and K. L. Shipley. 1980. A conversational-mode airline information and reservation system using speech input and output. *The Bell System Technical Journal* 59, 1 (1980), 119–137.

[24] Alain Loisel, Guillaume Dubuisson Duplessis, Nathalie Chaignaud, Jean-Philippe Kotowicz, and Alexandre Pauchet. 2012. A conversational agent for information retrieval based on a study of human dialogues. In *International Conference on Agent and Artificial Intelligence*. Vilamoura, Portugal, 312–317.

[25] Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne de Sevin, Stéphanie Bioulac, Alain Sauteraud, and Pierre Philip. 2016. Acceptability of Embodied Conversational Agent in a Health Care Context. In *Intelligent Virtual Agents*, David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski (Eds.). Springer International Publishing, 416–419.

[26] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*.

[27] R. N. Oddy. 1977. Information Retrieval through Man–Machine dialogue. *Journal of Documentation* 33, 1 (1977), 1–14.

[28] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*. 1532–1543.

[29] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*. 275–281.

[30] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway*. 117–126.

[31] Esteban A. Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Personality Recognition in Conversations using Capsule Neural Networks. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI'19), Thessaloniki, Greece*. 180–187.

[32] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA*. 109–126.

[33] Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology* 68, 9 (2017), 2101–2115.

[34] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA*. 235–244.

[35] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA*. 32–41.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 5998–6008.

[37] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA*. 2187–2193.

[38] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialogue Systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*. ACL, 515–522.

[39] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*. 404–413.

[40] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan*. ACM, 685–694.

[41] 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 2018, Torino, Italy*. 177–186.