
There may be differences between this version and the published version. You are advised to consult the publisher’s version if you wish to cite from it.

© 2021 Association for Computing Machinery. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ICTIR '21: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval

http://dx.doi.org/10.1145/3471158.3472237

http://eprints.gla.ac.uk/244187/

Deposited on: 16 June 2021
Effective Query Formulation in Conversation Contextualization: A Query Specificity-based Approach

Dipasree Pal  
dipasreepal@gmail.com  
Indian Statistical Institute  
Kolkata, India

Debasis Ganguly  
Debasis.Ganguly@glasgow.ac.uk  
University of Glasgow  
Glasgow, United Kingdom

ABSTRACT
Proactively retrieving relevant information to contextualize conversations has potential applications in better understanding the conversational content between communicating parties. Since, in contrast to traditional IR, there is no explicitly formulated user-query, a research challenge is to first identify the candidate segments of text that in fact require contextualization for better understanding of their content, and then make use of these identified segments to formulate a query and eventually retrieve the potentially relevant information to augment a conversation. In this paper, we propose a generic unsupervised framework that involves shifting overlapping windows of terms through a conversation and estimate a likelihood score of the indicator of an information need for each window. Within our proposed framework, we investigate a query performance prediction (QPP) based approach for scoring these candidate term windows with the hypothesis that a term window that indicates a higher specificity is likely to be indicative of a potential information need requiring contextualization. Our experiments revealed that the QPP approaches of scoring the term windows provide better contextualization than other term extraction approaches. Both pre-retrieval and post-retrieval QPP approaches were observed to yield comparable results in our experiments.

KEYWORDS
Conversational IR, Dialog contextualization, Query Specificity

1 INTRODUCTION
In contrast to traditional IR, where the interaction between a user and the system is essentially restricted to the users submitting queries comprised of keywords, and the system returning a list of potentially relevant documents or passages, the objective of conversational IR is to allow provision for a more engaging user experience, where the users, in order to satisfy their information seeking goal, could start with a broadly scoped query and then guide the search system to return a much focused set of relevant responses [1]. Similar to the evolution of information needs within search sessions of standard IR [8, 26], an increased user engagement during conversational search is likely to facilitate a more focused evolution of these information needs, e.g., a general information need on the planet Uranus can evolve to a question on its peculiarly tilted axis through conversation exchanges between a user and the system [13, 14].

Different from the existing notion of conversational IR that typically involves a human user with information needs and an automated agent seeking to find relevant answers to these information needs, the task that we address in this paper is that of exploring ways of leveraging search systems to better facilitate the comprehension of conversational exchanges between two or more humans, as proposed in [17]. Figure 1 shows an example conversation between two persons (excerpt from the script of the movie ‘Pi’). The objective of a conversational assistance agent, in this case, would be to identify concepts or entities, such as acacia tree and Ming Mecca, that are potential candidates requiring further elaboration. Although the entities requiring contextualization for better comprehension depend on the prior knowledge of the person to whom the conversation is directed to, in our work, we address the task from an objective point-of-view rather than a subjective one.

In contrast to the task of retrieval from verbose queries [18, 27, 30], which involves returning a single ranked list of documents for a query comprising a large number of words, the task of conversation contextualization, that we address in this paper, involves retrieving relevant information for each potentially ‘difficult to comprehend’ concept within a conversation, and then associating this information to each such segment of text. In our work, to identify such text segments indicative of concepts potentially requiring contextualization, we make use of the specificity estimates from the query performance prediction (QPP) literature. More specifically, we hypothesize that the segments of a conversation which leads to better QPP estimates (i.e., these segments representing queries for which an IR model retrieves top-k documents that are substantially different, content-wise, from the rest of the collection) are in fact those which require to be contextualized for a better comprehension.

The main contributions of our paper are summarized as follows.

(1) We propose a general methodology, which given a conversation, computes the likelihood scores of overlapping text segments (windows) for formulating the potential queries (segments of text that are likely to be hard to comprehend).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference '17, July 2017, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-xx/YY/MM ... $15.00
https://doi.org/10.1145/mmmmmm.mmmmmm

1https://www.imdb.com/title/t00138704/
We investigate two different QPP approaches, one pre-retrieval (specifically, average idf [19]) and the other post-retrieval (specifically, NQC [34]) to identify such segments.

Our experiment setup evaluates both the intermediate task of identifying these segments and the end-task of retrieving information relevant to a given conversational dialogue, and we find that effectively identifying the segments mostly leads to retrieving more relevant documents for contextualizing the conversation.

2 RELATED WORK

Somewhat similar to our objective of contextualizing a conversation, previous research has explored tasks of contextualizing short documents, such as tweets [5] or finding context of a query session to suggest the next query in that query session [9, 15, 36]. Retrieving relevant documents from conversations is also similar to the task of associating document retrieval, i.e., a task where a query assumes the proportions of a whole length document, the objective being then to retrieve other similar documents from the collection given such a verbose query (often called a query-document) [16, 37]. Associative document retrieval is particularly useful for related news search (TREC News Track3), and patent prior-art search, for which document reduction approaches such as sub-topic analysis [37], or pseudo-relevance feedback based reduction techniques have been used [16].

Retrieval with verbose queries is also similar to associative document search, a difference being that the verbose queries, in contrast to query-documents, are usually shorter in length, comprised usually of a small number of well-formed sentences [18, 30]. IR approaches specifically targeted for verbose queries usually employ a query length normalization component [3, 27], or transform the verbose query to a weighted term distribution (assigning higher weights to the terms that better describe the information need) estimated from the top-retrieved documents [30].

Selecting terms from verbose queries for query formulation has been reported to improve IR effectiveness, e.g., the use of tf-idf features for text extraction [6], or applying the clarity-based specificity measure at a term-level [22] (similar to the baseline approaches in our experiments), or the use of POS tags and named entities [40].

A query performance prediction (QPP) method, given a query and an IR system, yields a prediction (real-valued) of how easy the query is (known as its specificity), which is an estimate of how effective would the IR system be on the given query [7, 10–12, 12]. QPP approaches have found applications in reducing the length of verbose queries by retaining only the combinations of terms that yield the most well formulated queries. However, existing approaches are mostly supervised in nature trying out a number of different combinations of query terms in deciding their relative utilities based on whether their inclusion or exclusion contribute to increasing or decreasing an IR measure, e.g. average precision (AP) [2, 4, 23]. Our proposed approach directly applies a QPP estimate over a moving window of text to select the ones with higher scores as candidate text segments within a conversation, It thus differs from the existing thread of work on query reduction with QPP features [2, 4, 23] in three important ways. First, our approach is unsupervised and does not rely on a training set of queries. Second, our approach also does not rely on the existence of relevance assessments during the training process. Finally, instead of using a number of possible combinations of terms as sub-queries, which is exponential in the number of query terms, our method relies on formulating queries only with consecutive terms (since in our case, the sub-queries represent text segments that are usually difficult to interpret). This means that the number of sub-queries for which we compute the specificity scores is linear in the number of query terms.

Different to the aforementioned approaches which operate at the fine-grained level of individual terms, in our work we extract features at the level of fixed length sequences of terms (windows). Such a window-driven approach has been reported to work well for taking into account matches in query term positions [28], and also for improving relevance feedback [29, 39].

The task of conversation contextualization that we address in this paper was proposed as a shared task in [17]. However, different to focusing on designing the task itself, this paper differentiates itself from [17] by investigating ways of effectively approaching this task.

3 CONVERSATION CONTEXTUALIZATION

In this section, we describe our proposed approach of contextualizing a conversation with relevant information.

3.1 Extracting Candidate Queries

Given a conversation (Figure 1 shows an example), the first objective is to identify segments of text indicating potential scopes of information needs. To this end, we shift a window of a predefined size $k$ (a parameter) positioned at each word of the given text. Each instance of the window placed at position $p$ is considered to be a text segment, $S_{p,k} = \{w_p, w_{p+1}, \ldots, w_{p+k-1}\}$, where $w_i$ denotes the token at position $i$, $w_i \in \mathcal{V}$ (denoting the vocabulary).

Each segment, $S_{p,k}$, positioned at $p$, is then assigned a score with a generic function $\phi$ that takes as input a text of $k$ words and outputs a likelihood of it being indicative of a potential information need, i.e., $\phi : \mathcal{V}^k \mapsto \mathbb{R}$. We will discuss two concrete realizations of the function $\phi$ that we experimented with in Section 3.2.

A general characteristic of the function $\phi$ is that it should yield a high likelihood score for those segments of text that are lead to retrieving a set of top documents that are focused to a topic and are easily differentiated from the general topic of the collection. In other words, the specificity score of such text segments should be high. After computing the specificity scores for each text segment, we extract the top $m$ of them, where in the context of our problem, $m$ is a known number of concepts that are to be contextualized and is supplied as a part of the input (a more pragmatic approach corresponds to the situation when the number of concepts to be contextualized is not known, which we leave for future exploration). After sorting the text segments by the computed specificity scores ($\phi$), we extract the $m$-top segments with the constraint that the segments are non-overlapping, i.e., referring to back to the example of Figure 1, once ‘acacia tree East’ is selected as a query due to its highest specificity, the window ‘tree East Africa’ is not selected as

http://trec-news.org/
Figure 1: A schematic representation of the use of query specificity estimation for identifying candidate windows of text for query formulation from a given piece of conversation. The example shows that ‘Acacia tree in East Africa’ is a more likely candidate of an information need that the segment ‘I can be your greatest ally’ (another example shown in the figure is ‘Ming Mecca chip’ vs. ‘gold or diamond’). The segment of conversation from which no window eventually turn out to be one of the top-scoring ones is shown with the blue color.
We conduct our experiments using the data released as a part of the RCD (Retrieval from Conversational Dialogues) track [17] at FIRE Conference’17, July 2017, Washington, DC, USA Dipasree Pal and Debasis Ganguly

4 EXPERIMENTS

4.1 Setup

We conduct our experiments using the data released as a part of the RCD (Retrieval from Conversational Dialogues) track [17] at FIRE 2020. To give an overview, the conversations in the RCD dataset constitutes movie script extracts with manually annotated text spans representing information needs requiring contextualization. Each movie script in the RCD dataset is associated with a set of manually judged relevant documents from the Wikipedia target collection (dump of 2019).

Given a verbose query in the form of a movie script excerpt, the task then is to retrieve the relevant documents corresponding to all these manually annotated concepts in the script. Note that these ground-truth annotation spans of segments requiring contextualization are not available to the automated approaches. The only information which is made available as a part of the dataset is the number of such spans, i.e., the number of concepts, \( m \), that would require contextualization. Given the value of \( m \), which ranges from 1 to 3 in the dataset, the task is to retrieve a ranked list of documents for each. To evaluate the quality of retrieval, we adapt the same aggregated approach as used by the RCD track organizers [17], where the set of documents judged relevant for each concept in the conversation is considered relevant for the entire conversation. Again coming back to the example of Figure 1, this means that while evaluating the quality of retrieval for the conversation shown in the figure, the ground-truth set includes the relevant documents for both the concepts ‘acacia tree’ and ‘Ming Mecca chip’.

Since formulating the queries from a given conversation is the core component of our proposed methodology (computing specificity estimates by shifting windows), in addition to the retrieval effectiveness we also report the qualities of the identified queries themselves in terms of their overlap with the ground-truth. This was in fact an intermediate task in the RCD track, and we report the same metrics as also used in the track [17], namely the character n-gram based BLEU score and the word based Jaccard overlap.

As retrieval effectiveness metrics (aggregated over a conversation), we report the mean reciprocal rank (MRR).

4.2 Baselines and Parameter Settings

To test the effectiveness of the proposed window-based specificity approach, as a baseline we employ the standard methodology of term extraction from verbose queries [30]. Specifically, in contrast to selecting segments of text (contiguous terms) as potential queries, this baseline method forms the first query by grouping together the most discriminative terms (highest IDFs) and then forms the second query from the next group and so on. The size of a group (number of query terms), \( k \), is a parameter to the method. The parameter, \( k \), which is interpreted as the number of query terms for the baseline method, and the window size for our method, was varied in the range of 3 to 5. As retrieval model, we employed LM-Dirichlet [41] with \( \mu = 1000 \).

For the \( \varepsilon \)-weighted query formulation (Section 3.3), the value of \( \varepsilon \) was varied within the range of 0 to 0.4 in steps of 0.1, the case \( \varepsilon = 0 \) denoting the degenerate case when all terms outside the selected text spans are discarded.

In addition to investigating the retrieval effectiveness with LM-Dirichlet [41], we also conducted experiments with relevance feedback by applying the standard method of RLM (relevance model) [25], using the linear mixture with query terms [21] commonly known as RM3. We conducted a grid search over the number of pseudo-relevant documents, \( R \), and the number of top-scoring terms \( T \), and found that values of \( R = 10 \) and \( T = 10 \) turned out to be the best (in terms of MRR) for this task.

4.3 Results

In Table 1, we report the results for the query extraction effectiveness from the conversations, in terms of the overlap with the ground-truth information need spans. We observe that in this intermediate step of identification of potential information needs for contextualization, a window-based approach works more effectively, as expected, than a term selection approach which could yield non-contiguous terms as queries. A pre-retrieval specificity function (average IDF) was found to outperform the post-retrieval one (NQC-based) for this intermediate task. Our query extraction results are better than the submitted runs at the RCD track [17], where the best BLEU score was reported to be 0.1090.

Table 2 reports the effectiveness of the conversational contextualization task, which involves the subsequent step of retrieval after identification of the potential query spans (Section 3.1), and then formulating the weighted queries accordingly (Section 3.3). The first row of Table 2 presents the oracle case when the information need spans are known to a retrieval method and presents the effectiveness of the conversational contextualization task that could be achieved in an ideal situation.

The pre-retrieval based specificity also works the most effectively without the application of relevance feedback. It is seen, however, that with the application of RM3 the results obtained with the post-retrieval based specificity (NQC-based) outperforms the ones obtained with a pre-retrieval based specificity estimator. Again the best results obtained in our experiments are significantly higher than the method of employing summarization to extract the key concepts from a conversation, and using them as queries for retrieval.

<table>
<thead>
<tr>
<th>Term selection</th>
<th>Specificity</th>
<th>( k )</th>
<th>BLEU</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term-level Avg IDF</td>
<td>4</td>
<td>0.1459</td>
<td>0.0585</td>
<td></td>
</tr>
<tr>
<td>Term-level Avg IDF</td>
<td>4</td>
<td>0.1459</td>
<td>0.0585</td>
<td></td>
</tr>
<tr>
<td>Term-level Avg IDF</td>
<td>4</td>
<td>0.1459</td>
<td>0.0585</td>
<td></td>
</tr>
<tr>
<td>Term-level Avg IDF</td>
<td>4</td>
<td>0.1459</td>
<td>0.0585</td>
<td></td>
</tr>
<tr>
<td>Window-based Avg IDF</td>
<td>5</td>
<td>0.1623</td>
<td>0.0716</td>
<td></td>
</tr>
<tr>
<td>Window-based NQC</td>
<td>4</td>
<td>0.1113</td>
<td>0.0482</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Query extraction effectiveness from conversations. The optimal value of \( k \) (number of query terms) is shown alongside each method.

---

1http://fire.irs.ies.in/fire/2020/home
The method of soft-masking for computing the weighted distribution of the query terms also turns out to be particularly helpful as evident from the poor results obtained with the degenerate case of $\epsilon = 0$ (corresponding to the situation of hard-masking or only using the extracted segments as queries discarding the other terms) shown in Figure 2. Figure 2 also shows that the retrieval effectiveness is relatively insensitive to the value of $k$ (number of query terms) and that pseudo-relevance feedback (RM3) turns out to be more advantageous for the NQC-based query extraction method than the Avg-IDF one.

5 CONCLUSION

In this paper, we proposed a generic framework of conversation contextualization that first employs a specificity predictor function to identify potential candidates of information need within a conversation, and follows it up by soft-masking the identified regions to formulate a multiple number of weighted queries, one each for the identified text segments. These weighted queries are then for retrieval of potentially relevant documents corresponding to each identified text segments. These weighted queries are then for retrieval of potentially relevant documents corresponding to each identified information need within a conversation. A main advantage of our proposed method is that it is completely unsupervised in nature.

Our experiments showed that a term window based approach works particularly well in comparison to extracting terms independently from a conversation for the task of conversation contextualization. The experiments showed that a post-retrieval based specificity measure of query extraction coupled with pseudo-relevance feedback is the best performing method.

In future, we would like to work on techniques by which it could be possible to predict the likely number of information needs within a conversation (in our current work, we assumed that it is a part of the input).

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>LM-Dir RM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated spans (oracle)</td>
<td>-</td>
<td>0.3603 0.2946</td>
</tr>
<tr>
<td>Avg IDF (term level)</td>
<td>$\epsilon = 0.2, k = 5$</td>
<td>0.1142 0.0893</td>
</tr>
<tr>
<td>Avg IDF (window-based)</td>
<td>$\epsilon = 0.2, k = 5$</td>
<td>0.1807 0.2254</td>
</tr>
<tr>
<td>NQC (window-level)</td>
<td>$\epsilon = 0.2, k = 5$</td>
<td>0.1542 0.2281</td>
</tr>
</tbody>
</table>

Table 2: MRR values obtained with different methods for the task of conversation contextualization.

Figure 2: Sensitivity of the conversation contextualization with respect to the parameters $\epsilon$ (the weights assigned to the text segments with highest specificity scores) and the size of the text segments $k = 3$ (left), $k = 4$ (middle) and $k = 5$ (right).

REFERENCES


