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

ACM Reference Format:
Dipasree Pal and Debasis Ganguly. 2021. Effective Query Formulation in Conversation Contextualization: A Query Specificity-based Approach. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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'[1]). 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).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://www.imdb.com/title/tt0138704/

(2) 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.
(3) 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 associative 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 Track[2]), 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 term 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, 20, 24, 31, 33, 35, 38, 42, 43]. 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, $w_i \in \mathbb{V}$ (denoting the vocabulary).

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

A general characteristic of the function ϕ 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 (ϕ), 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

[2] http://trec-news.org/

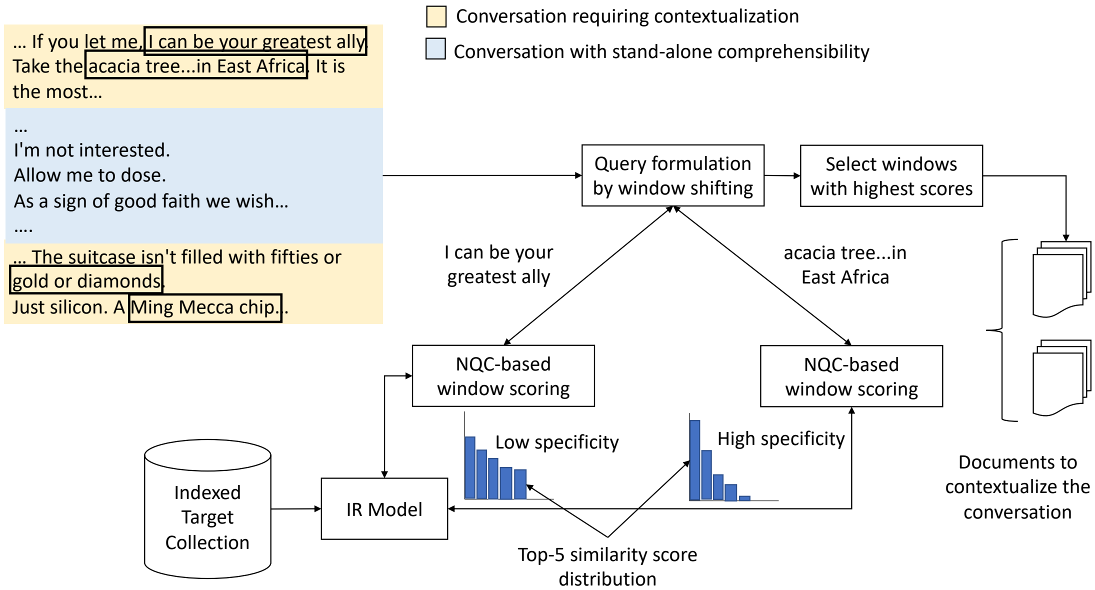

Figure 1: A schematic representation of the use of query specificity estimation for identifying candidate windows of text for query formulation from a given piece of conversation. The example shows that 'Acacia tree in East Africa' is a more likely candidate of an information need that the segment 'I 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.

the next candidate (because it overlaps with a previously selected window).

3.2 Query Scoring Functions

We now describe two concrete realizations of the function ϕ that, given a window of text, returns a specificity estimate. We employ two standard measures for specificity from the QPP literature, first the average IDF [19], and second, the NQC [32, 34].

For a given window of text of size k positioned at p, $S_{p,k}$, the specificity estimate obtained by the first method of average IDFs is given by

$$\phi_{\texttt{idf}}(S_{p,k}) = \frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{N}{d(w_i)}), \tag{1}$$

where $d(w_i)$ denotes the document frequency of the i^{th} token in the conversation, and N denotes the total number of documents in the target collection used for contextualization. Note that this method of specificity computation is a *pre-retrieval* approach, i.e., it relies solely on collection statistics and does not require retrieving a set of documents with the query $S_{p,k}$.

As the second approach, we employ a post-retrieval predictor - NQC (normalized query commitment), which estimates specificity by computing how non-uniform is the distribution of the retrieval status values (RSVs). In other words, in NQC [32, 34], a situation where the scores of a small number of top-retrieved documents are significantly higher than the rest of them (i.e., a skewed heavy tailed distribution associated with a high value of the variance), is assumed to indicate the case that the information need of the query itself is focused. In the context of our problem, this means that a high NQC for a text window S_p is likely to be indicative of a concept requiring further contextualization. Formally speaking,

$$\phi_{\text{NQC}}(S_{p,k}) = \frac{1}{|r(C,Q)|}\sum_{i=1}^{n}(r(D_i,S_{p,k})-\mu)^2, \quad \mu = \frac{1}{n}\sum_{i=1}^{n}r(D_i,S_{p,k}), \tag{2}$$

where $r(D_i, Q)$ denotes the similarity score (RSV) of document D_i retrieved in response to the query Q (in our case the query Q corresponds to the text segment $S_{p,k}$), $r(C, Q)$ represents the similarity of Q with respect to the collection (incorporating the collection statistics), and n denotes the number of top documents used to compute the specificity measure, which was set to 100 in our experiments as prescribed in [34].

Figure 1 schematically describes the idea, where it can be seen that the NQC-based method seeks to harness the different characteristic patterns of the RSV score distributions. While the RSV distribution for a more specific segment (and a likely concept requiring contextualization) is non-uniform (skewed with a heavy tail), the one for a less specific one is more uniform.

3.3 Weighted Query Formulation

After extracting the candidate query segments (the specified m top scoring ones), we assign a weight of $1 - \epsilon$ to the terms belonging to these segments, whereas for the other terms in the conversation, we assign a weight of $\epsilon \in [0, 0.5]$, where ϵ is a parameter. The rationale behind this step of *soft masking* is that we associate a higher emphasis of $1 - \epsilon$ on the terms within the text segments that are candidate information needs, while at the same time we do not completely discard the other terms within the conversation by associating a lower weight to them (note that $1-\epsilon > \epsilon$ because $\epsilon < 0.5$). For a conversation with multiple information need segments ($m > 1$), each candidate segment is soft-masked in its own turn.

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[3]. 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 $\mu = 1000$.

[3]http://fire.irsi.res.in/fire/2020/home

Term selection	Specificity	k	BLEU	Jaccard
Term-level	Avg IDF	4	0.1459	0.0585
Term-level	Avg IDF	4	0.1459	0.0585
Term-level	Avg IDF	4	0.1459	0.0585
Term-level	Avg IDF	4	0.1459	0.0585
Window-based	Avg IDF	5	**0.1623**	**0.0716**
Window-based	NQC	4	0.1113	0.0482

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

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

In addition to investigating the retrieval effectiveness with LM-Dir [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.

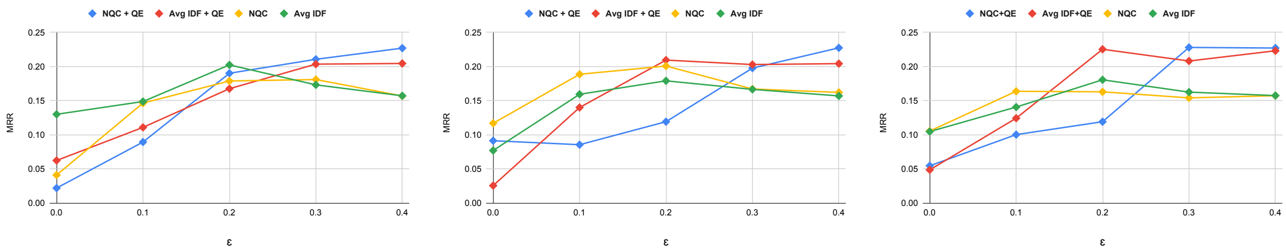

Figure 2: Sensitivity of the conversation contextualization with respect to the parameters - ϵ (the weights assigned to the text segments with highest specificity scores) and the size of the text segments - $k = 3$ (left), $k = 4$ (middle) and $k = 5$ (right).

Method	Parameters	LM-Dir	RM3
Annotated spans (oracle)	-	0.3603	0.2946
Avg IDF (term level)	$\epsilon = 0.2, k = 5$	0.1142	0.0893
Avg IDF (window-based)	$\epsilon = 0.2, k = 5$	**0.1807**	0.2254
NQC (window-level)	$\epsilon = 0.2, k = 5$	0.1542	**0.2281**

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

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 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).

REFERENCES

[1] 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* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3331184.3331265

[2] Jaime Arguello, Sandeep Avula, and Fernando Diaz. 2017. Using Query Performance Predictors to Reduce Spoken Queries. 27–39. https://doi.org/10.1007/978-3-319-56608-5_3

[3] Mozhdeh Ariannezhad, Ali Montazeralghaem, Hamed Zamani, and A. Shakery. 2017. Improving Retrieval Performance for Verbose Queries via Axiomatic Analysis of Term Discrimination Heuristic. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

[4] Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R. Carvalho. 2010. Exploring Reductions for Long Web Queries. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 571–578. https://doi.org/10.1145/1835449.1835545

[5] Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric SanJuan, and Xavier Tannier. 2016. INEX Tweet Contextualization task: Evaluation, results and lesson learned. *Inf. Process. Manag.* 52, 5 (2016), 801–819.

[6] Michael Bendersky and W. Bruce Croft. 2008. Discovering Key Concepts in Verbose Queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 491–498.

[7] David Carmel and Elad Yom-Tov. 2010. Estimating the Query Difficulty for Information Retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 911.

[8] Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. 2014. Overview of the TREC 2014 Session Track. In *Proc. of TREC 2014*.

[9] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2018. Attention-Based Hierarchical Neural Query Suggestion. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1093–1096. https://doi.org/10.1145/3209978.3210079

[10] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 299–306.

[11] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2006. Precision Prediction Based on Ranked List Coherence. *Inf. Retr.* 9, 6 (Dec. 2006), 723–755.

[12] Ronan Cummins, Joemon Jose, and Colm O'Riordan. 2011. Improved Query Performance Prediction Using Standard Deviation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, 1089–1090.

[13] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *CoRR* abs/2003.13624 (2020).

[14] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *SIGIR*. ACM, 1985–1988.

[15] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to Attend, Copy, and Generate for Session-Based Query Suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos

Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 1747–1756. https://doi.org/10.1145/3132847.3133010

[16] Debasis Ganguly, Johannes Leveling, Walid Magdy, and Gareth J. F. Jones. 2011. Patent query reduction using pseudo relevance feedback. In *CIKM*. ACM, 1953–1956.

[17] Debasis Ganguly, Dipasree Pal, Manisha Verma, and Procheta Sen. 2020. Overview of RCD-2020, the FIRE-2020 track on Retrieval from Conversational Dialogues. In *FIRE*. ACM, 33–36.

[18] Manish Gupta and Michael Bendersky. 2015. Information Retrieval with Verbose Queries. *Foundations and Trends® in Information Retrieval* 9, 3-4 (2015), 209–354. https://doi.org/10.1561/1500000050

[19] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A Survey of Pre-Retrieval Query Performance Predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 1419–1420.

[20] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A Survey of Pre-Retrieval Query Performance Predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. Association for Computing Machinery, 1419–1420.

[21] Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *TREC 2004*.

[22] Giridhar Kumaran and Vitor R. Carvalho. 2009. Reducing Long Queries Using Query Quality Predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 564–571.

[23] Giridhar Kumaran and Vitor R. Carvalho. 2009. Reducing Long Queries Using Query Quality Predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 564–571. https://doi.org/10.1145/1571941.1572038

[24] Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel. 2011. A Unified Framework for Post-Retrieval Query-Performance Prediction. In *Proceedings of the Third International Conference on Advances in Information Retrieval Theory (ICTIR'11)*. 15–26.

[25] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proc. of SIGIR '01* (New Orleans, Louisiana, USA). ACM, New York, NY, USA, 120–127.

[26] Nir Levine, Haggai Roitman, and Doron Cohen. [n.d.]. An Extended Relevance Model for Session Search. In *Proc. of SIGIR 2017*. 865–868.

[27] Yuanhua Lv. 2015. A Study of Query Length Heuristics in Information Retrieval. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1747–1750. https://doi.org/10.1145/2806416.2806592

[28] Yuanhua Lv and ChengXiang Zhai. 2009. Positional Language Models for Information Retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 299–306. https://doi.org/10.1145/1571941.1571994

[29] Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 206–214. https://doi.org/10.1145/290941.290995

[30] Jiaul H. Paik and Douglas W. Oard. 2014. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. Association for Computing Machinery, New York, NY, USA, 131–140. https://doi.org/10.1145/2661829.2661957

[31] Haggai Roitman. 2017. An Enhanced Approach to Query Performance Prediction Using Reference Lists *(Proc. SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 869–872.

[32] Anna Shtok, Oren Kurland, and David Carmel. 2009. Predicting Query Performance by Query-Drift Estimation. In *Advances in Information Retrieval Theory*, Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rüger, Milad Shokouhi, Dawei Song, and Emine Yilmaz (Eds.). Number 5766 in Lecture Notes in Computer Science. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-04417-5_30

[33] Anna Shtok, Oren Kurland, and David Carmel. 2010. 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 (SIGIR '10)*. 259–266.

[34] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (May 2012), 11:1–11:35 pages.

[35] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (2012), 35 pages.

[36] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/2806416.2806493

[37] Toru Takaki, Atsushi Fujii, and Tetsuya Ishikawa. 2004. Associative Document Retrieval by Query Subtopic Analysis and Its Application to Invalidity Patent Search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (Washington, D.C., USA) *(CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 399–405.

[38] Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In *Proceedings of the 22nd Australasian Document Computing Symposium (ADCS 2017)*. Association for Computing Machinery, Article 11, 4 pages.

[39] Jinxi Xu and W. Bruce Croft. 2000. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Trans. Inf. Syst.* 18, 1 (Jan. 2000), 79–112.

[40] Xiaobing Xue, Samuel Huston, and W. Bruce Croft. 2010. Improving Verbose Queries Using Subset Distribution. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1059–1068.

[41] Chengxiang Zhai and John Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. 334–342.

[42] Yun Zhou and W. Bruce Croft. 2006. Ranking Robustness: A Novel Framework to Predict Query Performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 567–574.

[43] Yun Zhou and W. Bruce Croft. 2007. Query Performance Prediction in Web Search Environments. In *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 543–550.