Document Summarization for Answering Non-Factoid Queries

Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft and Mark Sanderson

Abstract—We formulate a document summarization method to extract passage-level answers for non-factoid queries, referred as *answer-biased summaries*. We propose to use external information from related Community Question Answering (CQA) content to better identify answer bearing sentences. Three optimization-based methods are proposed: (i) query-biased; (ii) CQA-answer-biased; and (iii) expanded-query-biased, where expansion terms were derived from related CQA content. A learning-to-rank-based method is also proposed that incorporates features extracted from related CQA content. Our results show that even if a CQA answer does not contain a perfect answer to a query, their content can be exploited to improve the extraction of answer-biased summaries from other corpora. The quality of CQA content is found to impact on the accuracy of optimization-based summaries, though medium quality answers enable the system to achieve a comparable (and in some cases superior) accuracy to state-of-the-art techniques. The learning-to-rank-based summaries, on the other hand, are not significantly influenced by CQA quality. We provide a recommendation of the best use of our proposed approaches in regard to the availability of different quality levels of related CQA content. As a further investigation, the reliability of our approaches was tested on another publicly available dataset.

Index Terms— document summarization, answer-biased summaries, non-factoid queries, CQA, optimization, learning-to-rank

1 Introduction

Current search engines usually present single direct answers on a search result page for some popular *factoid* queries (e.g. current weather) [1], and for some entity queries. Some major search engines have also started to present single passages (so-called featured snippets) in response to more verbose informational queries. According to the Moz SERP features tracker[1], in 2017, these passages appears in 15% of queries submitted to Google, though with some errors.[2]

Such direct answers can improve a user's search experience [1]–[3]. They may also lead to *good abandonment* [4], where users find what they need in the result page and therefore do not need to read the full document. By removing the document reading step, user time can substantially be saved. This is as reported by Smucker and Clarke [5] that users spent 67% of their searching time reading webpages. Direct answers provide the most benefit to users who search on devices with limited screen size and low bandwidth (e.g. mobile search) as clicking through can incur additional costs at the user's end.

While *non-factoid* queries are the most frequently asked questions on the web [6], [7], research on finding answers for this type of query has not been extensively explored. Some past work was conducted to generate passage-level answers to a non-factoid query [3], [8]–[10]. However these approaches do not explore the idea of using automatic summarization. We argue that summarization techniques can be beneficial in tackling this problem because answers to non-factoid queries may consist of a number of sentences scattered in the underlying document (potentially with some overlap in the content) [9].

In this work, we explore the idea of extracting a summary from each retrieved document that is expected to contain answers to a non-factoid query, called an *answer-biased summary*. This idea is also motivated by a previous work [4] which suggested that improving search result summaries may potentially drive to good abandonment. It is as shown in [11], [12] that user can also address their information needs by viewing summaries only.

In contrast to query-biased summaries that are mainly indicative of user's queries (see Fig 1), our summaries are designed to hint at the whereabout of likely answers, in light of reducing the user's cognitive effort [5]. Displaying answer-biased summary from each retrieved document (as opposed to a single direct answer) is useful for some types of queries that may require users to see answers from various sources, such as a query displayed in Fig 1.

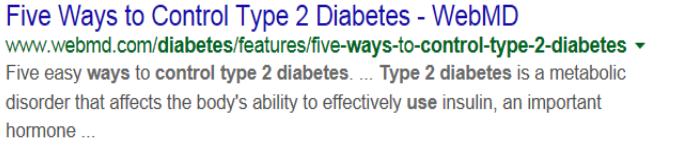

Fig. 1 An example of search result of Google search engine for query "what are methods to control type II diabetes?"

We propose to use external information from related Community Question Answering (CQA) content to guide the extraction of answer-biased summaries from retrieved documents. It is known that questions posted on CQA websites may be driven by similar information needs as

- *E. Yulianti is with RMIT University and Universitas Indonesia. Email: evi.yulianti@rmit.edu.au and evi.y@cs.ui.ac.id*
- *R.-C. Chen, F. Scholer, W. B. Croft and M. Sanderson are with RMIT University. E-mail: {ruey-cheng.chen, falk.scholer, bruce.croft, mark.sanderson}@ rmit.edu.au.*

[1]http://mozcast.com/features
[2]http://searchengineland.com/googles-one-true-answer-problem-featured-snippets-270549

the "tail queries" in web search [13]. CQA websites such as Yahoo! Answers contain a large number of questions and human-curated answers. This data is valuable for the research of non-factoid question answering not only because the curated answers can be reused, but the rich set of question-answer structures can be exploited to improve answer finding. This information could be useful to assist with the identification of answer-bearing sentences.

As will be shown in our experiments, our approach has a number of advantages over an approach of using CQA content directly as the answer [8]: (i) the accuracy of our summaries is better than CQA answers; (ii) the related CQA answers that do not necessarily contain a perfect answer to the query are useful in extracting answer-biased summaries from documents; and (iii) the learning-to-rank-based model can also improve the accuracy of summaries for which the related CQA answers are not available.

In this paper, we investigate the following research questions:

RQ1: Can we use related CQA content to extract better answer-biased summaries from documents?

RQ2: Does the quality of related CQA content affect the accuracy of generated summaries?

The followings are our contributions in this work:
- We propose a novel use of CQA content in a summarization algorithm for locating answer-bearing sentences in the document.
- We propose three optimization-based methods and a learning-to-rank-based method for answering non-factoid queries. These methods are empirically evaluated against state-of-the-art techniques.
- We analyse the effect of quality of related CQA content on our proposed methods. Then, we give recommendations on the best use of our methods in regard to the availability of different quality levels of CQA answers.
- We conduct a set of extensive experiments on two datasets: WebAP (Sec 5 & 6), and MSMARCO (Sec 8.2).

2 Related Work

2.1 Question Answering (QA)

Question Answering (QA) is an information retrieval task that returns answers in response to natural language questions. Commonly supported question types in this research include factoid, list, and definition questions [14]. Answering different types of questions generally relies on different techniques. Previously, much of the attention in the research of question answering has focused on answering factoid and list questions, which are the main themes of the TREC QA track [15].

Our work is different from TREC QA as we focus on non-factoid questions, such as the ones from TREC Terabyte topics: *"What allegations have been made about Enron's culpability in the California Energy crisis?"*, in which it may not be satisfied with just one or a list of factoids. A recent method in factoid QA that has superior performance on TREC QA track data has also been shown to perform poorly for these kinds of questions [16].

2.2 Community Question Answering (CQA)

Community Question Answering (CQA) is a service that allows users to post questions and elicit answers from other peers. Major CQA websites, such as Yahoo! Answers, Quora, and Stack Overflow, continue to see a growing user base. It is reported that Yahoo! Answers (YA) attracted 7,000 questions and 21,000 answers every hour in 2012.[3] This sheer amount of data has attracted a lot of research activities: predicting answer quality in CQA [17]; predicting the satisfaction of the original question asker [18] and web searcher [19] with CQA answers; answering factoid [20] and how-to web queries [8]; enhancing document summaries [21]; and summarizing CQA answers [22], [23].

2.3 Non-Factoid Question Answering

Finding answers for non-factoid queries remains a critical challenge in web question answering, and one difficulty is the vocabulary mismatch between questions and answers. Keikha et al. [9] has shown that state-of-the-art passage retrieval methods that focus on topical relevance are not effective for this task. A recent forum that is related to answering non-factoid questions (that came from real YA users) is the TREC LiveQA track [24]. The quality of the best performing run in this track is shown still far from human level, indicating the complexity of the task.

Most previous studies on this task retrieved answers from CQA collection [8], [25]–[27]. Xue et al. [25] combined a translation-based language model for the question part and a query likelihood language model for the answer part. Surdeanu et al. [26] used similarity, translation, density, and web correlation features for answer ranking. Weber et al [8] extracted tips answers for which the questions are perfect matches to the query. Shtok et al. [27] retrieved the most similar question and then validated the associated answer. Our work is different in which we extract an answer-biased summary from each retrieved document with the help of related CQA content.

Some past work have been conducted to extract passage level answers for non-factoid questions from documents. Soricut and Brill [10] studied question answering beyond factoid questions using statistical translation model. Bernstein et al. [2] extracted inline direct answers for search results using a paid crowdsourcing service, which is costly to be implemented for a huge number of questions. Wu et al. [28] exploited knowledge from a social Q&A collection to build a classifier of noisy sentences for each question focus. Yang et al. [16] proposed using semantic features for answer sentence retrieval. Our work mainly departs from these efforts in the way the task is formulated (as a document summarization problem).

2.4 Summarization

Most research in document summarization focusses on extracting relevant summaries [21], [29]–[36]. The document summaries displayed in search results are usually biased to the query and are aimed to help users identify relevant documents more quickly [29]. Little attention has

[3] http://searchengineland.com/yahoo-answers-hits-300-million-questions-but-qa-activity-is-declining-127314

been paid to generating document summaries that contain answers to the query.

Some previous work has utilized external resources to enhance relevant summaries of documents using social media [21],[32]. In our optimization-based methods, the related CQA answers are used to determine the importance of words in the documents, or apply query expansion, before an optimization is conducted to generate good coverage summaries [30]. It is important to note that here our methods are different from pseudo-relevance feedback [37], since the external information is not used for ranking purposes. In our learning-to-rank-based method, a CQA feature was incorporated into query-biased summarization features [31], [33].

Some approaches to document summarization exploit information from internal documents only [30], [34], [36]. Ko et al. [34] performed query expansion from assumed relevant sentences to generate snippet. Bando et al. [36] and Losada [35] in general used the same technique but obtained expansion terms from initially retrieved top documents [36]. Takamura and Okumura [30] formulated a summarization task as a maximum coverage problem.

Some previous work has also studied CQA answer summarization to address the "incomplete answer" problem [22], and summarize answers for yes/no questions [23]. In this work, we do not attempt to summarize CQA answers, and instead use this data to guide the extraction of answer-biased summaries from documents.

3 Data Collection

3.1 Queries, Documents, and Ground Truth Answers

Given a non-factoid query and a corresponding set of documents, our task is to generate an answer-biased summary from each document. In this case, a test collection of documents, queries, relevance judgments, and ground truth answers are needed. For this purpose, we use the WebAP[4] dataset which consists of 82 long queries, and their corresponding top 50 documents that were annotated for the answer passages.

Keikha et al. [9] created this dataset based on the GOV2 collection and description queries from the TREC Terabyte track. They ran the queries using a Sequential Dependence Model [38] retrieval function. The relevant documents were selected from the fifty top-ranked documents, which were then examined manually for answer passage annotation. The authors used five levels of relevance based on completeness and conciseness criteria (Perfect, Excellent, Good, Fair, and Bad). Across 8,027 answer passages identified, 43% were annotated as Perfect, 44% Excellent, 10% Good, and 3% Fair. Figure 2 shows a sample of queries and answer passages.

To examine the effectiveness of our approach to finding answers in a document, we only considered documents that are relevant to the query because answers do not exist in the non-relevant ones. We took up to ten documents for each query from the WebAP dataset. To ensure that we use high quality answer references, we created ground truth answers by focusing only on "Perfect" and "Excellent" answers. Queries that do not have ground truth answers were removed. In total, there are 80 queries, 710 documents and 3,298 ground truth answers in the final dataset.

Query : "What would cause a lowered white blood cell count?
Answer:
(*Perfect*) Cancer chemotherapy can dramatically lower the CD4 count. In general, the CD4 count goes down as HIV disease progresses.
(*Excellent*) HIV binds to the surface of CD4 cells, enters them, and either reproduces immediately killing them in the process, or remains in a resting state, reproducing when the cell becomes active.
(*Good*) The T-cell count can vary, depending on such differences as the time of day and the test was done, what laboratory did the test, and whether certain health problems were present the day the blood was drawn.

Fig. 2 Sample Question and Answers in WebAP Dataset

On average, only 7.1% of sentences in the document contain some level of answer to the query. The statistics for Perfect, Excellent, Good, and Fair are respectively 3.6%, 2.7%, 0.6%, and 0.1%. This indicates that the task of extracting answer-biased summaries is challenging.

3.2 Related CQA Answers

To obtain related CQA answers for a particular query, we simply use the retrieval function provided by the CQA website. We chose to use Yahoo! Answers (YA) because of the coverage of question topics, which suits our general web queries better. For simplicity, we stick with using only one such CQA backend in our experiment.

For retrieving related CQA answers, each query is first submitted to YA search engine and the top **10 results** are taken. Each of these results is a pair of YA <question, best answer>. A question in YA can have many answers, but only one of them can be the *best answer*, selected either by the original asker or the community. The best answers from the top 10 matching questions are then collectively served as **related CQA answers**.

Of 80 queries in our dataset, 45 have returned results from YA search engine, while the rest have 0 result. 305 results in total were retrieved for these queries, and the average number of results for each query is 6.8. These 45 queries correspond to 419 documents and 1,892 ground truth answers that are to be included in our experiment described in section 5. The remaining queries are also investigated in the experiment described in section 6.

3.3 Quality Judgment of Related CQA Answers

To analyze the influence of different levels of CQA answer quality on the generated summaries, we took a crowdsourcing approach to annotate the retrieved CQA data. We used a crowdsourcing service CrowdFlower[5] (CF) to collect the quality assessment of the related CQA answers. Note that automatically predicting the quality of CQA answers [17], [19] is beyond the scope of this paper.

[4] http://ciir.cs.umass.edu/downloads/WebAP/

[5] http://crowdflower.com

3.3.1 CrowdFlower Design

For each question we posted on CF, a search query and a pair of YA <question, best answer> related to the query was presented. CF workers were asked to assess the relevance of both the YA question and answer with respect to the query. A four level scale by Sormunen [39] was used for **question relevance**: 0=Irrelevant, 1=Marginally relevant, 2=Fairly relevant, and 3=Highly relevant. Then, a five level scale by Keikha et al. [9] was used for **answer relevance**: 0=Bad, 1=Fair, 2=Good, 3=Excellent and 4=Perfect. We used different scale for answer relevance because we also emphasize the consiceness of answer which is related with whether it contains noise (irrelevant materials). This is not covered in Sormunen's level.

Each of our questions in CF was judged by 5 workers. In total, there were 305 questions in our experiment, following the total of related CQA answers in our dataset. We created some gold questions to filter out untrusted answers, in two different ways. **First,** we displayed a random pair of YA <question, best answer> taken from different queries. The workers will fail this gold question if they do not judge "bad" for the answer relevance. **Second**, we replaced a YA answer for a particular query with a "perfect" answer taken from our ground truth answers. The workers will fail if they judge "bad" for the answer relevance. The average accuracy of trusted contributors in answering gold questions is 97%. We collected 1,528 trusted and 48 untrusted judgments. The total cost spent was $71. The ethic approval was obtained from RMIT University to conduct this crowdsourcing work.

3.3.2 CrowdFlower Result

To get an overall score of question and answer relevance, we took mean scores across five judgments and rounded it to the nearest integer. To analyze the agreement among judgments given by multiple contributors, we computed the ICC (Intra-Class Correlation) coefficient [40]. This measure was chosen as the judgments were ordinal variables, and each question was judged by a different set of random contributors [41]. The resulting ICCs for question and answer relevance are respectively 0.69 (good, p<0.01) and 0.76 (excellent, p<0.01). This shows that CF workers had a high degree of agreement [42].

Most of the questions as well as answers were judged as having a medium level of relevance. The statistics for question relevance are 21.6%, 33.8%, 36.4%, and 8.2%, respectively for Irrelevant, Marginally Relevant, Fairly Relevant, and Highly Relevant. The statistics for answer relevance are 16.1%, 26.2%, 33.1%, 20.3%, and 4.3%, respectively for Bad, Fair, Good, Excellent, and Perfect. To understand the relationship between relevance of both questions and answers with respect to a given query, a Spearman correlation was calculated. A strong relationship (r=0.76, p<0.01) [43] was found.

3.3.3 Splitting Data According to the Quality of Related CQA Answers

Two different proxies were used to identify the quality of related YA answers: (i) question relevance; and (ii) answer relevance. The data were split based on these quality proxies. Figure 3 illustrates an example of the data splitting process based on the answer relevance for a query for which the related YA answers were assessed as having 2 bad, 2 fair, 2 good, 2 excellent, and 2 perfect answers. In this case, the 2 bad YA answers, the query, and all documents associated with the query were put into group A_bad. The same procedure is applied to populate the other four answer relevance level groups: A_fair, A_good, A_excellent, and A_perfect.

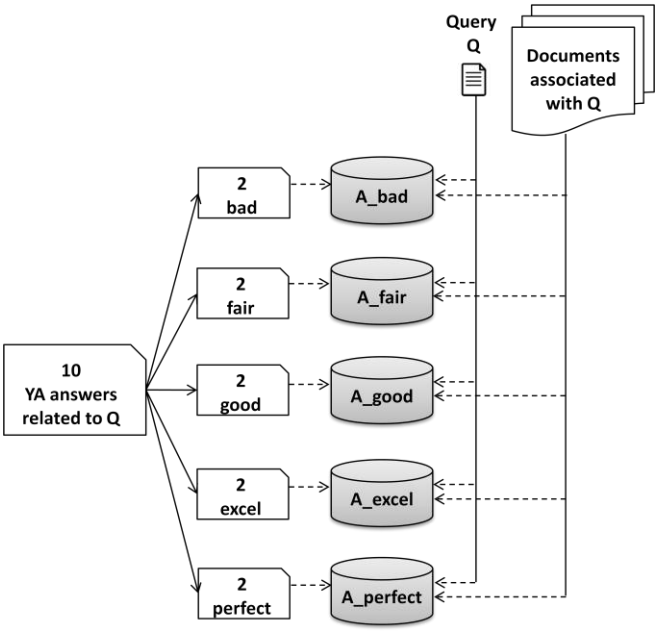

Fig. 3 An Example of Splitting Data Based on Answer Relevance for a Given Query

Note that it is most likely that the related YA answers for a particular query do not cover all the answer relevance levels. For example, the related YA answers for a particular query could be assessed as 4 bad, 3 good, and 4 perfect. In this case, the data corresponding to this query will not exist in two other groups: A_fair and A_excellent. As a result, the number of queries and documents in each group could be different.

TABLE 1
THE STATISTICS OF DATA IN EACH GROUP

Question Relevance Level Group				Answer Relevance Level Group			
Group	#Q	#D	#YA	Group	#Q	#D	#YA
Q_irrel	28	258	66	A_bad	22	205	49
Q_margin_rel	33	308	103	A_fair	33	306	80
Q_fair__rel	28	270	111	A_good	34	318	101
Q_high_rel	9	84	25	A_excel	24	233	62
				A_perfect	9	83	13

After splitting the data based on question relevance and answer relevance, we have four question relevance groups and five answer relevance level groups. The statistics of number of queries, documents, and YA answers in each group are described in Table 1. We use these groups in our subsequent experiment of extracting summaries using individual quality of CQA answers (see section 5.2).

4. SUMMARIZATION METHODS

4.1 Proposed Optimization-based Methods

This section first describes the original summarization model proposed by Takamura and Okumura [30] (see Eq 1), and then introduces three methods to improve this basic model. The model has been shown to beat the DUC 2004 best performing method [30]. It generates a summary containing a set of sentences that maximize coverage of important words while removing redundancy. It also takes into account the relevance to the document as controlled by λ. Lower λ indicates that summaries are generated with less redundancy.

It is important to note that this method is not based on sentence ranking, but an optimization of Eq 1 to select an optimal set of sentences with maximum coverage of important words. We used CPLEX[6] to solve the optimization problem using a branch-and-bound method.

$$(1-\lambda)\sum_j w_j z_j \; + \; \lambda \sum_i \left(\sum_j w_j a_{ij}\right) x_i \qquad (1)$$

$$s.t. \sum_i c_i x_i \le K \; ; \quad \forall j, \sum_i a_{ij} x_i \ge z_j \; ;$$

$$\forall i, x_i \in \{0,1\} \; ; \; \forall j, z_j \in \{0,1\}$$

Here, w_j is the weight of word e_j in the document; z_j is a binary value that denotes coverage of word e_j in a summary; a_{ij} is a binary value that denotes coverage of word e_j in sentence s_i ; x_i is a binary value that denotes the selection status of sentence s_i ; c_i denotes the cost of selecting sentence s_i, in this case it is the number of words in sentence s_i. The first constraint $\sum c_i x_i \le K$ specifies that the length of a summary cannot exceed K, which is set to fifty words in this work, following the settings in recent work [9] in answer finding. The length of each summary is then guaranteed to be less than or equal to fifty words. The second constraint $\sum a_{ij} x_i \ge z_j$ relates to word coverage in the summary, where word e_j is covered when at least one sentence containing this word is selected.

In the original model, words are weighted based on a document content. We use the TF-IDF formula in Eq 2 for word weighting. We call this original method ***DocOpt*** since this uses optimization of Eq 1 and determines the importance of words based on document content:

$$w_j = tf_{j,doc} \times idf_j \qquad (2)$$

where $tf_{j,doc}$ is the frequency of word e_j in the document; and idf_j is the inverse document frequency of word e_j in the GOV2 web collection, calculated as follows:

$$idf_j = ln(1 + \frac{n}{df_j}) \qquad (3)$$

where n is the total number of documents in the web collection and df_j is the number of documents in the web collection that contains word e_j. In our experiment, a preprocessing step was done by removing stop words, and stemming using the Krovetz stemmer on document as well as related CQA answers.

In this work, we propose three variants to this initial ***DocOpt*** summarization model:

1. ***QueryOpt (query-biased)***
   This method adapts the above model to generate *query-biased* summaries. Intuitively, it generates summaries that cover as many important query words as possible and simultaneously minimizes redundancy. Words in the document are weighted based on their occurrences in the query, as follows:

$$w_j = tf_{j,query} \times idf_j \qquad (4)$$

   where $tf_{j,query}$ is the frequency of word e_j in the query. As a result, sentences in the document are represen ted by query terms that they contain. So this method is more effective when the queries (or questions) are long, such as used in [24]. The calculated weights of words are then incorporated into the model in Eq 1 that is optimized to generate summaries. This method is highly efficient since the optimization stage only takes into account query words and the sentences containing those words. We set λ in Eq 1 to 0.1, following our best setting in TREC LiveQA track [44].

2. ***AnswerOpt (CQA-answer-biased)***
   This method identifies words that are likely to be used in these answers, and then generates summaries that cover as many of these words as possible. Intuitively, words that appear in many related CQA answers are assigned a high weight (see Eq 5). As specified in section 3, the related CQA answers for a particular query consist of the best answers for the top 10 matching questions. The answer for a matching question retrieved in the lower position is likely to be less relevant. Therefore, a penalty of the log of an answer's position in the CQA search result list is applied:

$$w_j = \left(\sum_{p=1}^{|CQA|} \frac{tf_{j,answe\ r_p}}{\ln(1+p)}\right) \times idf_j \qquad (5)$$

   where $tf_{j,answer\_p}$ is the frequency of word e_j in the answer at p-th position; $|CQA|$ is the total number of related CQA answers for the query (i.e. maximum of 10). We performed a 9-fold cross validation (CV) to optimize λ in Eq 1 in range [0.0,…, 1.0] with step of 0.1, that maximizes the ROUGE-2 score. The choice of 9-fold makes a balanced subdivision of the 45 queries. The average optimal λ value was 0.2.

3. ***ExpQueryOpt (expanded-query-biased)***
   This method applies query expansion from related CQA answers and then performs *QueryOpt* method as described above. For each document to summarize, Eq 5 is calculated to weight the words in the document based on its related CQA answers. Top k words with the highest weight are then used to expand the original query. Next, *QueryOpt* method is applied using the expanded query. We experiment with different number of expansion terms k, starting from 1 to 10, and report the results in section 5.1.

4.2 Proposed Learning-to-rank-based Method

This proposed method uses a learning-to-rank approach and a set of features. We incorporate a feature extracted from related CQA answers into six Metzler and Kanungo's features (MK) that have been shown to perform well in previous work [31], [33]. This proposed method is referred as ***MK++***. This CQA feature is computed using Eq 6. It is a total weight of words in the sentence, where the

[6] http://www-01.ibm.com

words weight calculation is based on the related CQA answers as described in Eq 5 in the AnswerOpt method above. Sentence S that contains many words that indicate answers to the query Q is assigned a high CQA feature value:

$$f_{CQA}(Q,S) = \sum_{j=1}^{|S|} w_j = \sum_{j=1}^{|S|} \left(\sum_{p=1}^{|CQA|} \frac{tf_{j,answer\_p}}{ln(1+p)}\right) \times idf_j \quad (6)$$

where w_j is the weight of word e_j in sentence S; and $|S|$ is the sentence length. The description of other variables is similar to those described in AnswerOpt method above.

The six MK features consist of 4 query-dependent and 2 query-independent features [31]:

- *Exact Match*. A binary feature indicating whether the exact query occurs in the sentence.
- *Term Overlap*. The fraction of query terms that occur in the sentence.
- *Synonyms Overlap*. The fraction of query terms as well as their synonyms, obtained from WordNet[7] that occur in the sentence.
- *Language Model Score*. Calculated as the log likelihood of the query being generated from the sentence.
- *Sentence Length*. The number of terms in the sentence.
- *Sentence Location*. The relative location of the sentence within the document.

We used the coordinate ascent (CA) algorithm implemented in *RankLib*[8] to train a linear model over the annotation of answer passages. Answer quality was mapped into numerical graded relevance, i.e., Bad=0, Fair=1, Good=2, Excellent=3, and Perfect=4. A 9-fold cross validation (CV) is applied to optimize the hyperparameters based on NDCG@*k*. Here, *k* was set to three based on the average number of sentences in the 50-word summaries in initial experiment (i.e. 2.7). The choice of 9-fold CV was as described in section 4.1. The CV was repeated for 10 times to avoid data overfitting.

4.3 Baseline Methods

To examine the effectiveness of our methods, the results for several baselines are reported.

4.3.1. Lead

This summary is based on the leading 50 words in the document (including title) which has often been used as a strong baseline of relevant summaries [29].

4.3.2. DocOpt

This is the generic summary produced using an approach proposed by Takamura and Okumura [30] that has been described in section 4.1. A 9-fold cross validation is performed to tune λ parameter, and the average value across 9 folds is found to be 0.28.

4.3.3. MEAD

This is a query-biased summary generated using MEAD summarizer [45] by implementing Centroid, Position, Length, and QueryCosine features. It has been used as a strong baseline in reflecting the document relevance [46].

4.3.4. RelSent

This is a query-biased summary that was generated using query expansion from assumed *relevant sentences* by Ko et al. [34]. Comparison with this baseline allows us to examine the merit of using external CQA resources in the expansion process. This method has been shown to be superior to search engine snippets [34]. We followed the best parameter setting reported in the original paper.

4.3.5 LCA

This is a novel summary that was generated using query expansion from related YA answers by applying LCA (Local Context Analysis) technique [35], [36], [47]. None of previous work has applied query expansion using external CQA resources for document summarization.

A comparison to this baseline helps to understand to what extent our proposed methods perform against other methods that use the same external CQA resources. We implemented LCA at document level [36] to weight terms in the related YA answers according to their co-occurrences with the query terms. Top N terms with the highest weights are used to expand the query. We set N equals to 45 following the best setting of Bando et al [36]. Following [35], [36], the sentences are then ranked according to their similarity with respect to the expanded query using the sentence ranking method by Allan et al [48].

4.3.6. QL

This is a state-of-the-art relevant passage that was retrieved using query likelihood passage retrieval method as implemented in *Galago*.[9] This passage was used in recent work by Keikha et al. [9] to examine the effectiveness of current passage retrieval method in finding answer. To generate a representative passage for a document, we first retrieve passages in the document with respect to the query, and took the top-ranked passage. The length of a passage was set to 50 words, with an overlap of 25 words, following the setting in [9].

4.3.7. MK

This is a state-of-the-art query-biased summary that was generated using a learning to rank approach and six features that have been shown to perform well in previous work [31], [33] that have been explained in section 4.2. They trained the MK model using an annotation of *relevant sentences* instead of answer passages, which therefore makes our MK summaries to be a stronger baseline than the summaries generated in the original paper [31].

4.4 Evaluation Metric

The quality of 50-word summaries was evaluated using ROUGE [49], by comparing the produced summaries with the ground truth answers. As our results on all ROUGE variants exhibit similar trends, we chose to report only the ROUGE-1 (unigram overlap) and ROUGE-2 scores (bigram overlap) for brevity. In the calculation of ROUGE scores, the maximum value of term overlap between the generated summary and the ground truth answers was used, as suggested in Keikha et al. [9].

[7] http://wordnet.princeton.edu
[8] https://sourceforge.net/p/lemur/wiki/RankLib/
[9] http://www.lemurproject.org/galago.php

5 Extracting Summaries When Related CQA Answers Are Available

A series of experiments were carried out to investigate the effectiveness of our proposed methods. The availability of related CQA answers to the query was taken into account because certain methods were built upon such resources. We used a subset of the WebAP data that contains 45 queries that have related CQA answers. The 419 documents and 1,892 ground truth answers were associated with these queries. In section 5.1, an original set of related CQA answers returned from YA search engine for each query, was used to generate summaries. This setting simulates a condition of using mixed quality answers. In section 5.2, summaries were generated using individual quality of related CQA answers that have been identified in section 3.3 above. This setting simulates the existence of "answers quality predictor" [17], [19] and is conducted to better understand the accuracy of our proposed methods when using different quality of CQA answers.

5.1 Using Mixed Quality of CQA Answers

All our proposed methods were set up following the descriptions in sections 4.1 and 4.2. Table 2 summarizes the experimental results. *Lead*, *MEAD*, and *DocOpt* were found the least effective. While *RelSent* was shown effective in producing relevant snippets [34], it failed to extract accurate answer-biased summaries. The *LCA* baseline, which uses CQA answers to perform query expansion, also can only achieve comparable performance to the *QL*. The *MK* baseline gave a stronger result than *LCA* and *QL*.

TABLE 2
Summary Accuracy Using Mixed Quality CQA Answers

Method	ROUGE-1	ROUGE-2
Lead	0.305	0.087
MEAD	0.375	0.128
DocOpt	0.404	0.175
RelSent	0.431	0.212
LCA	0.476	0.264
QL	0.494 ^{r}	0.278 ^{r}
MK	0.515 ^{r,l}	0.297 ^{r,l}
QueryOpt	**0.528** ^{r,l,q}	**0.325** ^{r,l,q,m}
AnswerOpt	0.517 ^{r,l,q}	0.291 ^{r}
ExpQueryOpt	0.519 ^{r,l,q}	0.302 ^{r,l}
MK++	0.520 ^{r,l,q}	0.310 ^{r,l,q,m}

The proposed methods are printed in boldface. Superscripts r, l, q, and m indicate a significant difference respectively against RelSent, LCA, QL and MK as measured by paired t-test ($p<0.05$).

The two methods with the highest accuracy were *QueryOpt* and *MK++*, which managed to outperform all the baselines significantly. Although the scores of *QueryOpt* were a little higher than *MK++*, the differences were not significant based on a t-test (The corresponding *p*-value for ROUGE-1 and ROUGE-2 are 0.39 and 0.28, respectively). The fact that *QueryOpt* achieved the highest scores in this experiment echoed our previous results on the TREC LiveQA track [44].

All of our variants of optimization-based methods (i.e. *QueryOpt*, *AnswerOpt*, and *ExpQueryOpt*) enhanced the original *DocOpt*. The enhanced method *MK++* was also found to significantly outperform the original *MK*, affirming that using CQA in answer finding has a clear advantage that benefits the state-of-the-art techniques.

ExpQueryOpt failed to improve *QueryOpt* method. *ExpQueryOpt* and *AnswerOpt* also could not improve state-of-the-art *QL* and *MK*, but their accuracies were still shown to be comparable. We suspect that the CQA quality may have an influence on this result. This hypothesis is investigated in the next subsection. The effect of query expansion in *ExpQueryOpt* technique on the accuracy of summaries generated using different quality of CQA answers is discussed in detail in section 5.2.3.

ExpQueryOpt result reported in Table 2 was generated using optimal parameter value by tuning k=1. The ROUGE-2 result of ExpQueryOpt on different k value is shown in Table 3. We can see that the lowest result is still comparable to *QL*.

TABLE 3
The Accuracy Of ExpQueryOpt On Different K Value

k	1	2	3	4	5	6	7	8	9	10
R-2	.302	.298	.295	.278	.277	.279	.279	.285	.289	.289

5.2 Using Individual Quality of CQA Answers

To investigate the effect of quality of related CQA answers on summary accuracy, we generated summaries of documents using different quality of CQA answers. As described in section 3.3.3, the data were split into groups based on the quality of their related CQA answers. In this experiment, *LCA*, *AnswerOpt*, *ExpQueryOpt* and *MK++* methods were re-run to generate summaries of documents in each group using their related CQA answers in the corresponding group. The parameters estimated earlier for *AnswerOpt*, *MK++* and *MK* through 9-fold CV were averaged and used to generate the summaries. This is due to the group size imbalance, which makes parameter optimization ineffective for smaller groups. Note that for each group, all methods generated summaries from the same set of documents.

This follow-up experiment focused on only top baseline methods, such as *RelSent*, *LCA*, *QL*, and *MK*, according to the previous experimental results. For brevity, we only report the result of ROUGE-2 scores in this section.

5.2.1. Summary Accuracy

Recall that we use two different proxies of CQA quality: question relevance and answer relevance. To get an idea of which proxy correlated more with summary accuracy, we calculated Spearman correlation between the quality proxies and ROUGE-2 scores of summaries that were generated using CQA answers in such quality. The correlation coefficients between question relevance and ROUGE-2 scores of *LCA*, *AnswerOpt*, *ExpQueryOpt*, and *MK++* are respectively 0.102, 0.168, 0.099, and 0.094. When answer relevance is used as the quality proxy, the corresponding coefficients become 0.091, 0.262, 0.120, and 0.072, respectively. All the correlation coefficients are significant ($p<0.01$). On average, answer relevance correlates slightly better with summary accuracy than question relevance does. Therefore, considering the space limit in

this paper, we describe in more detail the result of summaries generated using individual quality of CQA answers, with answer relevance is used as the quality proxy. The result of using question relevance as the quality proxy is briefly described at the end of this section.

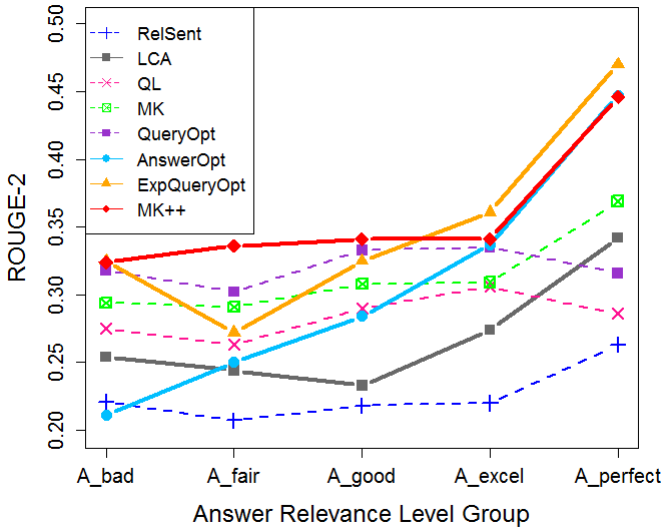

Fig. 4 Summary Accuracy on Each Answer Relevance Level Group (Note: y-axis is started from 0.20)

Figure 4 illustrates the result of each method in each answer relevance level group. Methods with access to related CQA answers are indicated by solid lines, and those without such access are indicated by dashed lines. The access to related CQA answers (solid lines) tends to respond with increased summary quality to the improved underlying answer relevance, and this trend is not observed for those without such access (dashed lines). The accuracy of *AnswerOpt* and *ExpQueryOpt* summaries improved with increasing quality of CQA answers. The accuracy of learning-to-rank-based *MK++* summaries appears stable across various levels of CQA answer quality, but increased markedly on perfect answers.

The results of *MK++* appear to be comparable to *QueryOpt* on A_bad, A_good, and A_excel groups. Although the *MK++* seems to be superior to *QueryOpt* on A_fair group, the difference is not significant based on a t-test (p=0.06). These results are inline with the results obtained in section 5.1 when these summaries were generated using a mixed quality of CQA answers. However, when the summaries are generated using perfect answers, *MK++* significantly outperforms *QueryOpt*.

To examine whether there was a significant effect of CQA answers quality on the accuracy of summaries, we calculated a one-way ANOVA (using 0.05 significance level) for each system summaries that was generated using CQA answers. The result shows that there is a significant effect of CQA answers quality on the accuracy of *LCA* (F(4,1140)=2.531, p=0.039), *AnswerOpt* (F(4,1140)= 10.809, p<0.01) and *ExpQueryOpt* summaries (F(4,1140)= 6.746, p<0.01). No significant effect of CQA quality was found with *MK++* summaries (F(4,1140)=2.311, p=0.056).

To understand which pairs of answer relevance level groups are significantly different, a Tukey post-hoc test was then performed on *LCA*, *AnswerOpt* and *ExpQueryOpt*. For *LCA*, the significance is found in one pair of groups: A_fair and A_perfect. For *ExpQueryOpt*, the significance was found in four pair of groups: (i) A_bad and A_perfect; (ii) A_fair and A_excellent; (iii) A_fair and A_perfect; and (iv) A_good and A_perfect. The highest number of significance case was found for *AnswerOpt* with six pair of groups: (i) A_bad and A_excellent; (ii) A_bad and A_perfect; (iii) A_fair and A_excellent; (iv) A_fair and A_perfect; (v) A_good and A_perfect; and (vi) A_excellent and A_perfect.

We also computed a one-way ANOVA test for other summaries that do not use related CQA answers: *RelSent*, *QL*, *MK*, and *QueryOpt*. If the significant effect of CQA quality was found in these methods, then the one found above for *AnswerOpt* and *ExpQueryOpt* might not be due to the effect of the quality of CQA answers. However, the result of ANOVA test shows that none of these methods encounter significant effect of CQA quality.

TABLE 4
THE SIGNIFICANT DIFFERENCE OF EACH METHOD

	A_bad	A_fair	A_good	A_excel	A_perfect
RelSent					
LCA		r		r	
QL	R	r	r,l	r	
MK	r,l	r,l	r,l	r	r,q
QueryOpt	r,l,q	r,l,q	r,l,q	r,l	
AnswerOpt		r	r,l	r,l	r,l,q
ExpQueryOpt	r,l	r	r,l	r,l,q,m	r,l,q,m
MK++	r,l,q,m	r,l,q,m	r,l,q,m	r,l,m	r,l,q,m

Letters r, l, q, and m mean significantly better than RelSent, LCA, QL, and MK as measured by paired t-test (p<0.05).

Table 4 describes the significant difference of each method against *RelSent*, *LCA*, *QL*, and *MK* in each answer relevance level group. From Figure 4 and Table 4, we can conclude that less than perfect CQA answers are useful in extracting answer-biased summaries from documents. Summaries generated using *AnswerOpt* with medium quality answers were significantly more accurate than *RelSent* and *LCA*, while they also managed to outperform *QL* when high quality answers were used. *ExpQueryOpt* were significantly better than *LCA* in almost all groups, while it further improved *QL* and *MK* when higher quality answers were used. *MK++* was quite robust, in which they were significantly more accurate than *MK* for all quality level of CQA answers.

Using question relevance as a CQA quality proxy in general reveals a similar trend. A significant effect of CQA answers quality was found on our proposed optimization-based *AnswerOpt* (F(3,915)=5.234, p=0.007) and *ExpQueryOpt* (F(3,915)=2.663, p=0.047) summaries. Again, no significance was found with our proposed learning-to-rank-based *MK++* (F(3,915)=2.355, p=0.071) summaries.

5.2.2 Effect of Varying k

The result of *ExpQueryOpt* displayed in Figure 4 is based on an optimized k in the range of 1 to 10. The result for all other k values is illustrated in Figure 5. It appears that summaries generated using high quality CQA an-

swers tend to achieve high accuracy for medium *k* value. It is because when the related CQA answers are of high quality, expansion terms are likely to be more relevant.

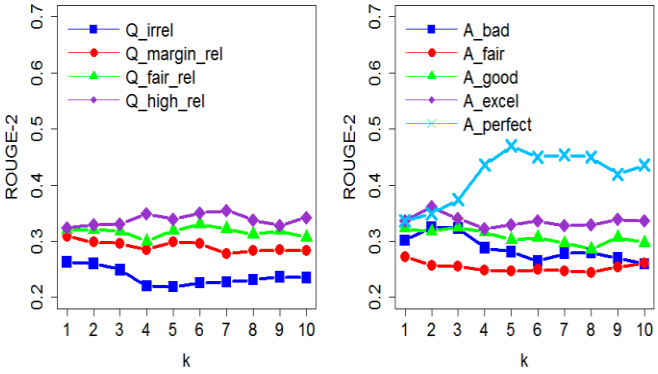

Fig. 5 Varying *k* Value on Each Question Relevance Level Group (left) and Answer Relevance Level Group (right)

The right side of the figure shows that the line for A_bad is higher than A_fair. Recall that *ExpQueryOpt* is an extension of *QueryOpt*, and in Figure 4 it also appears that that the accuracy of *QueryOpt* for A_bad was also higher than A_fair. Then, notice that this trend also appears for other summaries that were generated based on query-based techniques: *RelSent, LCA* and *QL*. This indicates that the ground truth answers in A_bad share many similar vocabularies with the queries than A_fair.

5.2.3 Effect of Query Expansion

To better understand the effect of query expansion on the accuracy of *ExpQueryOpt* summaries on each group, we calculated the proportion of summaries for which the accuracies are improved (I), decreased (D), and unchanged (U) by the expansion. This is measured by comparing the accuracy of summaries generated with and without using query expansion. The version without using query expansion is actually *QueryOpt*.

TABLE 5
THE EFFECT OF QUERY EXPANSION ON SUMMARY ACCURACY

Question Relevance Level Group	I (%)	D (%)	U (%)	Answer Relevance Level Group	I (%)	D (%)	U (%)
Q_irrel (k=1)	15.1	28.7	56.2	A_bad (k=2)	19.0	27.8	53.2
Q_margin_rel (k=1)	16.6	23.7	59.7	A_fair (k=1)	17.0	26.8	56.2
Q_fair_rel (k=6)	31.1	37.4	31.5	A_good (k=3)	26.7	34.9	38.4
Q_high_rel (k=7)	34.9	32.5	32.5	A_excel (k=2)	27.0	26.6	46.4
				A_perfect (k=5)	44.6	26.5	28.9

By using query expansion, the proportion of summaries with improved accuracies is found to be higher for groups with better CQA answer quality. An exception is for the statistics of A_fair as compared to A_bad. This could be explained because there is minimal difference between "bad" and "fair" CQA quality. The proportion of improved summaries is the highest in group Q_highly_relevant and A_perfect.

6 EXTRACTING SUMMARIES WHEN RELATED CQA ANSWERS MAYBE UNAVAILABLE

In the previous experiments, we focused on queries with related CQA answers. This setting however does not provide much insight as to how the propose methods respond to the availability of CQA resources.

In this section, we simulate the situation where the related CQA answers for some queries may be unavailable. We used 35 queries in our dataset that do not have related CQA answers, together with the corresponding 291 documents and 1,406 ground truth answers. Because the CQA answers are unavailable for this data, the *LCA*, *AnswerOpt*, and *ExpQueryOpt* summaries could not be generated in this experiment. To generate *MK++* summaries, initially the training process was conducted using 45 queries that have related CQA answers. Then, the model was used to generate summaries of documents associated with 35 queries. For consistency, this proportion of training and testing data was also applied to *MK*.

TABLE 6
SUMMARY ACCURACY OF DOCUMENTS ASSOCIATED WITH QUERIES WITHOUT RELATED CQA ANSWERS

Method	ROUGE-1	ROUGE-2
Lead	0.336	0.120
MEAD	0.415	0.160
DocOpt	0.428	0.188
RelSent	0.425	0.188
QL	0.483 r	0.262 r
MK	0.539 r,q	0.331 r,q
QueryOpt	0.526 r,q	0.310 r,q
MK++	**0.553** r,q,m	**0.352** r,q,m

Superscripts r, q, and m indicate a significant difference respectively against RelSent, QL and MK, as measured by paired t-test ($p<0.05$).

As can be seen in Table 6, our proposed methods *QueryOpt* and *MK++* consistently perform well and significantly outperform state-of-the-art techniques. *MK++* is shown to be the best system, and it is also significantly better than *QueryOpt*. This suggest that the *MK++* model trained on queries with related CQA answers remains useful for enhancing answer-biased summaries for queries without access to related CQA answers. *MK++* seems to be less sensitive to the change in CQA quality.

7 ANALYSIS

7.1 Document Summaries vs CQA Answers

We also investigated the effectiveness of applying the idea of Weber et al [8] that used the most relevant answer from CQA as an answer to the query. The quality judgment of the CQA answer retrieved in the top result for each of 45 queries were examined. The numbers of Bad, Fair, Good, Excellent, and Perfect answers are 9, 8, 19, 7, and 2, respectively. Following Weber et al., there are only

2 queries that can be answered perfectly using CQA answers. The top related CQA answers for these 2 queries, however, were found to be lengthy: 202 and 680 words. It is not clear whether this long text is effective to be presented as answers in search result page.

We conducted an analysis to demonstrate the advantage of using related CQA answers to extract answer-biased summaries over using CQA answers directly. We made this comparison with respect to the approach of Weber et al. by comparing our summaries and the top related CQA answer against ground truth answers. To be fair, the length of CQA answers is made into 50 words by taking the leading text. The number of cases where the ROUGE-2 score of our summaries are better, worse, and equal to the score of top related CQA answer are displayed in Table 7. We can see that the number of cases where our summaries are found better than CQA answers is much higher than the number where they are worse.

TABLE 7
DOCUMENT SUMMARIES VS CQA ANSWERS

	Better	Worse	Equal	Total
AnswerOpt	313	80	26	419
ExpQueryOpt	332	67	20	419
MK++	340	65	14	419

7.2 Example of Answer-biased Summaries

Figure 6 displays an example of answer-biased summaries that were extracted from doc "GX013-38-0620682" associated to query: "*Who is Pol Pot and what did he do?*".

***AnswerOpt* and *ExpQueryOpt* (ROUGE-2: 1.00)**
***Pol Pot** led the Khmer Rouge regime that is held responsible for the deaths of up to two million Cambodians.* All the senior Khmer Rouge leaders responsible for the atrocities committed in Cambodia during the Khmer Rouge rule from 1975 to 1979 should be brought to justice," Rubin said.

***MK*++ (ROUGE-2: 1.00)**
POL POT/CAMBODIA -- The US government cannot confirm that **Pol Pot** is dead, but "we have no reason to dispute the rather compelling reports of his death," Rubin said. ***Pol Pot** led the Khmer Rouge regime that is held responsible for the deaths of up to two million Cambodians.*

***QL* (ROUGE-2: 0.08)**
[daiywash.gif] *EPF401 04/16/98 TRANSCRIPT: STATE DEPARTMENT NOON BRIEFING, APRIL 16, 1998 (**Pol Pot**/Cambodia, Russia/Iran/missiles) (4620) State Department Spokesman James Rubin briefed. **POL POT**/CAMBODIA -- The US government cannot confirm that **Pol Pot** is dead, but "we have no reason to dispute the rather compelling reports…

***MK* (ROUGE-2: 0.2)**
POL POT/CAMBODIA -- The US government cannot confirm that **Pol Pot** is dead, but "we have no reason to dispute the rather compelling reports of his death," Rubin said. We have no basis for suggesting that **Pol Pot** is alive. Although **Pol Pot** was the most notorious leader, the Khmer …

Fig 6. An example of answer-biased summaries

The *AnswerOpt*, *ExpQueryOpt*, and *MK*++ summaries benefit from high quality related CQA answers and achieved a perfect ROUGE-2 score. They significantly improve state-of-the-art *QL* and *MK* by extracting sentences that contain a ground truth answer to the query (italized in the figure). *ExpQueryOpt* improves *QueryOpt* by expanding the query with terms that were extracted from related YA answers: "*cambodia*", "*khmer*", "*rouge*", "*1979*", and "*leader*". This example shows that using external information from related CQA answers could help to bridge the lexical chasm between queries and answers.

8 FURTHER INVESTIGATION

8.1 Using the Key Concept of Query

Much research explored the advantage of using key concept of queries (also called question focus [28]) for improving the ranking of documents [50] and CQA answers [51]. Focussing on the more important part of the query makes more effective term weighting and may, as a result, improve the ranking. Recall that in this work we used long queries in WebAP dataset for the sake of consistency with the ones that were used to retrieve documents and generate human annotated answers in the dataset. Here, we want to investigate whether the use of key concepts can help in improving the accuracy of generated summaries. We used a similar set of documents and ground truth answers as used in section 5.1.

TABLE 8
SUMMARY ACCURACY USING KEY CONCEPT OF QUERIES

Method	**R-1**	**R-2**	**Δ R-1**	**Δ R-2**
MEAD	0.375	0.128	0.000	0.000
RelSent	0.435	0.215	+ 0.004	+ 0.003
LCA	0.471	0.265	- 0.005	+ 0.001
QL	0.511	0.295	+ 0.017	+ 0.017
MK	0.517	0.300	+ 0.002	+ 0.003
QueryOpt	0.514	0.310	- 0.014	- 0.015
AnswerOpt	**0.542***	**0.330***	**+ 0.025**	**+ 0.039**
ExpQueryOpt	0.534	0.311	+ 0.015	+ 0.009
MK++	0.513	0.299	- 0.007	- 0.011

*Symbol * indicates a significant difference over respective methods that use original queries, as measured by paired t-test ($p<0.05$).*

All methods that use query information were re-run using short query taken from title field in the query specification file of TREC. Here, we assume that such title can be used as a good representative of key concept for our queries. The related CQA answers were obtained by submitting the key concept of query to YA search engine. The results are displayed in Table 8. Column "ΔR-1" and "ΔR-2" contains the difference of ROUGE scores between using key concept and not, by subtracting the scores in Table 8 with those in Table 2. For any method, a positive difference means that the method becomes more effective when key concepts are available.

QueryOpt, is shown to have decreasing accuracy. It is because removing some query terms which may also important, can reduce the coverage of summaries. Besides that, when queries are short, there are many sentences that will have the same representation in *QueryOpt*. Therefore the algorithm lacks of knowledge in accurately

selecting answer sentences. The score of learning-to-rank-based *MK* and *MK++* changes little, indicating that they are less sensitive to the effect of key concept.

The accuracy of *QL*, *ExpQueryOpt* and *AnswerOpt* showed some level of improvement. This is shown to be significant for *AnswerOpt* which can be explained because as described in section 5.2, *AnswerOpt* is the system which correlates the most with the quality of related CQA answers, and using key concept of queries may improve the retrieval of such answers by CQA search engine. It is worth mentioning that when submitting key concept of queries to YA search engine, there are more queries that have returned results (i.e. 79 vs 45) and there are relatively higher average results per query (i.e. 9.6 vs 6.8).

8.2 Using Another Dataset

8.2.1 MSMARCO Dataset

To investigate the robustness of our methods, we apply our techniqus on another dataset, called MSMARCO [7]. Nguyen et al. [7] performed the following steps to build this dataset: (1) filter the queries from Bing logs that ask for a question; (2) retrieve top 10 relevant passages using their IR system; and (3) ask judges to select the passages that contain answers to the query, and then write natural language answers based on the selected passages.

The MSMARCO dataset contains 100K queries with each has a set of relevant passages, the document URLs for which each passage is extracted from, the annotated answers, and the query type. There are five query types in the dataset: numeric, entity, location, person, and description. In our experiment, we select a subset of queries from MSMARCO data, based on the following criteria:

- Queries that have one or more valid answers (some were left without any answer in the original data).
- Queries with all their respective answers coming from only one underlying document. This is to be made consistent with the assumption behind single document summarization.
- Queries with the type of description, and with answers of sufficient length. Following the guideline for average sentence length for plain English [52], we applied 15 words as the minimum answer length.

The above steps result in 23,999 queries. For our experiment, a random sampling was performed to select 1,000 queries for which the document URL can be crawled and the related YA answers are available. The human generated answers are then used as ground truth answers. Web documents are parsed using JSoup[10] and split into sentences using Lingua[11]. To generate training data for the learning-to-rank-based approach, i.e. *MK* and *MK++*, sentences in the selected passages and in the web documents were manually aligned to generate label of answer sentences in 1000 web documents in our dataset.

Most of the queries in our dataset are associated with one answer where the average number of answer per query is 1.1. The average answer length is 31.4 words. The Web documents on average consist of 284 sentences.

[10] http://jsoup.com
[11] http://search.cpan.org

8.2.2 Result

The similar methods described in section 4 are then applied to generate a 50-word answer-biased summary from each document. The results are displayed in Table 9.

MK++ consistently performs well, and it achieves the best result that is significantly better than *RelSent*, *LCA*, *QL*, and *MK* baselines. This confirms the reliability of this technique that is shown in section 5 and 6 above.

TABLE 9
SUMMARY ACCURACY USING MIXED QUALITY CQA ANSWERS

Method	ROUGE-1	ROUGE-2
Lead	0.285	0.146
MEAD	0.453	0.260
DocOpt	0.348	0.143
RelSent	0.510	0.341
LCA	0.538	0.363
QL	0.452	0.280
MK	0.529	0.360
QueryOpt	0.415	0.232
AnswerOpt	0.487 q	0.272
ExpQueryOpt	0.478 q	0.263
MK++	**0.551** r,q,m	**0.385** r,l,q,m

The proposed methods are printed in boldface. Superscripts r, l, q, m indicate a significant difference respectively against RelSent, LCA, QL and MK, as measured by paired t-test ($p<0.05$).

It appears from the table that the three optimization-based approaches, in general, have less accuracy than the baselines using query-based techniques such as: *RelSent*, *LCA*, and *MK*. However, in comparison to *QL* result, *AnswerOpt* and *ExpQueryOpt* can give significant improvement on ROUGE-1, but are comparable on ROUGE-2. This is analysed in more detail in next subsection.

ExpQueryOpt can significantly improve *QueryOpt* by 15.2% and 13.4% respectively on ROUGE-1 and ROUGE-2. *ExpQueryOpt* result reported in Table 9 is generated using optimal parameter value by tuning k=5. The ROUGE-2 result of *ExpQueryOpt* on different k value is shown in Table 10.

TABLE 10
THE ACCURACY OF EXPQUERYOPT ON DIFFERENT K VALUE

k	1	2	3	4	5	6	7	8	9	10
R-2	.220	.236	.252	.258	.263	.261	.249	.250	.251	.248

8.2.3 Analysis

The accuracy of *QueryOpt* is lower than the one displayed in Table 2 when using WebAP data. It can be explained because the average non-stopword query length in MSMARCO data (e.g. 2.6) is relatively shorter than the query in WebAP data (e.g. 6.1). An explanation for this result is similar to the one described in section 8.1 above when we use shorter query format.

In contrast to the result in Table 2, here *ExpQueryOpt* can improve *QueryOpt*. We suspect that besides the CQA quality effect that was investigated earlier in section 5.2, query verbosity may also influence the effectiveness of query expansion performed in *ExpQueryOpt*. When the queries are relatively short, expansion terms are likely to be more accurate. This then enables *ExpQueryOpt* to im-

prove *QueryOpt* technique in selecting answer sentences.

One highlight from Table 9 is almost all query-based baselines have high ROUGE scores, even *MEAD* system which is shown very inferior in section 4. We analyze that this might be influenced by the way the answers in the dataset was generated. Recall that the human answers are written only based on the information contained in the relevant passages that were initially retrieved using separate IR system. As commonly known, current IR system still relies on query-based technique to retrieve the results as a response to the query. To investigate this issue, we calculated ROUGE scores between query and ground truth answers in MSMARCO. As a comparison, the calculation is also performed for WebAP dataset (see Table 11). It appears that the query term overlap in ground truth answers in MSMARCO data is significantly higher than the ones in WebAP data. It confirms our presumption on the bias of the ground truth answers in MSMARCO data.

TABLE 11
QUERY TERM OVERLAP IN GROUND TRUTH ANSWERS

Dataset	ROUGE-1	ROUGE-2
WebAP	0.361	0.109
MSMARCO	0.573†	0.208†

Symbol † indicates significant difference against WebAP (p<0.05).

In addition, ground truth answers in MSMARCO data may have incompleteness issue. They possibly missed other sentences that also contain answers to the query (but are not retrieved in the initially relevant passages). This may underrate the summaries. Fig 7 presents an example of *AnswerOpt* summary whose ROUGE-2 scores are zero, although they essentially contain answers. Sentences that contain answers are boldfaced.

QID 28002: what causes childhood epilepsy
Important Information about epilepsy and seizures…. Causes of Epilepsy in Childhood. **Less common causes of childhood epilepsy include brain tumors or cysts and degenerative disorders (progressive and deteriorating conditions, often associated with loss of brain cells).** All people are capable of having a seizure. New to Epilepsy and Seizures?

Fig. 7 An Example of AnswerOpt Summary that Are Underrated Because of Incomplete Ground Truth Answers

Our results show that our proposed optimization-based methods are superior to query-based techniques when they are applied to WebAP data, but this is not the case when MSMARCO data is used. Recall that WebAP data was built based on GOV2 collection which contains ~ 25 million documents, and MSMARCO data was built based on Bing index which has billions of documents. We argue that when the data size is relatively small, so that the probability to include documents containing high number of co-occurences with query terms is smaller, our optimization-based methods could be more beneficial than query-based techniques. Another situation that may be benefited by our approach is when queries are getting more complex and longer, so that relying on co-occurrence statistics with query terms may not be sufficient to find the answers.

At last, it is important to note that our proposed method MK++ can consistently gain high accuracy on both WebAP and MSMARCO dataset. This confirms the robustness of this method to exploit CQA content for extracting answer-biased summaries from documents.

9 DISCUSSION AND FUTURE WORK

As described in section 7, the idea of extracting answer-biased summaries from retrieved documents with the help of CQA content can benefit from an approach such as used by Weber et al. [8] that directly takes the most relevant answer in CQA as the answer to the query. Next in section 6, we showed another benefit of our approach: using the model learned in our method can help to improve the accuracy of summaries for which the related CQA content are unavailable.

The queries and the CQA answers that we used in this work came from different collections that were independent of each other. However, the coverage of CQA for our queries is found to be reasonable which further seems to indicate a promising sign of our proposed approach.

When the retrieved documents are CQA pages, one may wonder the possibility of directly using answers displayed in the pages as the summaries. Because WebAP does not contain any CQA pages, we analyzed this issue only using MSMARCO data. We found there are 56 (out of 1000) documents in the data are CQA pages. We manually extracted the best answer displayed on the page, and took the leading sentences to be compared with our summaries generated earlier. According to ROUGE scores, the leading text of answers was shown to be more accurate than summaries. This suggests that when retrieved documents come from CQA, it may be preferable to directly use the leading text of answers as the summaries. This approach is effective when the CQA answers are concise and focused in answering the queries. Despite this result, it is worth noting that the proportion of retrieved documents that are CQA pages in our dataset is relatively low, implying that our summarization techniques still give high benefit for the rest of the documents.

Our results reveal that as measured by ROUGE, leading sentences are not accurate to serve as answer-biased summaries. While some previous research, such as [29], found that leading sentences are effective in creating snippets that can help searchers to judge the relevance of documents, this is not the case when they are aimed to present answer. The result of a state-of-the-art query likelihood passage retrieval method was also not satisfactory, confirming the result of previous work [9].

We suggest few things that could improve *ExpQueryOpt* and *AnswerOpt* methods. First, by enhancing question retrieval in CQA for long queries such as by identifying key concepts [50]. Second, by incorporating "answer quality predictor" [17], [19] to remove related CQA answers with bad quality. These suggestions are based on the result described in section 8.1 and 5.2.

We made some recommendation about the best use of our proposed methods in regard to the quality of related CQA answers that are available. When a large number of training data are available, then *MK++* is preferred. This

is as shown in our results that this technique is consistent to perform well in different quality level of CQA and in two different datasets tested in this work. As building such training resources are expensive, it is worth considering other alternatives. When related answers are either not available or all in low quality, *QueryOpt* is preferred if the queries are long, but query-based techniques (e.g. *QL* or *RelSent*) may be more beneficial if the queries are short. When the quality of related answers is medium to high, then *AnswerOpt* or *ExpQueryOpt* may be more accurate.

In this work, we used a bag-of-words representation of sentences, and then related CQA content was used to weight more important words that are expected to bear answers. In future, it is possible to use more advanced representation of words (e.g. word embedding [53]) or sentences (e.g. Explicit Semantic Analysis [54]). We may, for example calculate semantic similarity between sentences and queries, such as performed by Yang et al. [16], and incorporate this feature into *MK++*.

Some past work has studied the effectiveness of answers for popular factoid queries using log-based analysis [1], [4] and controlled user studies [2], [12]; and found that it lead to good abandonment and increased user satisfaction. Another work [3] conducted a similar study using tail queries (mostly non-factoid) and drew the same conclusion as previous work [1], [2], [4], [12]. All of this past work confirms that presenting answers as a response to the query has a significant positive effect on the user search experience. In future, we plan to conduct user studies to evaluate the accuracy of our summaries. We may run pairwise comparisons of summaries [32], and ask users to rate their satisfaction rating [3], [12].

10 CONCLUSION

We propose to use external information from related CQA content to guide the extraction of an answer-biased summary from each retrieved document. Three optimization-based methods and a learning-to-rank-based method were proposed. Our results show that the related CQA content, that do not necessarily contain perfect answer to the query, are useful to extract better answer-biased summaries from documents. This answers **RQ1**. The quality of CQA content is shown to have significant effect to the accuracy of optimization-based summaries. In contrast, the significant effect of CQA quality is not found on the accuracy of learning-to-rank-based summaries. The learning-to-rank-based method consistently performs well on different level of CQA quality. This answers **RQ2**.

ACKNOWLEDGMENTS

This research is supported in part by the Australian Research Council (DP140102655) and the Indonesia Endowment Fund for Education (LPDP).

REFERENCES

[1] L. B. Chilton and J. Teevan, "Addressing People's Information Needs Directly in a Web Search Result Page," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 27–36.

[2] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam, "Towards Better Measurement of Attention and Satisfaction in Mobile Search," in *Proc. 37th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 113–122.

[3] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz, "Direct Answers for Search Queries in the Long Tail," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 237–246.

[4] J. Li, S. Huffman, and A. Tokuda, "Good abandonment in mobile and PC internet search," in *Proc. 32nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 43–50.

[5] M. D. Smucker and C. L. Clarke, "Time-based calibration of effectiveness measures," in *Proc. 35th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 95–104.

[6] V. Jijkoun and M. de Rijke, "Retrieving answers from frequently asked questions pages on the web," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 76–83.

[7] T. Nguyen *et al.*, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," *in Proc. NIPS Workshop*, 2016.

[8] I. Weber, A. Ukkonen, and A. Gionis, "Answers, Not Links: Extracting Tips from Yahoo! Answers to Address How-to Web Queries," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 613–622.

[9] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson, "Retrieving Passages and Finding Answers," in *Proc. Australasian Document Computing Symposium*, 2014, pp. 81–84.

[10] R. Soricut and E. Brill, "Automatic Question Answering Using the Web: Beyond the Factoid," *Inf. Retr.*, vol. 9, no. 2, pp. 191–206, Mar. 2006.

[11] E. Cutrell and Z. Guan, "What are you looking for?: an eye-tracking study of information usage in web search," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 407–416.

[12] K. Williams, J. Kiseleva, A. C. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa, "Detecting good abandonment in mobile search," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 495–505.

[13] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor, "When web search fails, searchers become askers: understanding the transition," in *Proc. 35th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 801–810.

[14] O. Kolomiyets and M.-F. Moens, "A survey on question answering technology from an information retrieval perspective," *Inf. Sci.*, vol. 181, no. 24, pp. 5412–5434, Dec. 2011.

[15] H. T. Dang, D. Kelly, and J. J. Lin, "Overview of the TREC 2007 Question Answering Track.," in Proc. *TREC*, 2007, vol. 7, p. 63.

[16] L. Yang *et al.*, "Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval," in *Adv. Inf. Retrieval: 38th European Conf. Inf. Retrieval*, 2016, pp. 115–128.

[17] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *Proc. 33rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 411–418.

[18] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 483–490.

[19] Q. Liu *et al.*, "Predicting Web Searcher Satisfaction with Existing Community-based Answers," in *Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 415–424.

[20] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the right facts in the crowd: factoid question answering over social media," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 467–476.

[21] C. L. Chiang, S. Y. Chen, and P. J. Cheng, "Summarizing Search Results with Community-Based Question Answering," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) and Intell. Agent Tech. (IAT)-Volume 01*, 2014, pp. 254–261.

[22] V. Pande, T. Mukherjee, and V. Varma, "Summarizing Answers for Community Question Answer Services," in *Lang. Proc. Knowl. Web*, Springer, 2013, pp. 151–161.

[23] J. He and D. Dai, "Summarization of Yes/No Questions Using a Feature Function Model.," in *ACML*, 2011, pp. 351–366.

[24] E. Agichtein, D. Carmel, D. Harman, D. Pelleg, and Y. Pinter, "Overview of the TREC 2015 LiveQA Track," in *Proc. TREC.*, 2015.

[25] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 475–482.

[26] M. Surdeanu, M. Ciaramita, and H. Zaragoza, "Learning to Rank Answers on Large Online QA Collections," in *Proc. ACL*, 2008, pp. 719–727.
[27] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the Past: Answering New Questions with Past Answers," in *Proc. WWW*, 2012, pp. 759–768.
[28] Y. Wu, C. Hori, H. Kashioka, and H. Kawai, "Leveraging social Q&A collections for improving complex question answering," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 1–19, 2015.
[29] A. Tombros and M. Sanderson, "Advantages of query biased summaries in information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 2–10.
[30] H. Takamura and M. Okumura, "Text Summarization Model Based on Maximum Coverage Problem and Its Variant," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2009, pp. 781–789.
[31] D. Metzler and T. Kanungo, "Machine learned sentence selection strategies for query-biased summarization," in *SIGIR Learning to Rank Workshop*, 2008, pp. 40–47.
[32] E. Yulianti, S. Huspi, and M. Sanderson, "Tweet-biased summarization," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 6, pp. 1289–1300, 2016.
[33] M. Ageev, D. Lagun, and E. Agichtein, "Improving Search Result Summaries by Using Searcher Behavior Data," in *Proc. 36th. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 13–22.
[34] Y. Ko, H. An, and J. Seo, "Pseudo-relevance feedback and statistical query expansion for web snippet generation," *Inf. Process. Lett.*, vol. 109, no. 1, pp. 18–22, 2008.
[35] D. E. Losada, "Statistical query expansion for sentence retrieval and its effects on weak and strong queries," *Inf. Retr.*, vol. 13, no. 5, pp. 485–506, 2010.
[36] L. Leal Bando, F. Scholer, and A. Turpin, "Query-biased summary generation assisted by query expansion," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 5, pp. 961–979, 2015.
[37] F. Diaz and D. Metzler, "Improving the Estimation of Relevance Models Using Large External Corpora," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 154–161.
[38] D. Metzler and W. B. Croft, "A Markov Random Field Model for Term Dependencies," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2005, pp. 472–479.
[39] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2002, pp. 324–330.
[40] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability.," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.
[41] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.
[42] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.," *Psychol. Assess.*, vol. 6, no. 4, p. 284, 1994.
[43] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Brooks/Cole, 1996.
[44] R.-C. Chen *et al.*, "RMIT at the TREC 2015 LiveQA Track," in Proc. TREC, 2015.
[45] D. Radev *et al.*, "MEAD — A platform for multidocument multilingual text summarization," in *Conf. Lang. Res. Eval. (LREC)*, Lisbon, Portugal, 2004.
[46] J. He, P. Duboue, and J.-Y. Nie, "Bridging the Gap between Intrinsic and Perceived Relevance in Snippet Generation," in *Proc. COLING 2012*, Mumbai, India, 2012, pp. 1129–1146.
[47] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Trans. Inf. Syst.*, vol. 18, no. 1, pp. 79–112, 2000.
[48] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 314–321.
[49] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL-04 Workshop*, 2004, pp. 74–81.
[50] M. Bendersky and W. B. Croft, "Discovering key concepts in verbose queries," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 491–498.
[51] J. H. Park and W. B. Croft, "Using Key Concepts in a Translation Model for Retrieval," in *Proc. 38th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 927–930.
[52] M. Cutts, *Oxford guide to plain English*. OUP Oxford, 2013.
[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *in Proc. ICLR Workshop*, 2013.
[54] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis," in *Proc 20th Int. Joint Conf. Artif. Intell.*, San Francisco, CA, USA, 2007, pp. 1606–1611.

Evi Yulianti is a Ph.D. student in the Discipline of Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia. She is a lecturer at Faculty of Computer Science, Universitas Indonesia. She received her B.C.S. degree from Universitas Indonesia in 2010, and dual M.C.S. degree from Universitas Indonesia and RMIT University in 2013.

Ruey-Cheng Chen is a Research Fellow in the Discipline of Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia. He received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Falk Scholer is an Associate Professor in the Discipline of Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia. His research interests include the evaluation of information retrieval systems; relevance and user perceptions; document summarisation; and interactive search.

W. Bruce Croft is a Distinguished Professor at the faculty of the Department of Computer Science at the University of Massachusetts, Amherst, USA. Since 2011, he is also a Senior Research Fellow at RMIT University, Melbourne, Australia. He received a B.Sc. (Hons) degree in 1973, and an M.Sc. in Computer Science in 1974 from Monash University in Melbourne, Australia. His Ph.D. in Computer Science was from the University of Cambridge, England in 1979.

Mark Sanderson is a Professor in the Discipline of Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia. He received B.Sc. (hons) and Ph.D. degrees in computer science from the University of Glasgow, Glasgow, U.K., in 1988 and 1997, respectively. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering and ACM Transactions on the Web; and co-editor of Foundations and Trends in Information Retrieval